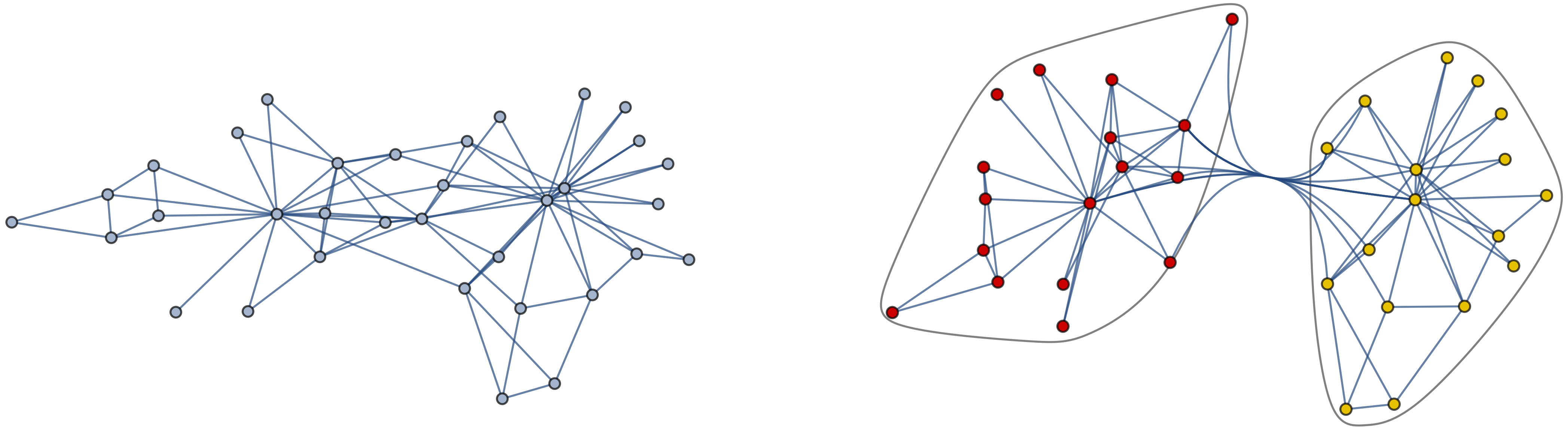
Community analytics

Communities (Zachary karate club)



Could the boundaries of the two clubs been been predicted from the structure?

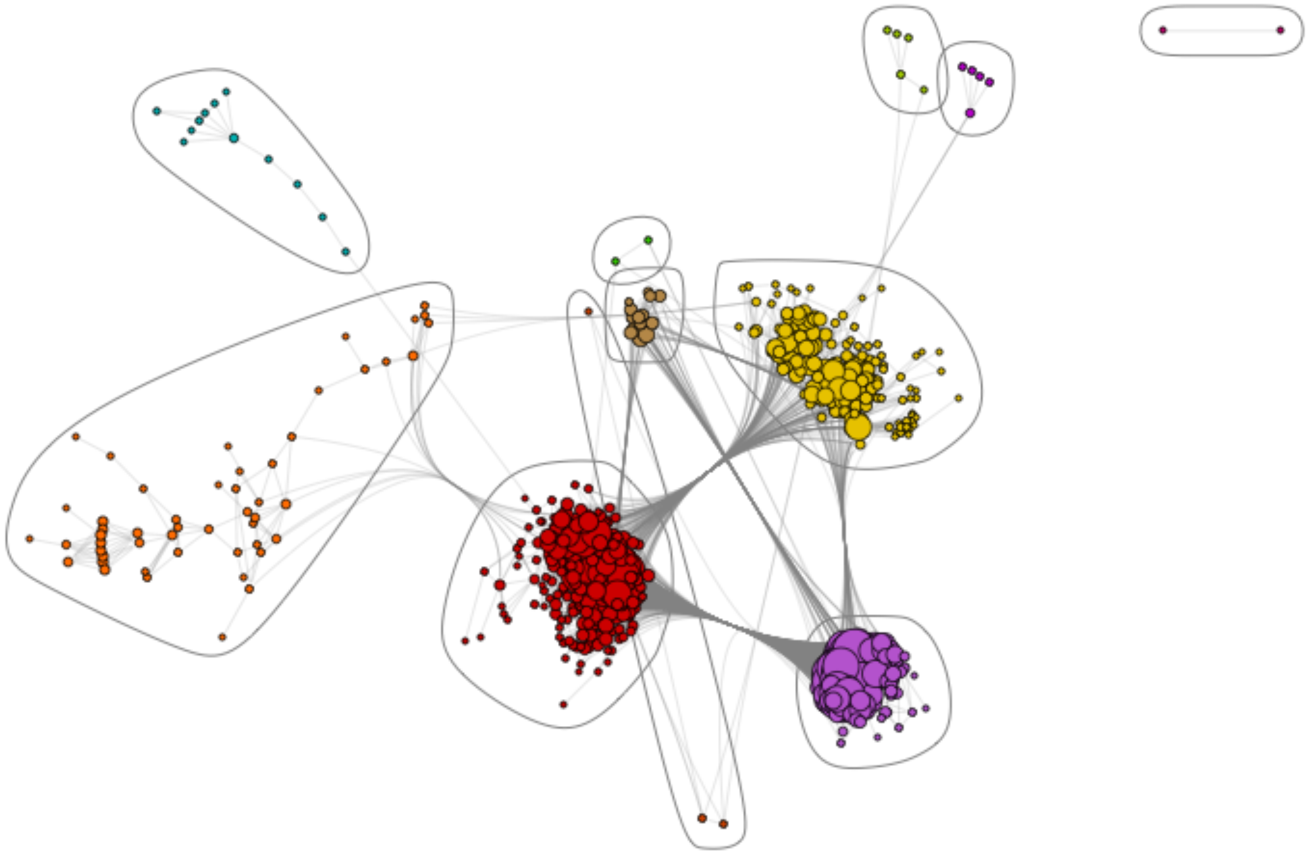
- In some scenarios, yes!
- No explicit definition of node types
- Division based on edge density
- More specifically, modularity

Community analytics

- No agreed-upon definition of community!
- Quantitative definitions of community
  - based on internal and external degrees
  - definition of community in a strong/weak sense
- Communities based on modularity (commonly-used measure)
- Analyzing modularity
  - Conditions for community structure to satisfy different definitions
  - Distribution of nodes among modules
  - Limits of modularity
- Clique percolation
- Communities with overlapping (link partitioning)

Communities (Facebook)

Groups of friends that have a higher affinity to connect to each other than to friends from other communities



Community: Generally thought of as subgraph where internal connections are denser than external ones

Communities

- Non-overlapping communities
- Weak and strong definitions of communities[1]
- Consider a graph $G = (V,E)$ and a subgraph $G_s = (V_s,E_s)$

k_i^{in} : number of edges connecting node i to other nodes belonging to G_s

k_i^{out} : number of edges connecting node i toward nodes in the rest of the network

Strong definition of community

A subgraph G_s is a community of G in the strong sense if

$$k_i^{\text{in}} > k_i^{\text{out}} \text{ for all } i \in V_s$$

Weak definition of community

A subgraph G_s is a community of G in the strong sense if

$$\sum_{i \in V_s} k_i^{\text{in}} > \sum_{i \in V_s} k_i^{\text{out}} = m_s^{\text{out}}$$

---

[1] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, Proc. Natl. Acad. Sci. 101, 2658-2663, 2004

Modularity

Idea: number of edges within a group minus the expected number in an equivalent network with edges placed at random

$V = \{v_1, \ldots, v_n\}$: set of nodes

k_i : degree of node i

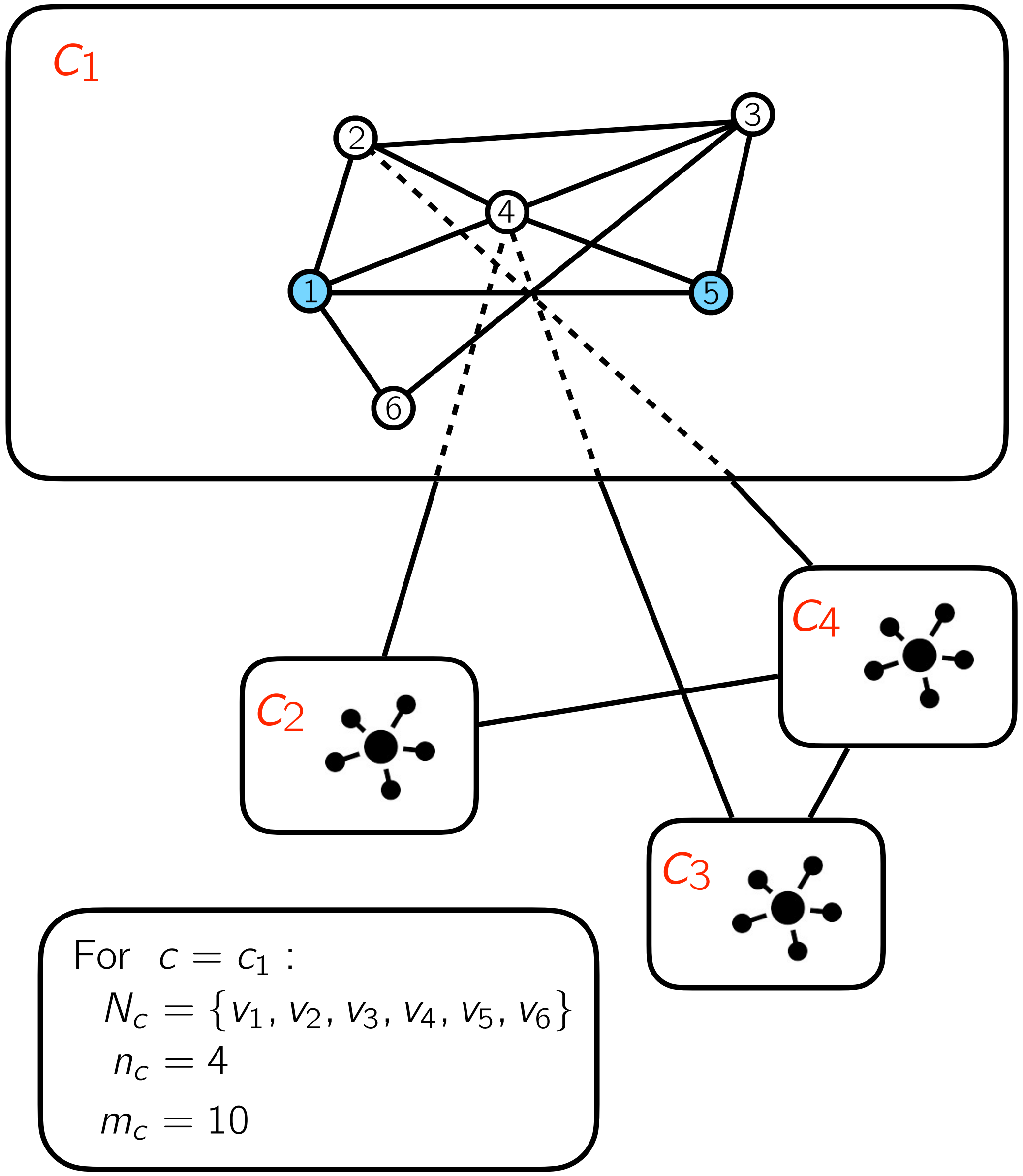
$$m = |E| = \frac{1}{2}\sum_i k_i \quad : \text{ total number of edges}$$

a_{ij} : number of edges between nodes i and j

n_c : number of communities

V_c : nodes belonging to community c

m_c : number of edges between $|V_c|$ nodes



For $c = c_1$:
$N_c = \{v_1, v_2, v_3, v_4, v_5, v_6\}$
$n_c = 4$
$m_c = 10$

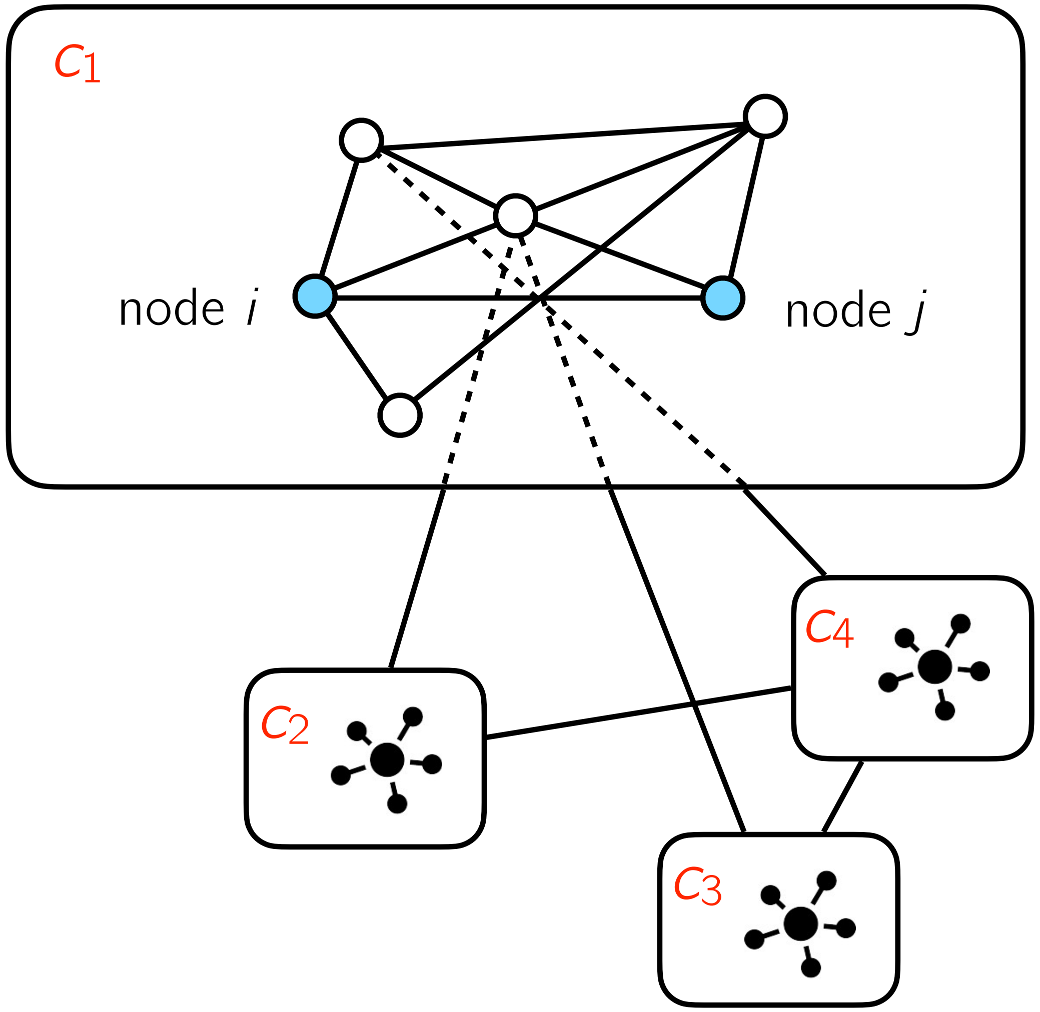
If m_c is greater than the expected value, then V_c is considered a module/community

Modularity

- Membership function

$$s \ : \ V \mapsto \{1, \ldots, n_c\}$$

$$\delta(v_i, v_j) = \begin{cases} 1, & \text{if } s(v_i) = s(v_j) \\ 0, & \text{otherwise} \end{cases}$$



- Number of edges within group c

$$\frac{1}{2} \sum_{v_i, v_j \in V_c} a_{ij}$$

- Modularity of group c

$$M_c = \underset{\text{normalization factor}}{\frac{1}{2m}} \sum_{v_i, v_j \in V_c} (a_{ij} - \underset{\text{probability of a random edge (based on degree preserving model)}}{p_{ij}})$$

normalization factor

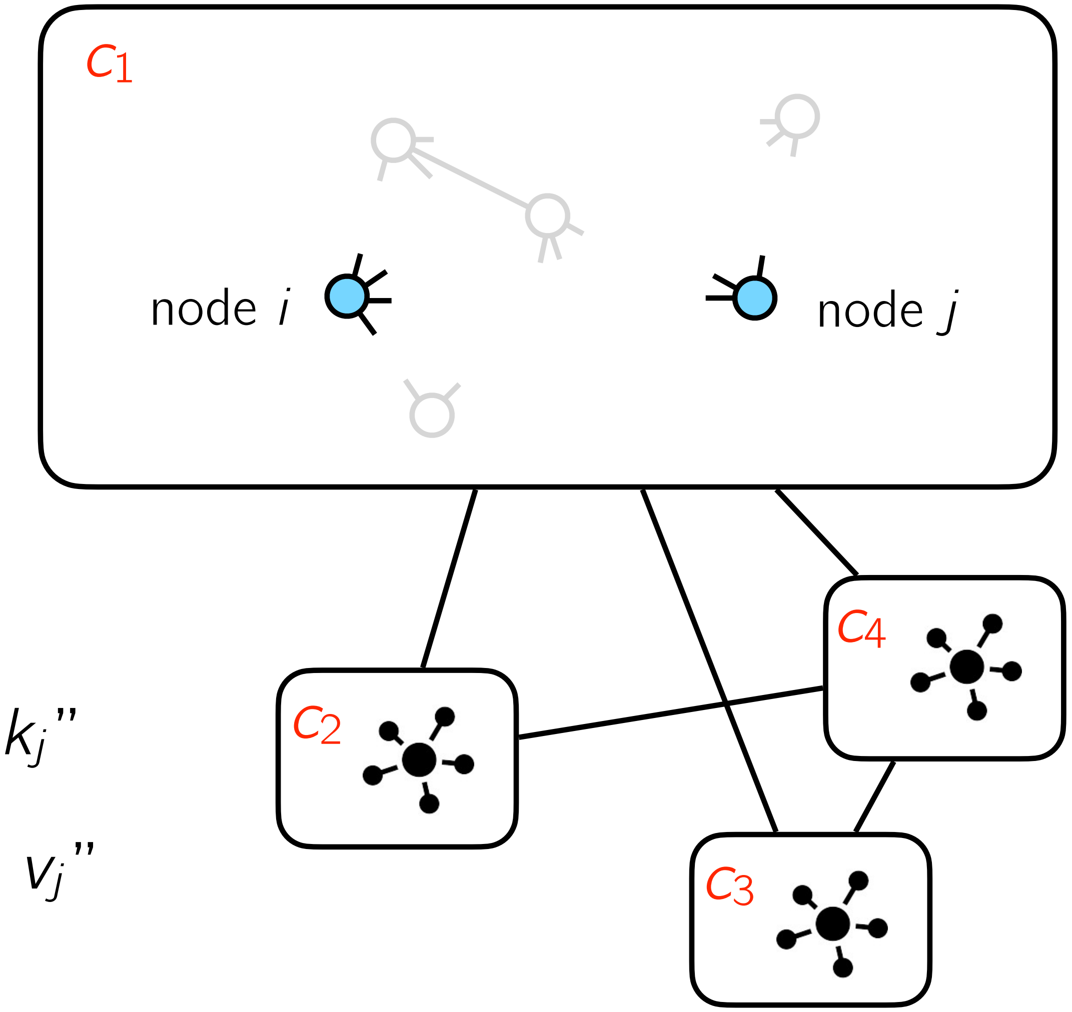
probability of a random edge
(based on degree preserving model)

Network modularity

$$M = \sum_{c=1}^{n_c} M_c = \sum_{v_i, v_j \in V} (a_{ij} - p_{ij})\delta(v_i, v_j)$$

Modularity

- Degree preserving null model:
  - equivalent network
  - same number of edges m
  - edges placed at random
- Consider v_i and v_j with k_i and k_j
- Not necessary within the same group
- Event A = "node v_i with k_i is linked to v_j *with* k_j"
- A = "half-edges of v_i connect to half-edges of v_j"
- $|A| = k_i k_j$ (rule of product)
- $|\Omega| = 2m$ (total number of half-edges)
- Probability that a half-edge of node i connects to a half-edge of node j



$$p_{ij} = \frac{|A|}{|\Omega|} = \frac{k_i k_j}{2m}$$

Modularity

- For a given group

$$M_c = \frac{1}{2m} \sum_{v_i, v_j \in V_c} \left(a_{ij} - \frac{k_i k_j}{2m} \right)$$

- If $M_c > 0$, then group c is a potential community

- For the entire network

$$M = \frac{1}{2m} \sum_{v_i, v_j \in V} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(v_i, v_j)$$

- Consider the first term (aggregating by community)

$$\frac{1}{2m} \sum_{v_i, v_j \in V} a_{ij} \delta(v_i, v_j) = \sum_{c=1}^{n_c} \frac{1}{2m} \sum_{v_i, v_j \in V_c} a_{ij}$$

$$= \sum_{c=1}^{n_c} \frac{m_c}{m}$$

— number of edges within group c

— total number of edges in the network

Modularity

- For a given group

$$M_c = \frac{1}{2m} \sum_{v_i, v_j \in V_c} \left(a_{ij} - \frac{k_i k_j}{2m} \right)$$

- If $M_c > 0$, then group c is a potential community

- For the entire network

$$M = \frac{1}{2m} \sum_{v_i, v_j \in V} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(v_i, v_j)$$

- Similarly, for the second term

$$\frac{1}{2m} \sum_{v_i, v_j \in V} \frac{k_i k_j}{2m} \delta(v_i, v_j) = \sum_{c=1}^{n_c} \frac{1}{(2m)^2} \sum_{v_i, v_j \in V_c} k_i k_j = \sum_{c=1}^{n_c} \frac{1}{(2m)^2} k_c^2$$

where k_c is the total degree of the nodes inside group c

Analyzing modularity

- Network modularity can be written as

$$M = \sum_{c=1}^{n_c} \left(\underbrace{\frac{m_c}{m}}_{\text{fraction of edges inside group } c} - \underbrace{\left(\frac{k_c}{2m} \right)^2}_{\text{expected fraction of edges if established at random}} \right) \quad \text{where } k_c = (2+\alpha) m_c$$

- Number of (outgoing) edges joining nodes in group c to the rest of the network

$$m_c^{\text{out}} = \alpha m_c \text{ for some } \alpha \geq 0$$

- If $\alpha = 0$, then group c is disconnected

- Total degree of all nodes inside group c (internal and external degree)

$$k_c = 2m_c + m_c^{\text{out}} = (2+\alpha) m_c$$

Analyzing modularity

- Network modularity can be written as

$$M = \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \left(\frac{k_c}{2m} \right)^2 \right) \quad \text{where } k_c = (2+\alpha) m_c$$

- The subgraph of group c is a module if

$$\frac{m_c}{m} - \left(\frac{k_c}{2m} \right)^2 > 0 \qquad \Rightarrow \qquad \frac{m_c}{m} - \left(\frac{(2+\alpha) m_c}{2m} \right)^2 > 0$$

- Rearranging terms

$$\frac{m_c}{m} > \frac{(2+\alpha)^2 m_c^2}{4m^2} \qquad \Rightarrow \qquad m_c < \frac{4m}{(2+\alpha)^2}$$

- Upper bound on edge density for subgraph to be a module
- If $\alpha = 0$ the disconnected graph is a module for $m_c < m$ (true for $\alpha \geq 0$)
- If $\alpha > 0$ the bound limits the number of internal edges that a module can have!

Analyzing modularity

- Consider $0 < \alpha < 2$

$$m_c < \frac{4m}{(2+\alpha)^2} = m \quad \text{for} \quad \alpha = 0$$

$$m_c < \frac{4m}{(2+\alpha)^2} = \frac{1}{4}m \quad \text{for} \quad \alpha = 2$$

- Total degree of all nodes inside group c for $0 < \alpha < 2$

$$\begin{aligned} 2m_c + m_c^{\text{out}} &= (2+\alpha)m_c \\ &< 4m_c \qquad \Rightarrow \qquad 2m_c > m_c^{\text{out}} \quad \left(\sum_{i \in V_c} k_i^{\text{in}} > \sum_{i \in V_c} k_i^{\text{out}} = m_c^{\text{out}} \right) \end{aligned}$$

- Total degree of internal subgraph larger than its external degree
- Perfect balance between internal and external degree $2m_c = m_c^{\text{out}}$
- Consistent with “weak” definition of community by Radicchi[1]
- If $\alpha = 3$, then $3m_c > m_c^{\text{out}}$, which implies stronger conditions
- The higher α, the “stronger” the community (based on the definition by Radicchi[1])

---

[1] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, Proc. Natl. Acad. Sci. 101, 2658-2663, 2004

Analyzing modularity

- Communities are defined within the modularity framework and according to the weak definition, if for any group c and $\alpha = \alpha_c$

1. Number of edges within group c satisfies

$$m_c < \frac{1}{4}m$$

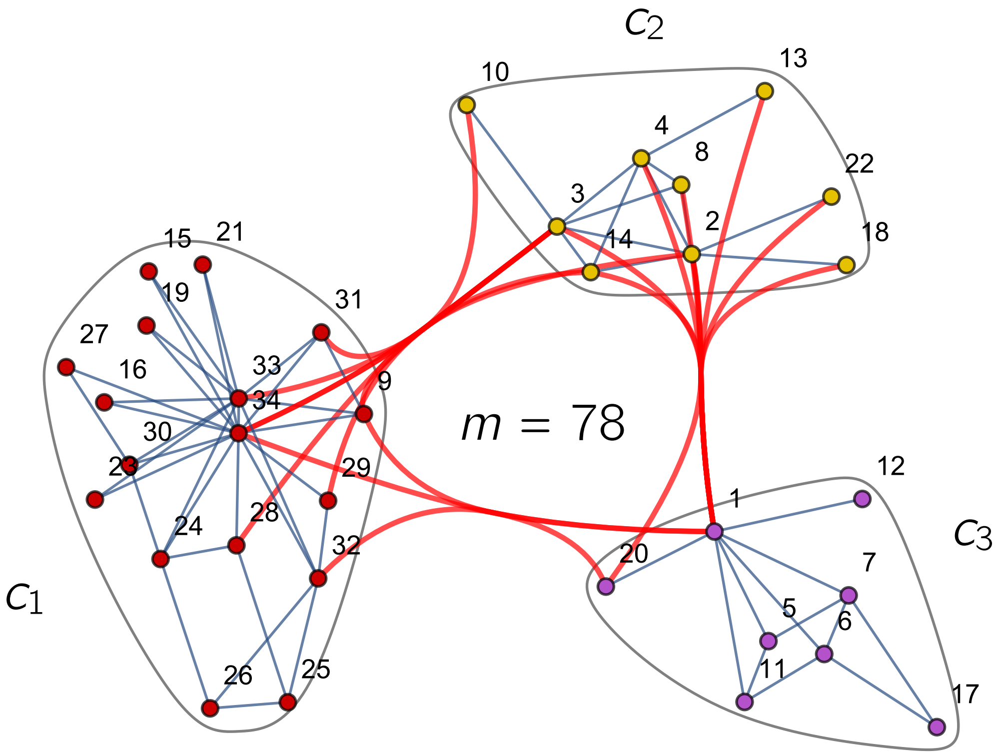
2. Fraction of edges joining nodes in group c to the rest of the network satisfies

$$\alpha < 2$$

Analyzing modularity

External degree: $m_c^{\text{out}} = \alpha_c m_c$ and $\sum_{c=1}^{3} m_c^{\text{out}} = 38$

Total degree: $2m = 2\sum_{c=1}^{3} m_c + \sum_{c=1}^{3} m_c^{\text{out}} = 118 + 38 = 156$



from node (in c)	to node (in c)
$2(c_2)$	$1(c_3)$
$3(c_2)$	$1(c_3)$
$4(c_2)$	$1(c_3)$
$8(c_2)$	$1(c_3)$
$9(c_1)$	$1(c_3)$
$9(c_1)$	$3(c_2)$
$13(c_2)$	$1(c_3)$
$14(c_2)$	$1(c_3)$
$18(c_2)$	$1(c_3)$
$20(c_3)$	$2(c_2)$
$22(c_2)$	$1(c_3)$
$28(c_1)$	$3(c_2)$
$29(c_1)$	$3(c_2)$
$31(c_1)$	$2(c_2)$
$32(c_1)$	$1(c_3)$
$33(c_1)$	$3(c_2)$
$34(c_1)$	$10(c_2)$
$34(c_1)$	$14(c_2)$
$34(c_1)$	$20(c_3)$

How "weak" are the communities?

	c_1	c_2	c_3	total
$2m_c$	68	26	24	118
m_c	34	13	12	-
m_c^{out}	10	16	12	38
α_c	0.29	1.23	1	-

Analyzing modularity

- Back to modularity equation

$$M = \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \left(\frac{k_c}{2m} \right)^2 \right)$$

- Consider a network made up of a single community
- Number of edges within group c satisfies $m_c = m$
- Total degree of the nodes inside the only community $k_1 = 2\text{m}$
- For a single community, $M = 0$
- In general, $M > 0.3$ for community structure to be evident

For any network, maximizing modularity requires $M_{\text{max}} = \max_{n_c, m_c, m} M \leq 1$

Analyzing modularity

- Network with a ring topology between communities
- Note that

$$\sum_{c=1}^{n_c} m_c = m - n_c$$

- Modularity

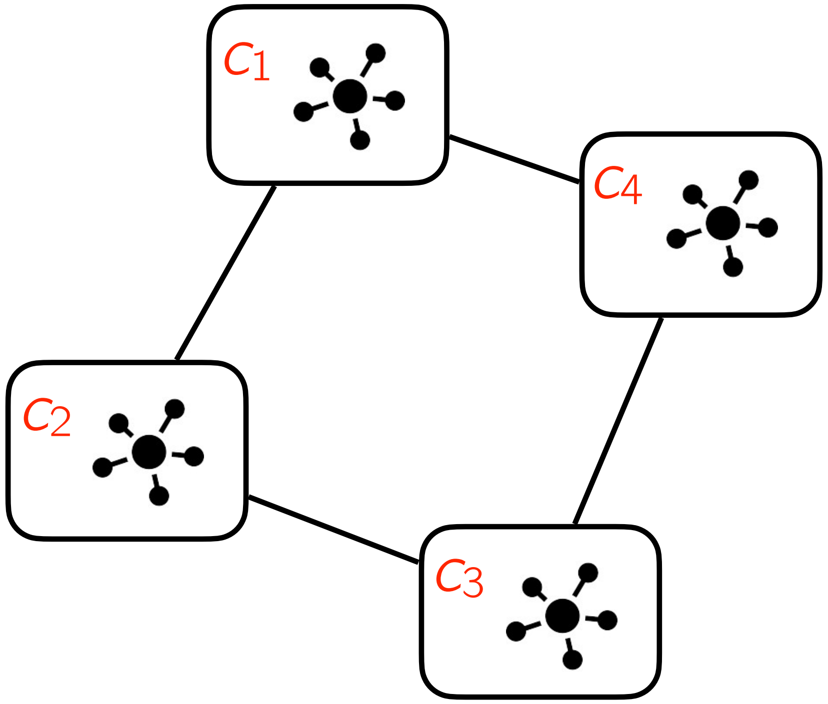
$$M = \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \left(\frac{k_c}{2m} \right)^2 \right)$$

$$= \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \left(\frac{2m_c + 2}{2m} \right)^2 \right)$$

attains its maximum value when

$$m_c = \frac{m}{n_c} - 1 \qquad \text{(Lagrange multipliers)}$$

Ring topology: minimum number of edges between communities



$$k_c = 2m_c + m_c^{\text{out}} = 2m_c + 2$$

M is maximized when all modules have the same number of edges!

Analyzing modularity

- Lagrange multipliers (constrained optimization)
- Maximize

$$f(m_c) = \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \left(\frac{2m_c + 2}{2m} \right)^2 \right)$$

subject to

$$g(m_c) = \sum_{c=1}^{n_c} m_c - m + n_c = 0$$

- Define the Lagrangian

$$\begin{aligned} \mathcal{L}(m_c, \lambda) &= f(m_c) - \lambda g(m_c) \\ &= \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \frac{(2m_c + 2)^2}{2m} \right) - \lambda \left(\sum_{c=1}^{n_c} m_c - m + n_c \right) \end{aligned}$$

- Solve

$$\nabla_{m_c,\lambda} \mathcal{L}(m_c, \lambda) = 0 \iff \begin{cases} \nabla_{m_c} f(m_c) = \lambda \nabla_{m_c} g(m_c) \\ g(m_c) = 0 \end{cases}$$

Analyzing modularity

- To satisfy $\nabla_{m_c} f(m_c) = \lambda \nabla_{m_c} g(m_c)$
- For any group c

$$\frac{\partial}{\partial m_c} \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \left(\frac{2m_c + 2}{2m} \right)^2 \right) = \frac{1}{m} + \frac{2m_c + 1}{2m^2}$$
$$= \lambda \left(\frac{\partial}{\partial m_c} \sum_{c=1}^{n_c} m_c - m + n_c \right) = \lambda$$

- So

$$m_c = m_{c'} \text{ for any } c, c' \in \{1, ..., n_c\}$$

- All groups have the same number of internal edges

- Moreover, since $g(m_c) = 0$

$$\sum_{c=1}^{n_c} m_c - n + n_c = n_c m_c - m + n_c = 0 \quad \Longrightarrow \quad m_c = \frac{m - n_c}{n_c}$$
$$= \frac{m}{n_c} - 1$$

Analyzing modularity

- For $m_c = \frac{m}{n_c} - 1$

$$M_{\text{max}}(n_c, m) = \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \left(\frac{2m_c + 2}{2m} \right)^2 \right)$$

$$= \sum_{c=1}^{n_c} \left(\frac{\frac{m}{n_c} - 1}{m} - \left(\frac{2(\frac{m}{n_c} - 1) + 2}{2m} \right)^2 \right)$$

$$= n_c \left(\frac{1}{n_c} - \frac{1}{m} - \frac{1}{n_c^2} \right)$$

$$= 1 - \frac{n_c}{m} - \frac{1}{n_c}$$

- Since

$$\frac{dM_{\text{max}}(n_c, m)}{dn_c} = -\frac{1}{m} + \frac{1}{n_c^2} = 0 \quad \text{when} \quad n_c = \sqrt{m}$$

- Let n_c^* be the closest integer to $\sqrt{m}$

$$M_{\text{max}}(m) = M_{\text{max}}(n_c^*, m) = 1 - \frac{2}{\sqrt{m}} \rightarrow 1 \text{ as } m \rightarrow \infty$$

Intrinsic scale of order $\sqrt{m}$

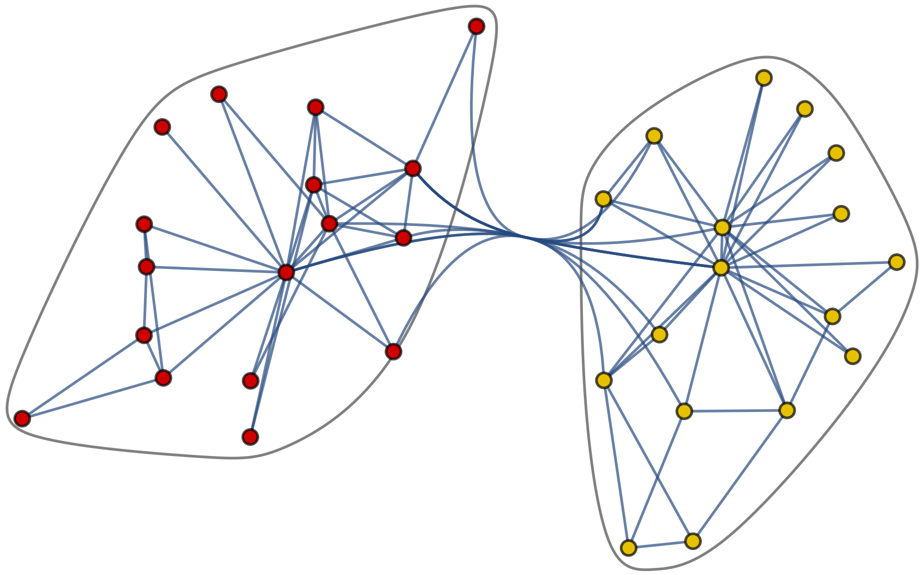
Remarks

- Maximum value of M depends on number of edges m
- Maximized when all modules have the same number of edges
- Number of edges of each module

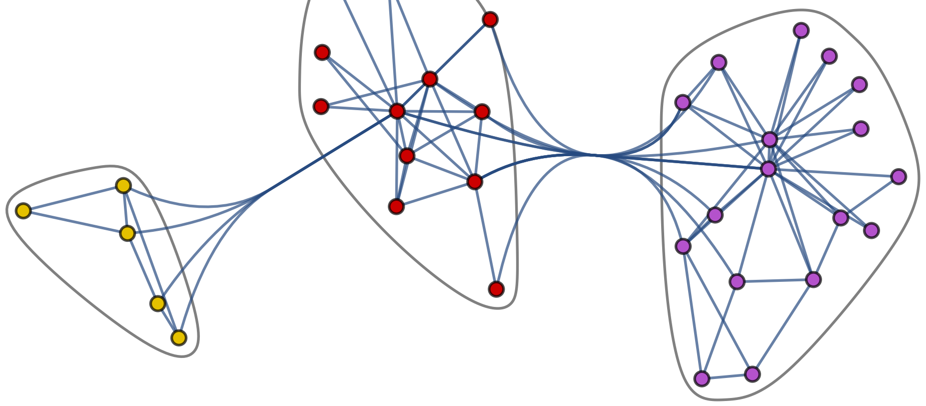
$$\frac{m}{n_c^*} - 1 = \frac{m}{\sqrt{m}} - 1 = \sqrt{m} - 1$$

- Not necessarily the same number of nodes!
- Modularity does not depend on the distribution of nodes among modules
- Exact solutions to maximizing modularity are computationally expensive
- Approximations
  - Greedy algorithm
  - InfoMap
  - Louvain

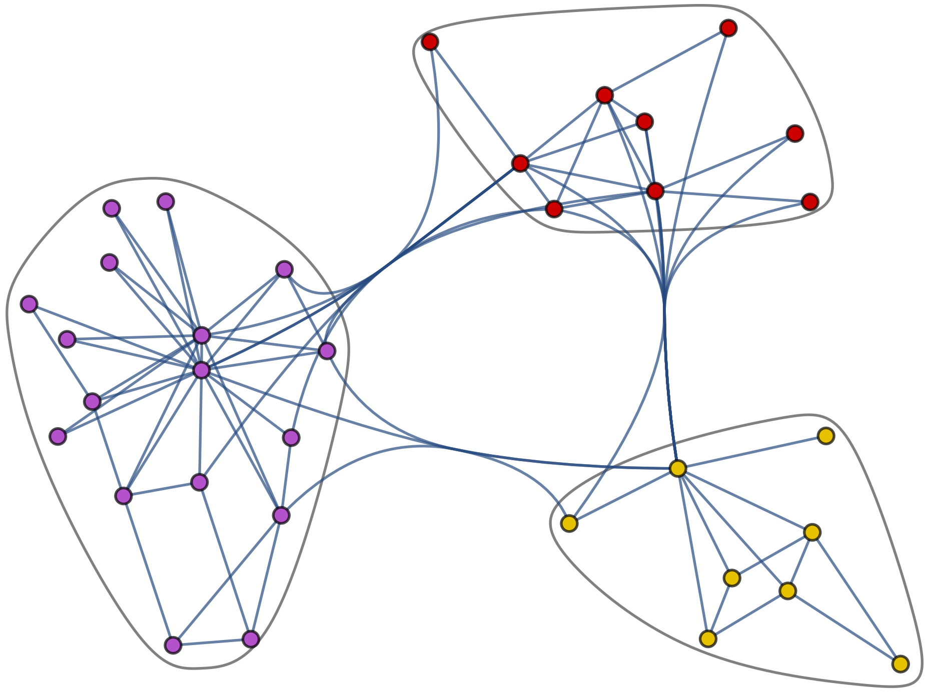
Maximizing modularity (karate club)



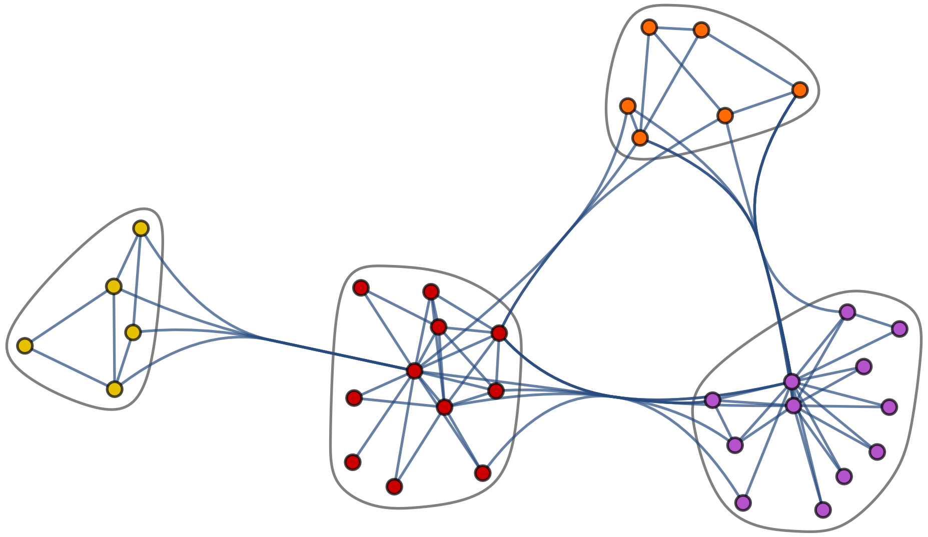
Two communities ($M = 0.37$)



InfoMap ($M = 0.40$)



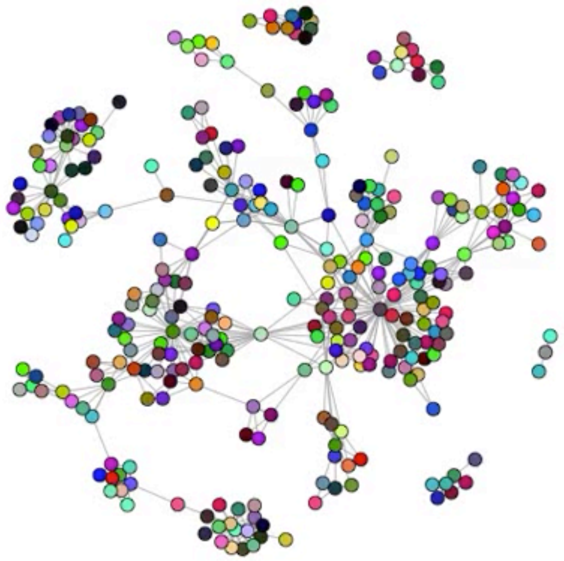
Greedy algorithm ($M = 0.38$)

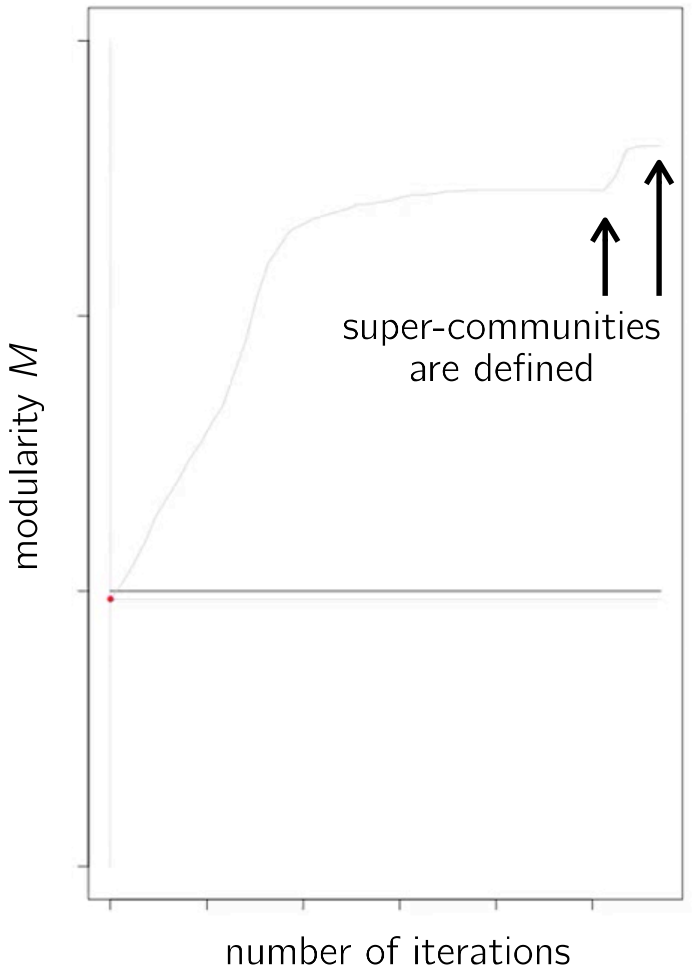


Louvain ($M = 0.42$) - optimal

Louvain community detection

- Assign each node to a unique community
- Pick a nodes at random
- Evaluate placing the node in the community of each neighbor
- Select community that yields the greatest positive modularity change
- Continue picking nodes at random until minimum increase in modularity
- Define super-communities
- Repeat previous steps

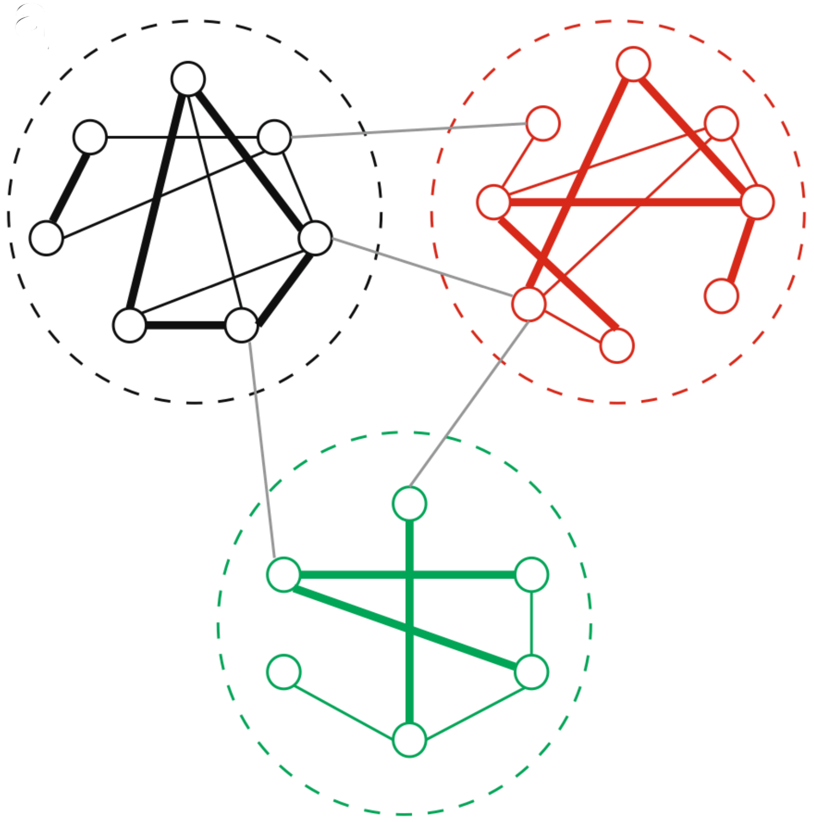




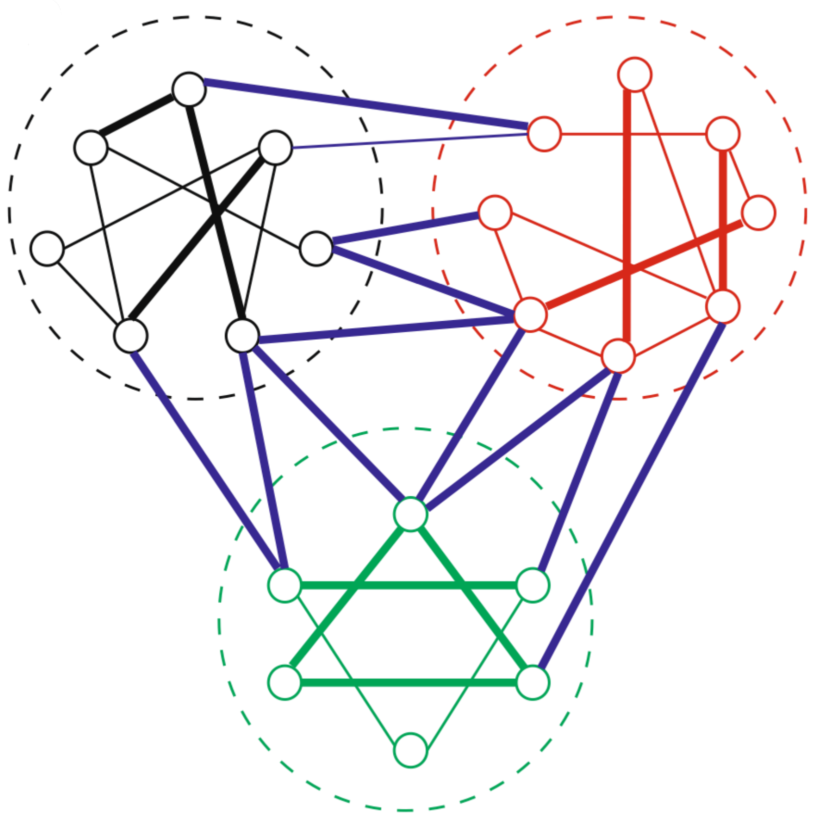
http://youtu.be/dGa-TXpoPz8

Community structure: applications

- Global events trigger viral information (spreading of viral information)
- Distinct diffusion patterns with respect to community structure
- Viral information easily crosses community boundaries
- **Example:** email data from Enron in the wake of bankruptcy
- **Idea:** detect events by monitoring intra- and inter-community communications



no event: most communication takes place within communities

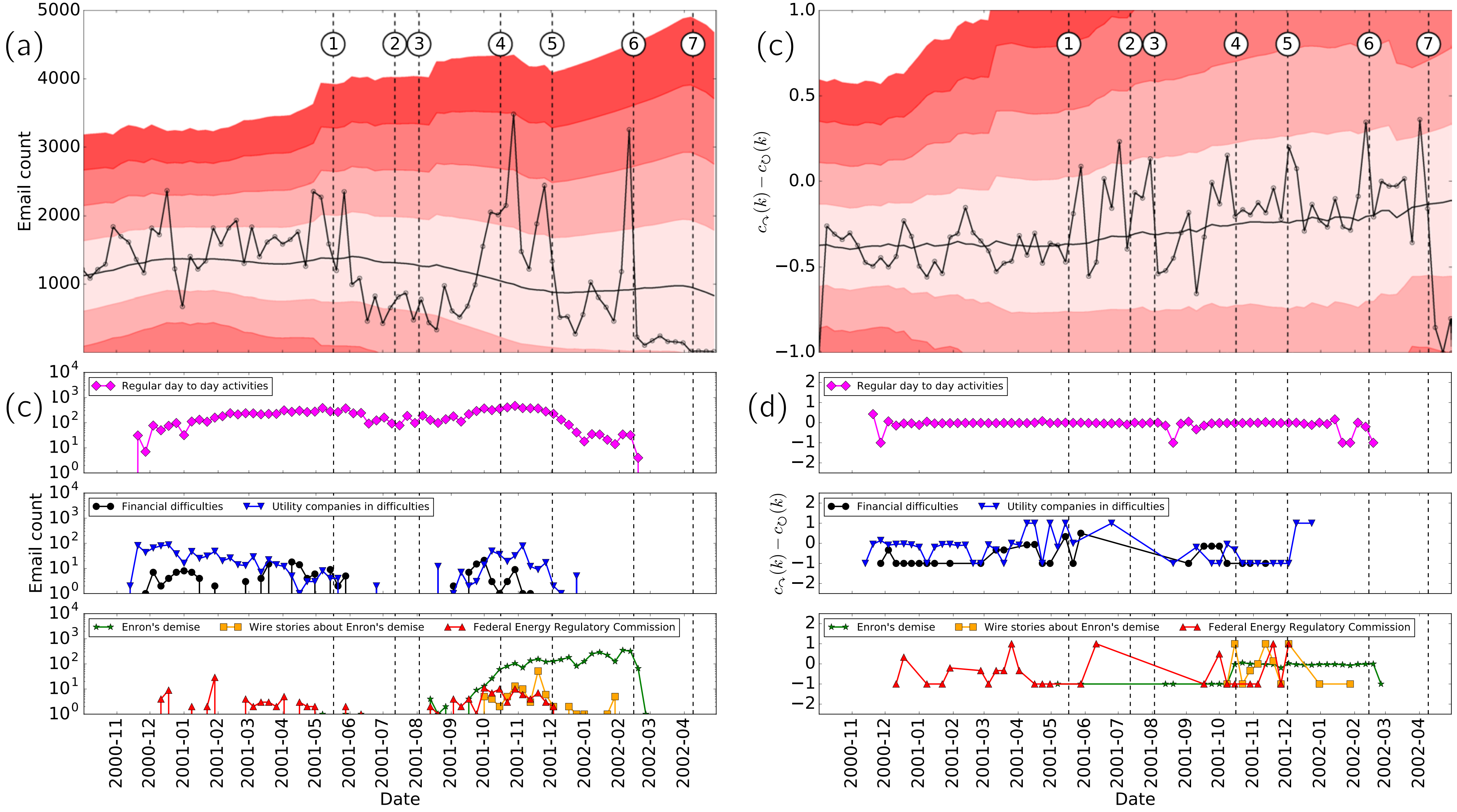


global event: more communication takes place across communities

P. Moriano, J. Finke, Y-Y., Ahn. Community-Based event Detection in temporal Networks. Scientific Reports, 2019

Community structure: applications

- Inter- and intra-community link ratios
- Community structure based on InfoMap



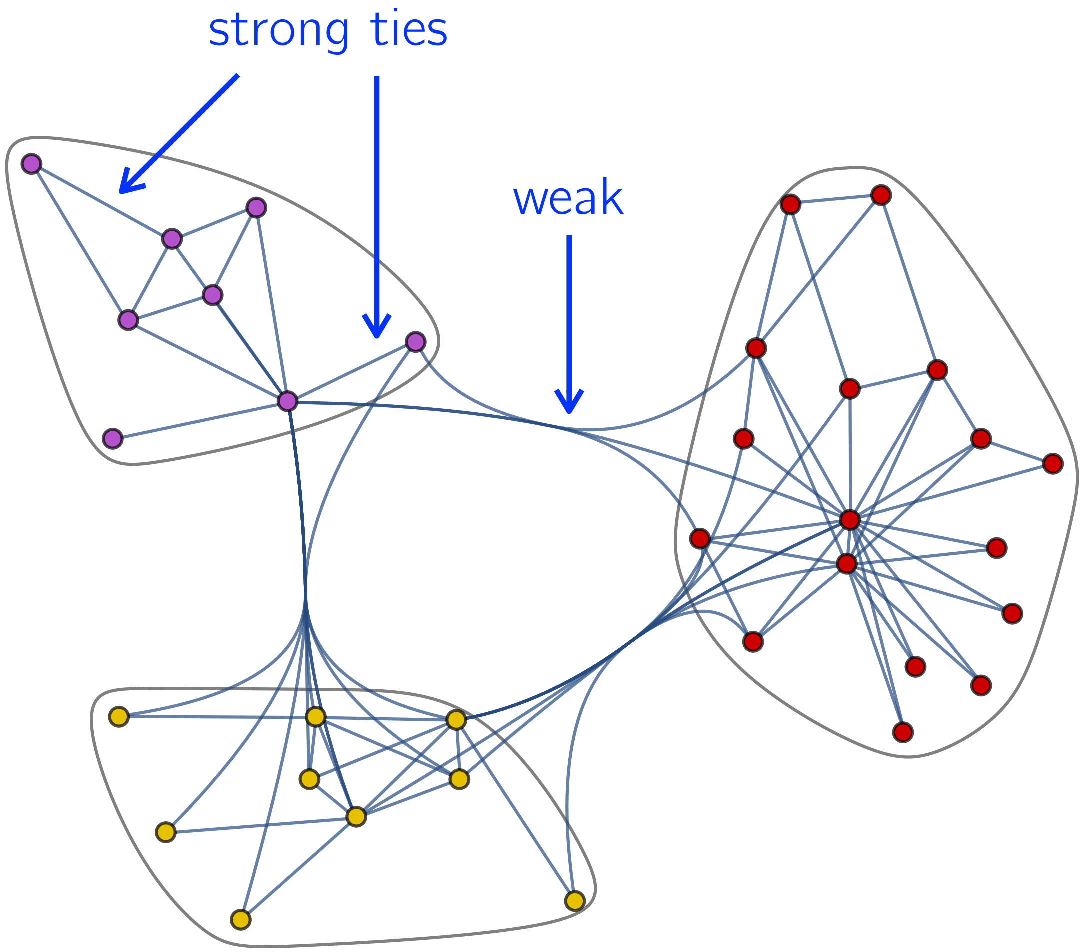
Time series of Enron events. (a) Time series of the number of emails. (b) Time series of the difference between the inter- and intra-community link ratios. (c) Time series of the number of emails classified by topics. (d) Time series of the difference between the inter- and intra-community link ratio classified by topics.

P. Moriano, J. Finke, Y-Y., Ahn. Community-Based event Detection in temporal Networks. Scientific Reports, 2019

Community structure: applications

Dunbar's number:
Theoretical cognitive limit to the number of people with whom one can maintain strong ties (stable social relationships)

Granovetter's theory:
Suggests that weak ties define "diffusion of influence and information, and mobility opportunities."

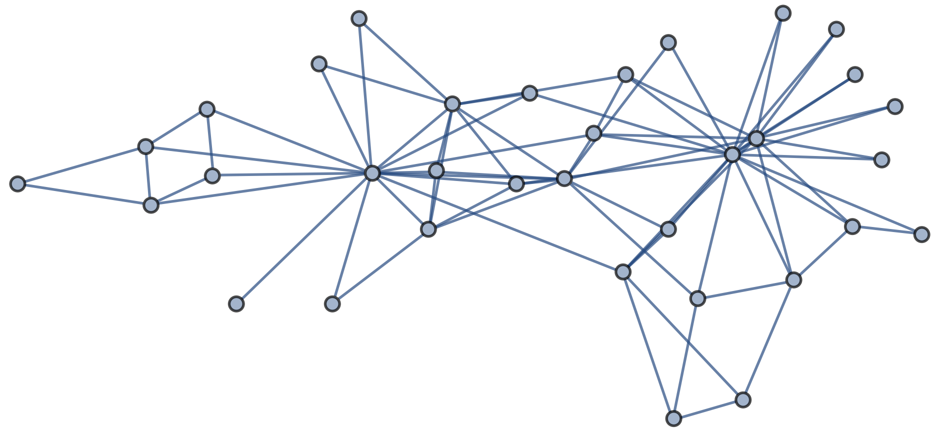


Dunbar's number commonly cited as 150

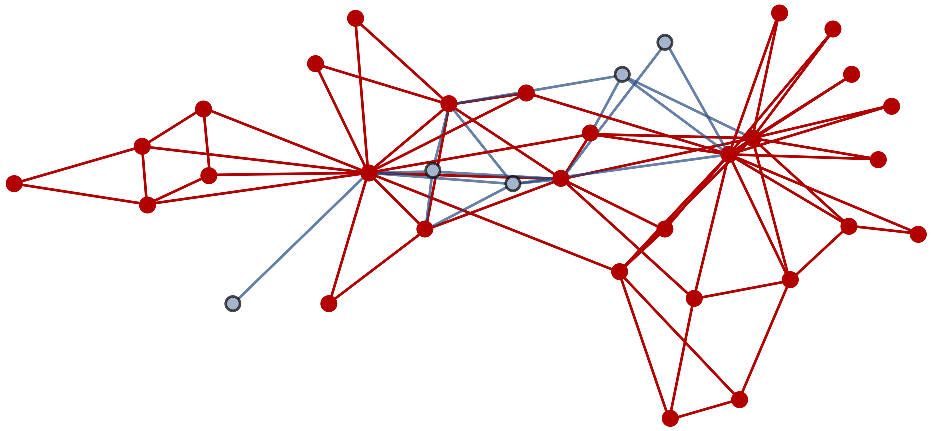
Alternative methods to compute communities?

Clique percolation method (CPM)

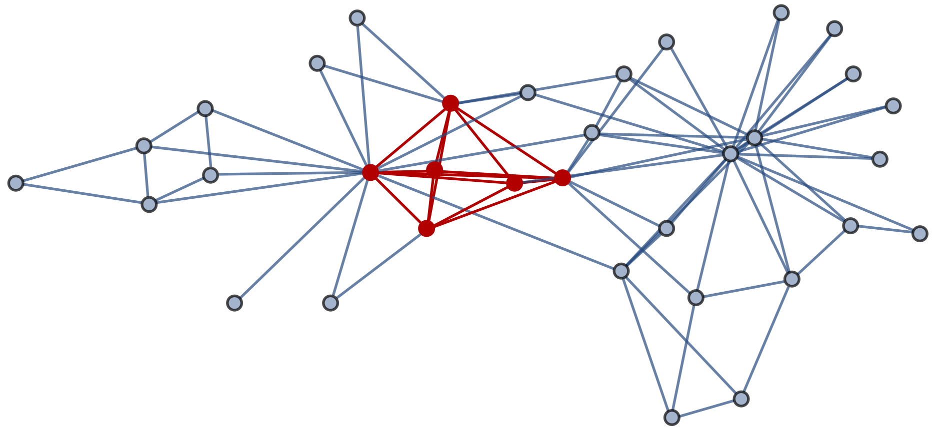
- **Idea:** identify all cliques (fully connected subgraphs) of size *k*
- **A community:** overlapping sets of cliques of size *k*-1



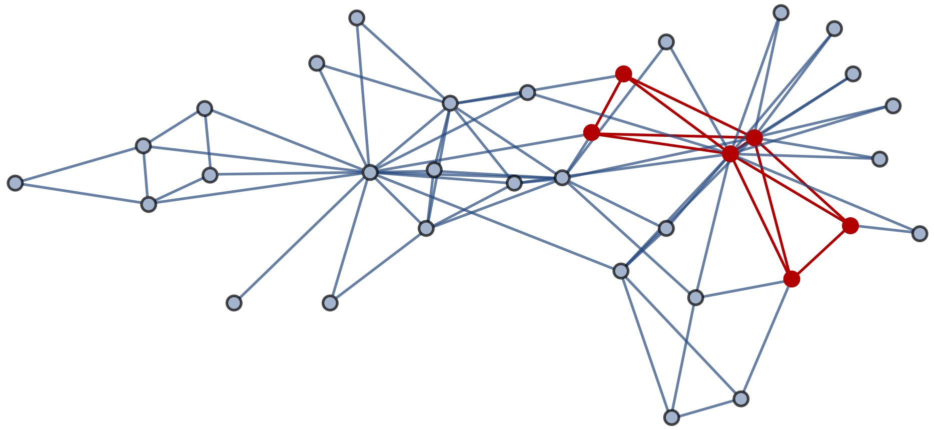
Zachary Karate Club



k = 3 (21 total)

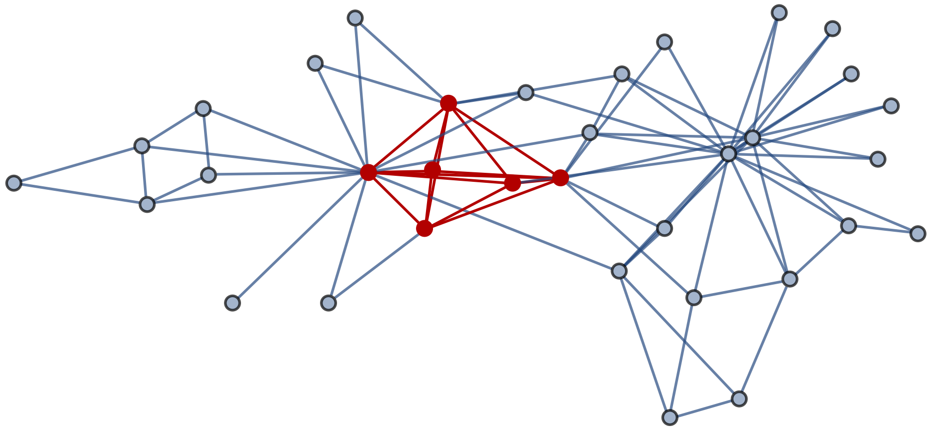


k = 4 (2 total)

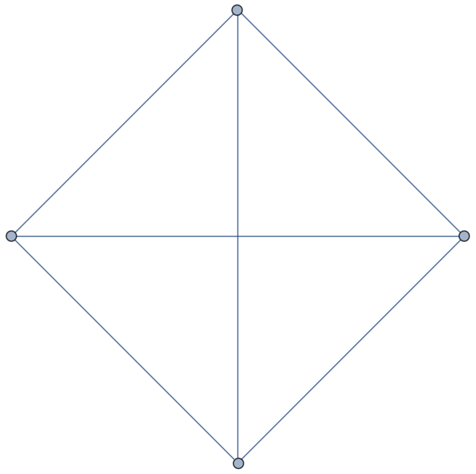


k = 5 (2 total)

Cliques for $k = 4$ and $k = 5$



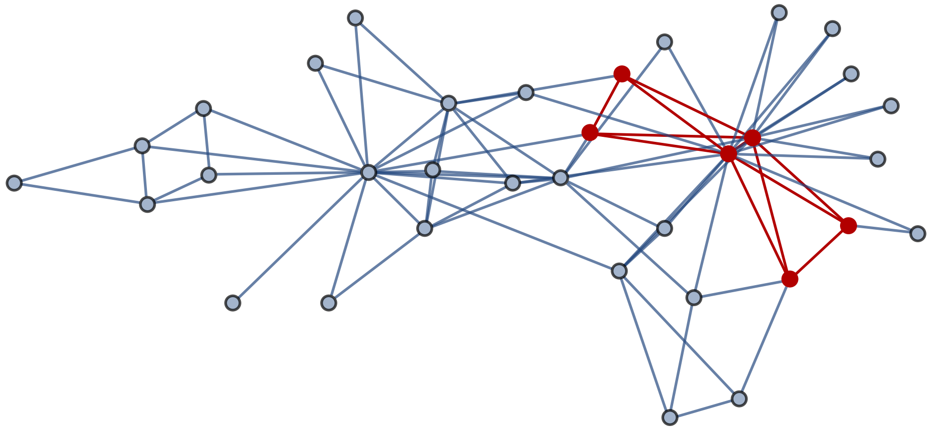
$k = 4$ (2 total)



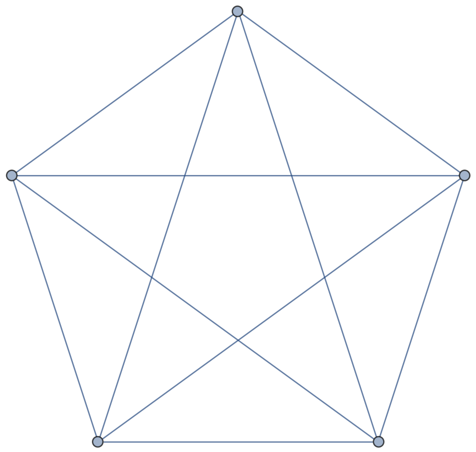
$c_1 = \{v_{24}, v_{30}, v_{33}, v_{34}\}$

$c_2 = \{v_9, v_{31}, v_{33}, v_{34}\}$

No two cliques (of size $k = 4$)
share k-1 members



$k = 5$ (2 total)



$c_1 = \{v_1, v_2, v_3, v_4, v_{14}\}$

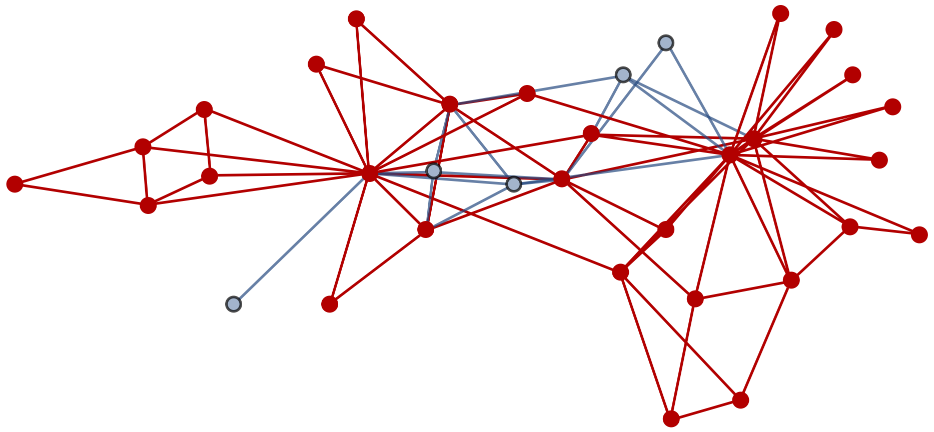
$c_2 = \{v_1, v_2, v_3, v_4, v_8\}$

Nodes v_1 to v_4 belong to both cliques

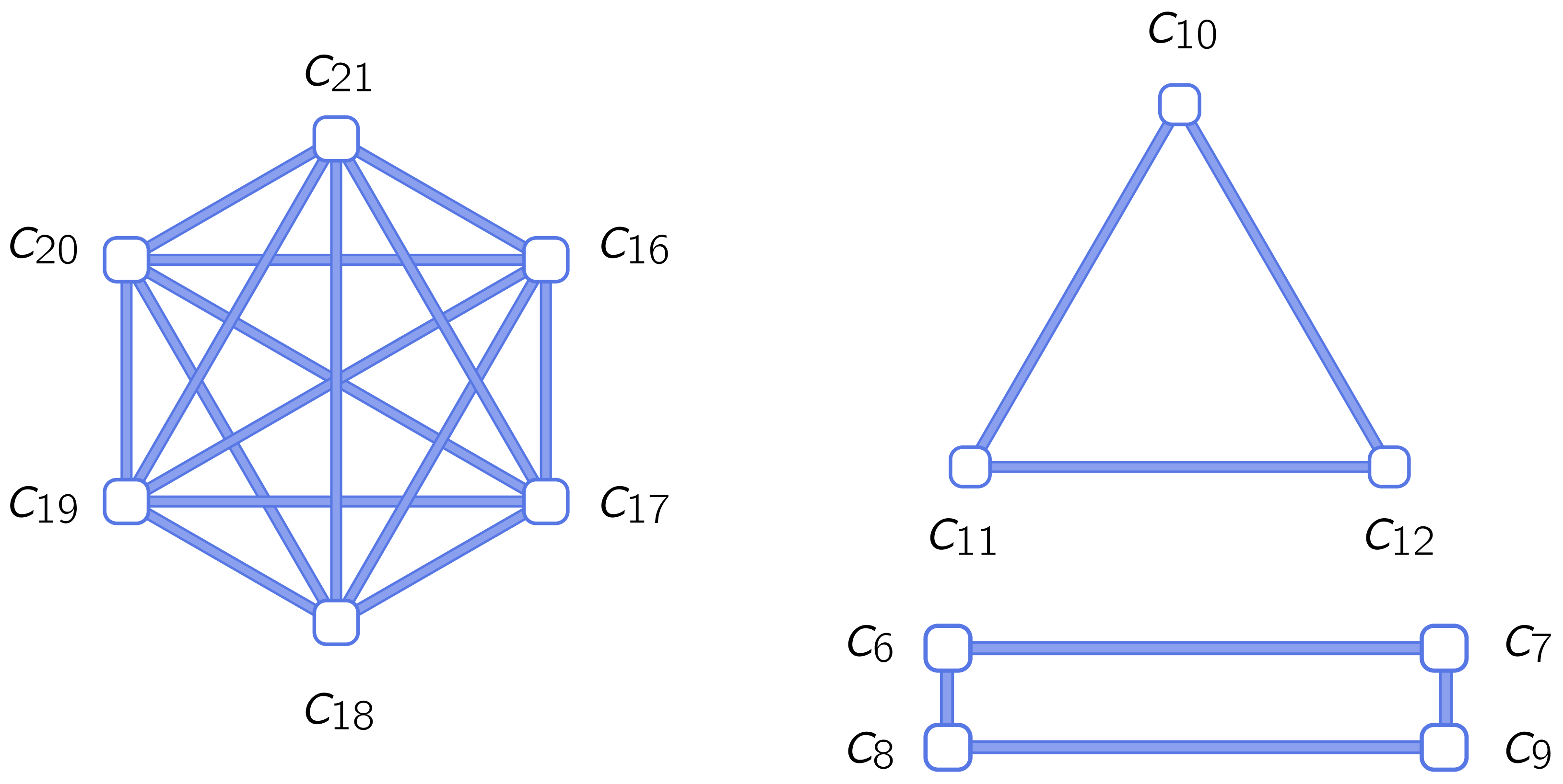
Cliques for $k = 3$

list of cliques:

$C_1 =$ ~~$\{v_{26}, v_{25}, v_{32}\}$~~
$C_2 =$ ~~$\{v_6, v_7, v_3\}$~~
$C_3 =$ ~~$\{v_3, v_9, v_{33}\}$~~
$C_4 =$ ~~$\{v_1, v_4, v_{13}\}$~~
$C_5 =$ ~~$\{v_1, v_3, v_9\}$~~
$C_6 = \{v_1, v_6, v_{11}\}$
$C_7 = \{v_1, v_6, v_7\}$
$C_8 = \{v_1, v_5, v_{11}\}$
$C_9 = \{v_1, v_5, v_7\}$
$C_{10} = \{v_1, v_2, v_{22}\}$
$C_{11} = \{v_1, v_2, v_{20}\}$
$C_{12} = \{v_1, v_2, v_{18}\}$
$C_{13} =$ ~~$\{v_{30}, v_{27}, v_{34}\}$~~
$C_{14} =$ ~~$\{v_{29}, v_{32}, v_{34}\}$~~
$C_{15} =$ ~~$\{v_{24}, v_{28}, v_{34}\}$~~
$C_{16} = \{v_{33}, v_{23}, v_{34}\}$
$C_{17} = \{v_{33}, v_{21}, v_{34}\}$
$C_{18} = \{v_{33}, v_{19}, v_{34}\}$
$C_{19} = \{v_{33}, v_{16}, v_{34}\}$
$C_{20} = \{v_{33}, v_{15}, v_{34}\}$
$C_{21} = \{v_{33}, v_{32}, v_{34}\}$



Overlapping sets of cliques:



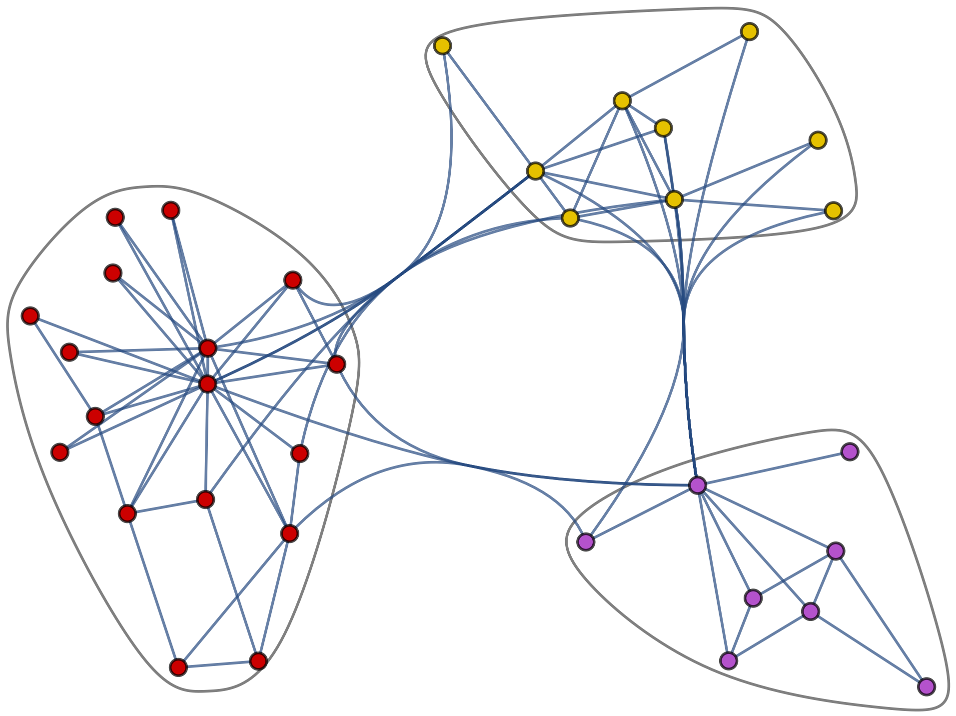
Remarks

- CPM suitable for dense networks
- For a network with n nodes and m edges:

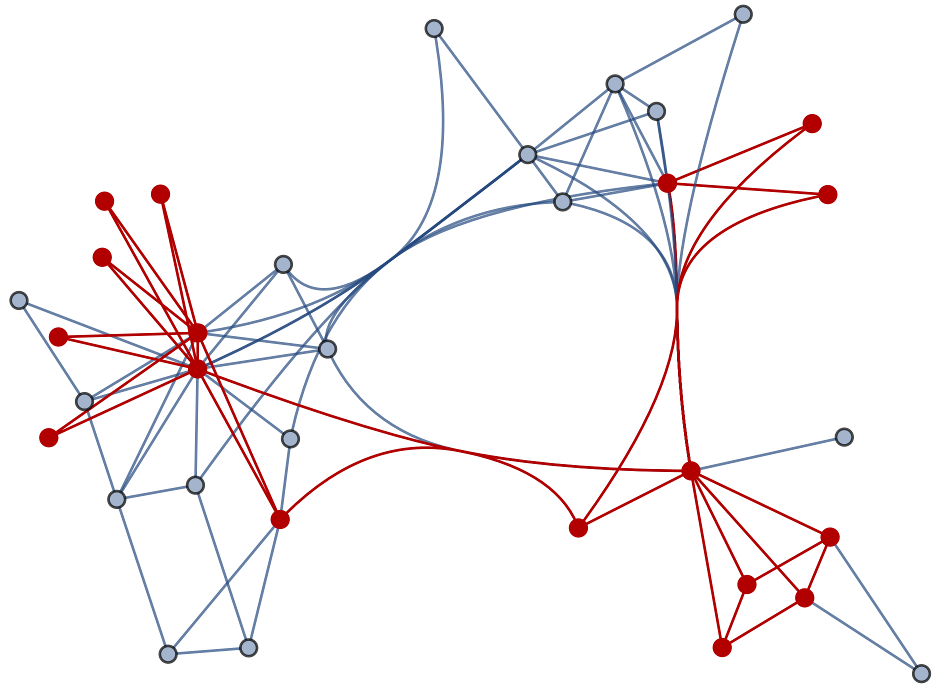
$$m = O(n^2)$$

- Typically, small values of k $(2 < k < 7)$
- CPM for weighted networks
    - k-cliques with intensity larger than a threshold are included into a community
- Aims to find specific localized structure (pattern matching)
- High-quality communities (but for a small fraction of the network)
- Highly overlapping communities can have many more external than internal connections

Karate club (greedy algorithm)

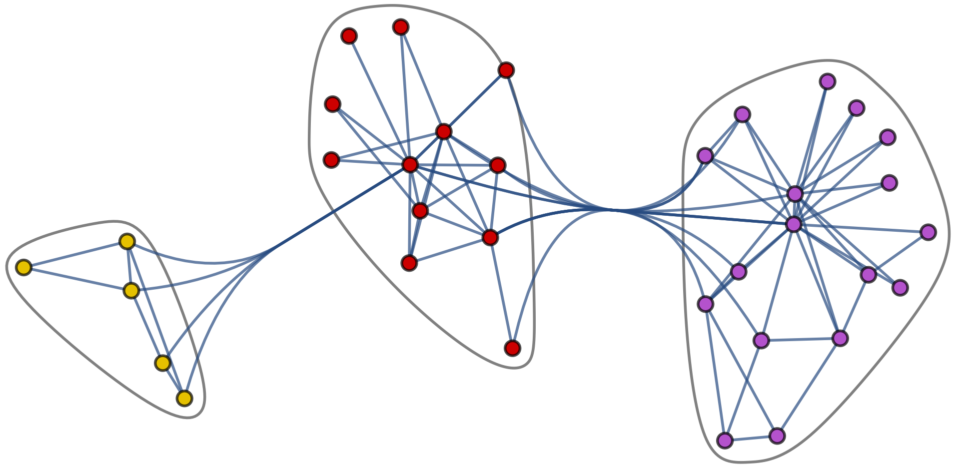


modularity based

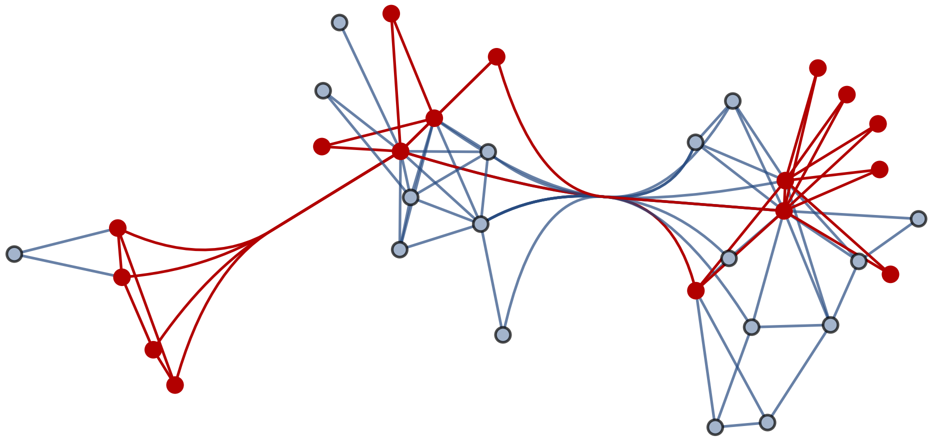


non-modularity based

Karate club (InfoMap)

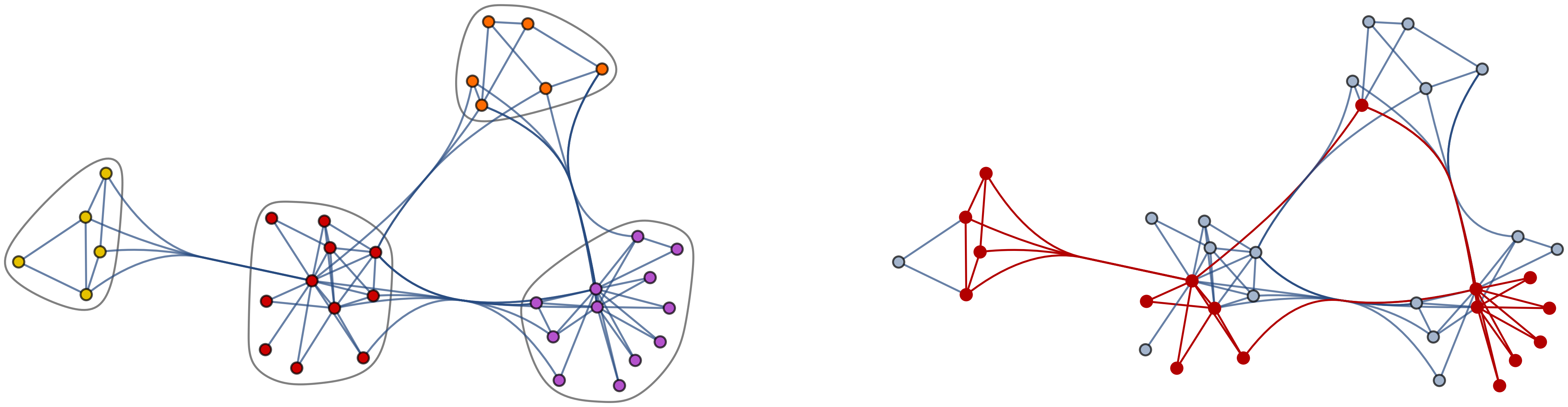


modularity based



non-modularity based

Karate club (Louvain)



modularity based

non-modularity based

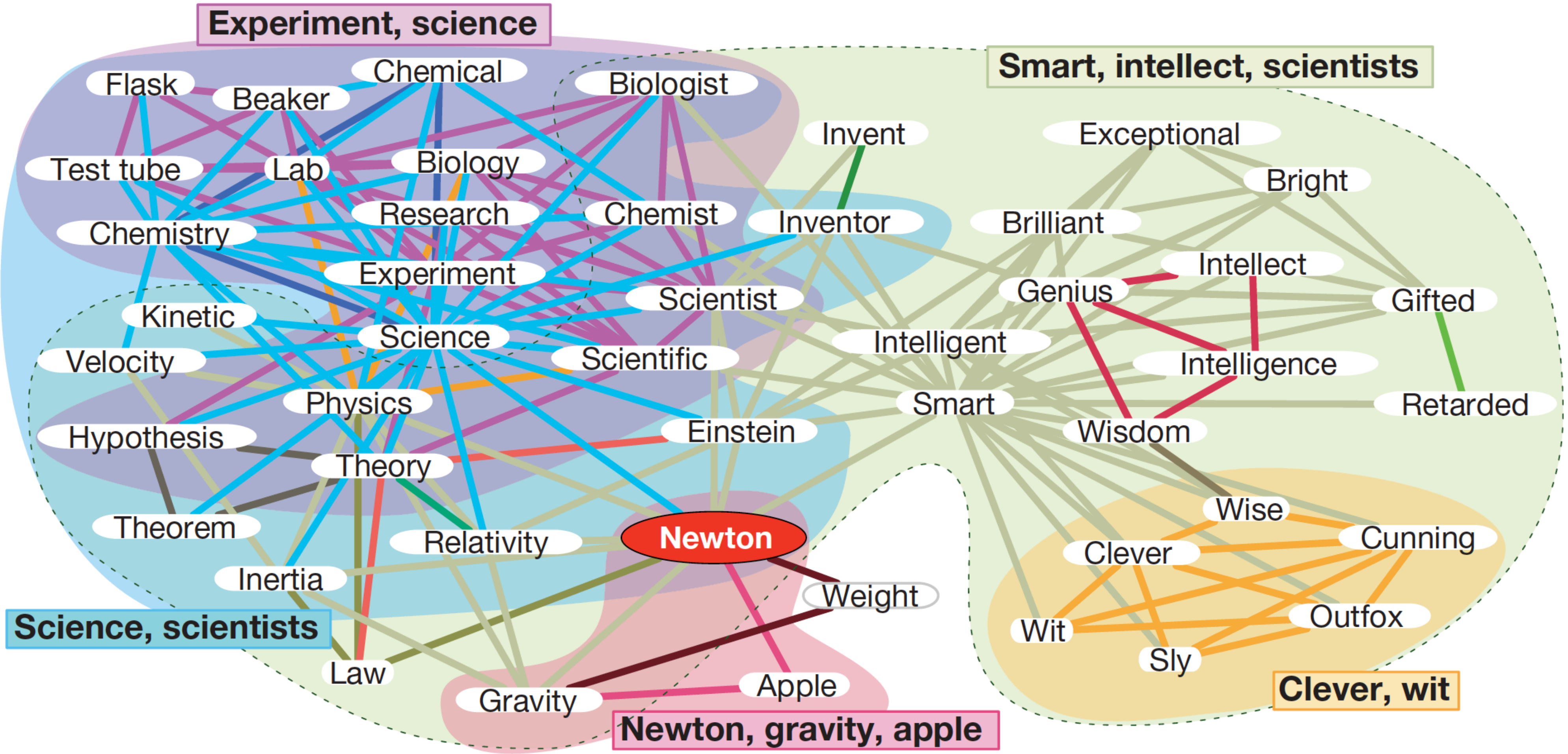
Problem

As we have seen, one needs to be careful when defining community structures based on modularity. Complete the following steps:

1. From SNAP, download the social network facebook_combined.txt
2. Create the corresponding graph using networkx
3. Compute the division of the network that maximizes modularity (use greedy modularity maximization)
4. Verify that the resulting communities are defined within the modularity framework and according to the weak definition
5. Generate a random network based on the degree preserving model (where nodes have similar degrees as the original network)
6. Compare the formation of communities and modularity values the for both networks
7. Define communities based on CPM. Do the formation of communities from modularity-based and non-modularity based methods overlap?

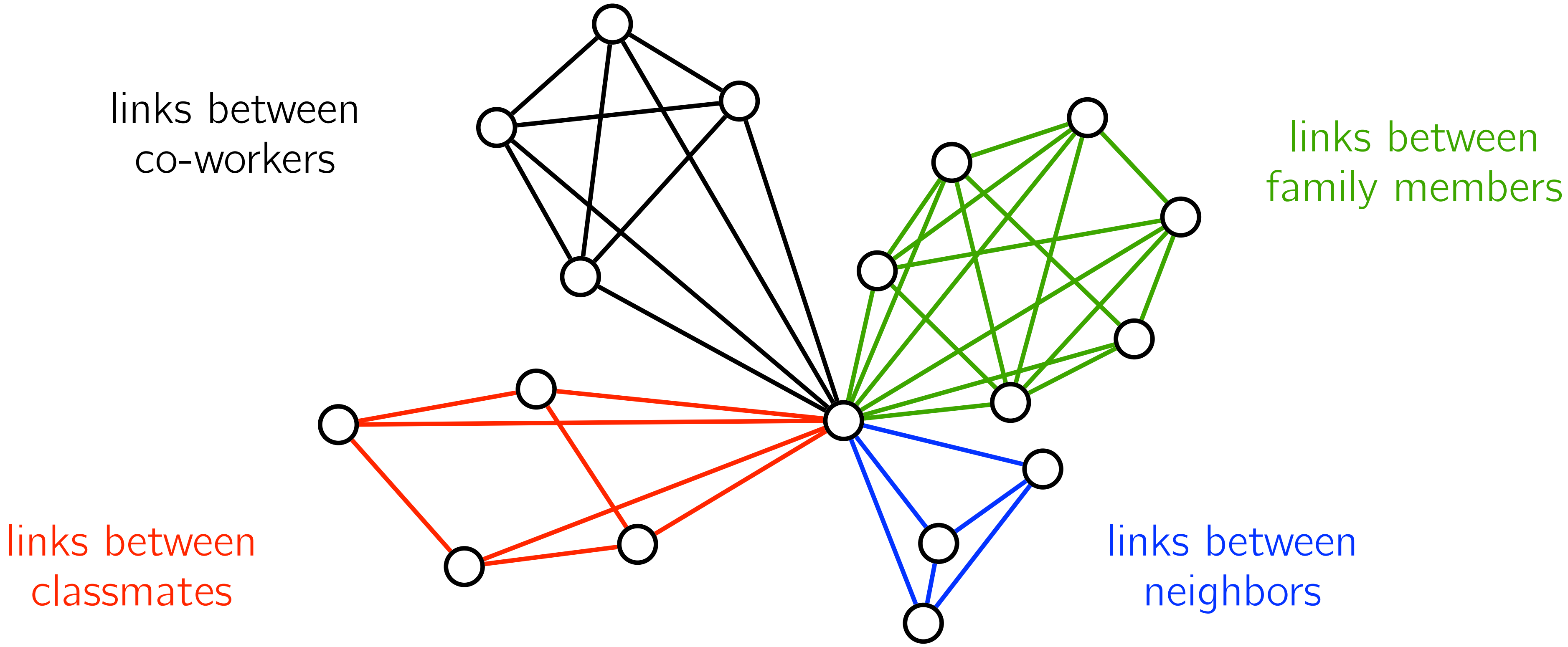
Overlapping community detection

- **Idea:** Allow nodes to simultaneously belong to several groups
- Evidence for hierarchical organization of large network
- Contain hierarchically organized community structures
- **Example:** Commonly associated English words



Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann, "Link communities reveal multiscale complexity in networks," Nature, vol. 466, no. 5, August, 2010.

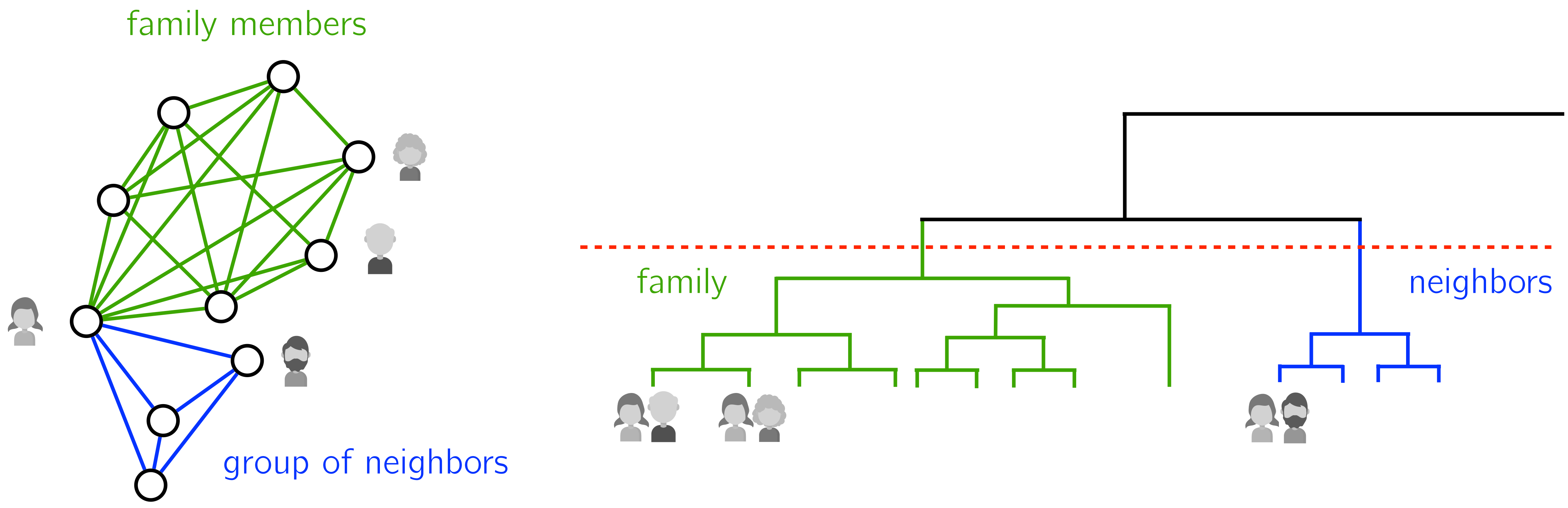
Overlapping community detection



- An individual node sees the communities it belongs to
- Nodes tend to belong to multiple groups
- Links exist for one dominant reason (social networks)
  - same family
  - work together
  - share common interests
  - live in the same neighborhood

Overlapping community detection

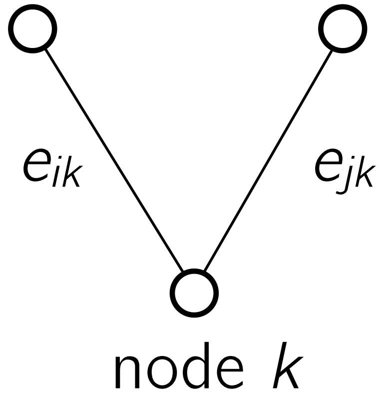
- Community: set of closely interrelated links
- Places each link in a single context (community)
- Each leaf of a dendrogram is a link from the original network
- Extract link communities at multiple levels by cutting at various thresholds
- Each brach represents link partition



Nodes can occupy multiple leaves of the link dendrogram

Link partitioning

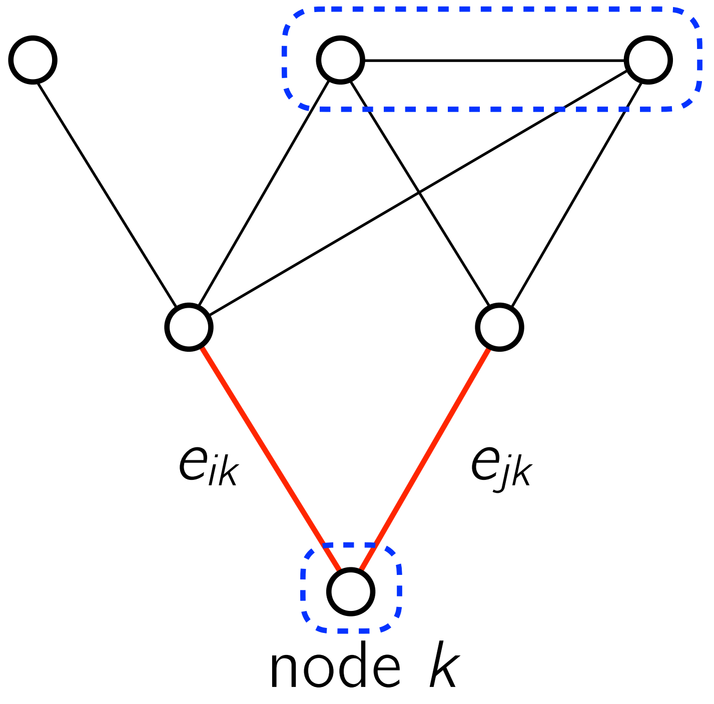
- **Idea:**
  - partition link instead of nodes
  - discover groups of links with similar characteristics
- **Overlapping node:** a node with edges belonging to more than one link partition
- **Hierarchical link clustering:**
  - Given links e_{ik} and e_{jk} (incident on node k)
  - N_i : Neighborhood of node i (itself included)
  - Edge similarity (Jaccard) between two edges:



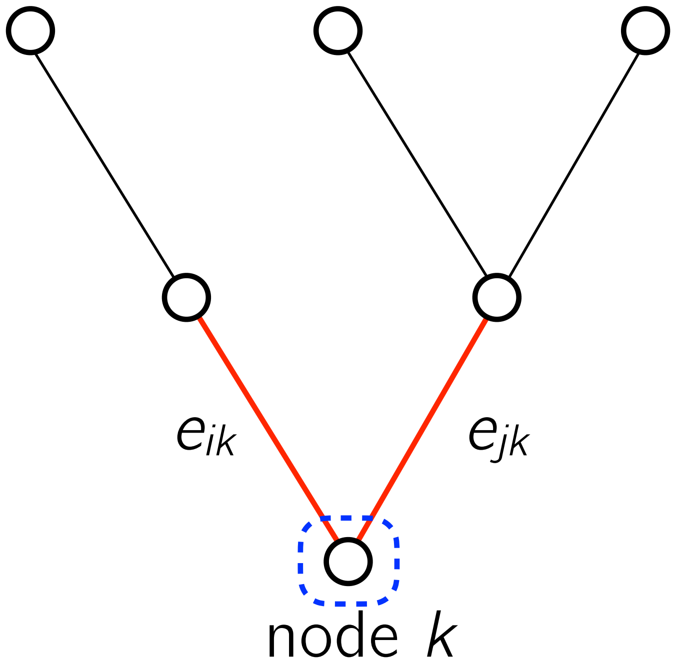
$$S(e_{ik}, e_{jk}) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

- **Dendrogram:** cut at some threshold

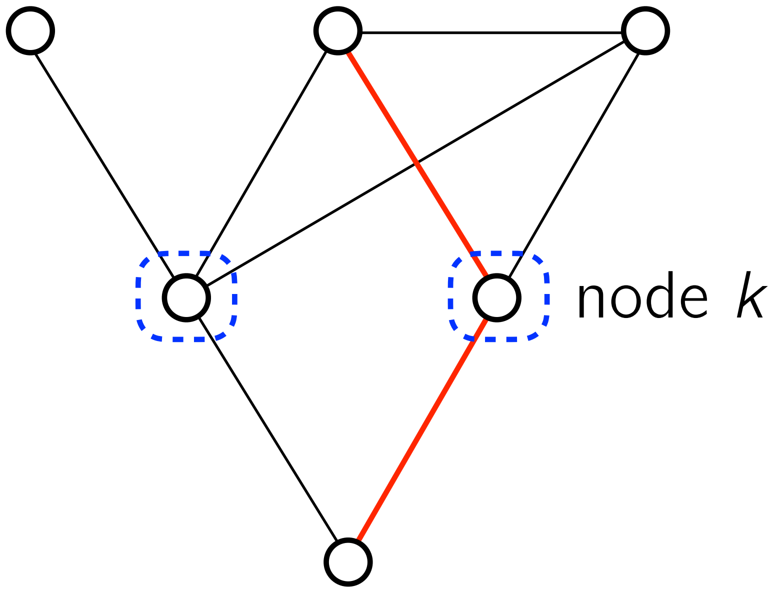
Link partitioning



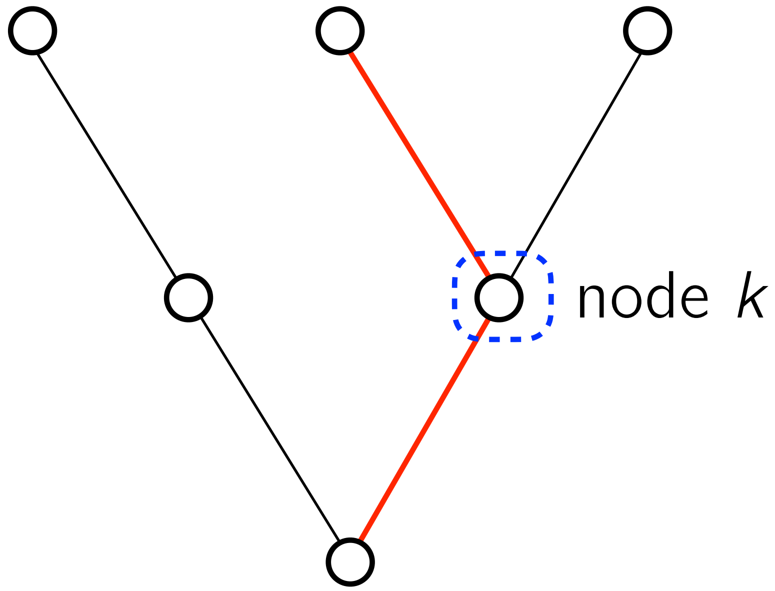
$$S(e_{ik}, e_{jk}) = S(e_{jk}, e_{ik}) = \frac{1}{2}$$



$$S(e_{ik}, e_{jk}) = S(e_{jk}, e_{ik}) = \frac{1}{6}$$

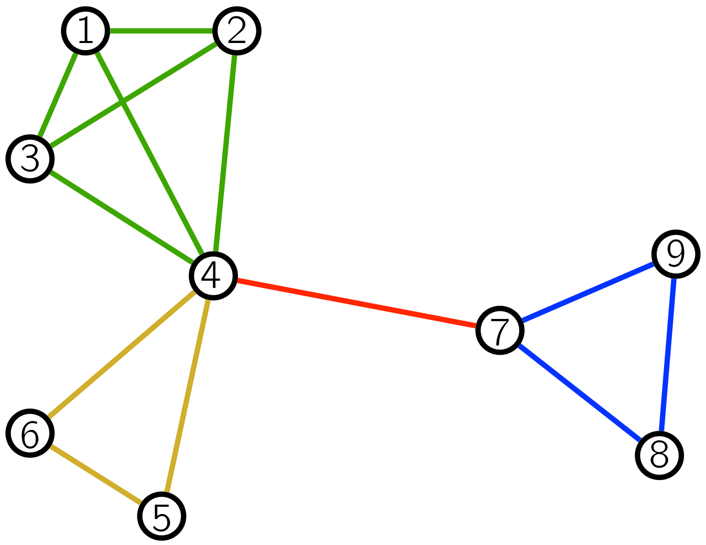


$$S(e_{ik}, e_{jk}) = S(e_{jk}, e_{ik}) = \frac{2}{5}$$

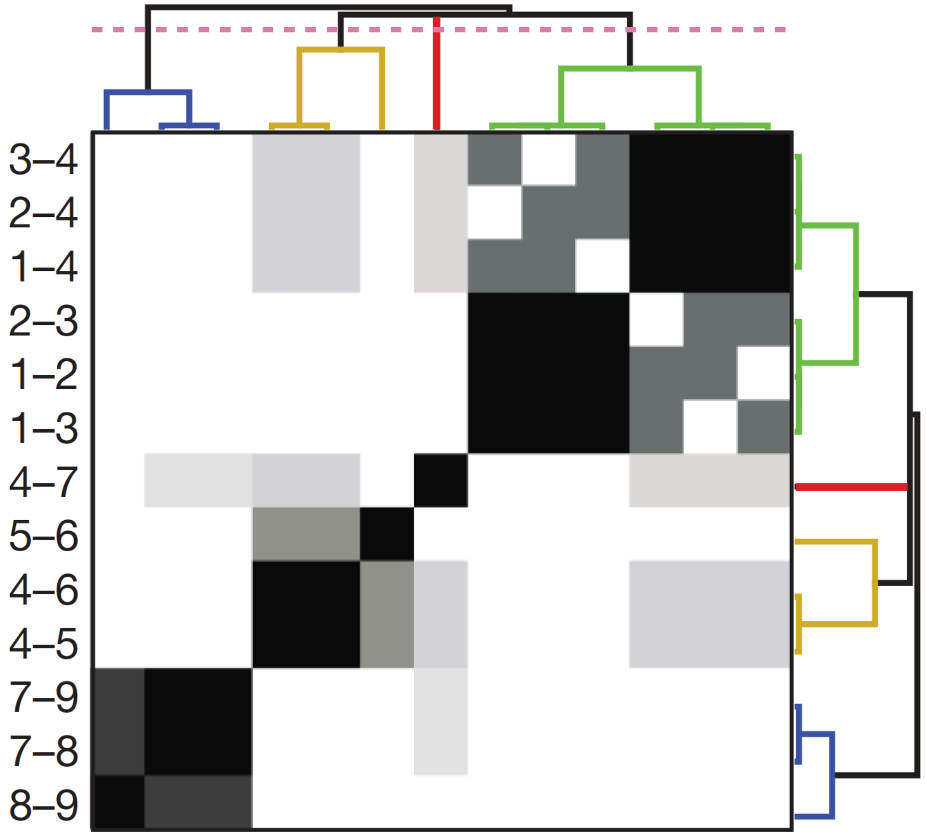


$$S(e_{ik}, e_{jk}) = S(e_{jk}, e_{ik}) = \frac{1}{4}$$

Link communities

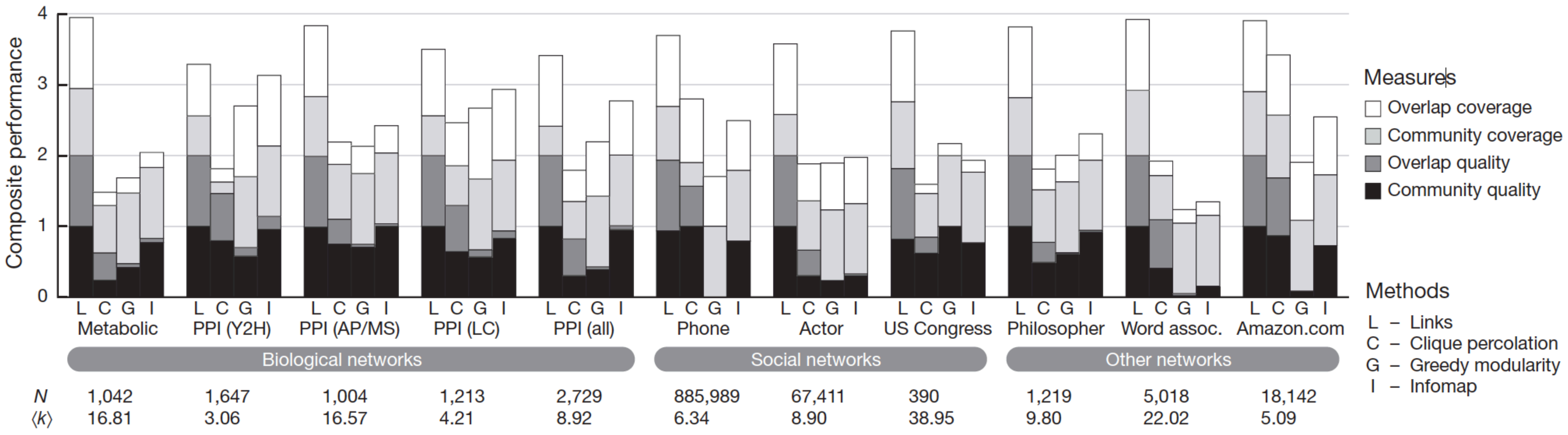


network with overlapping communities



link similarity matrix

Performance of various algorithms

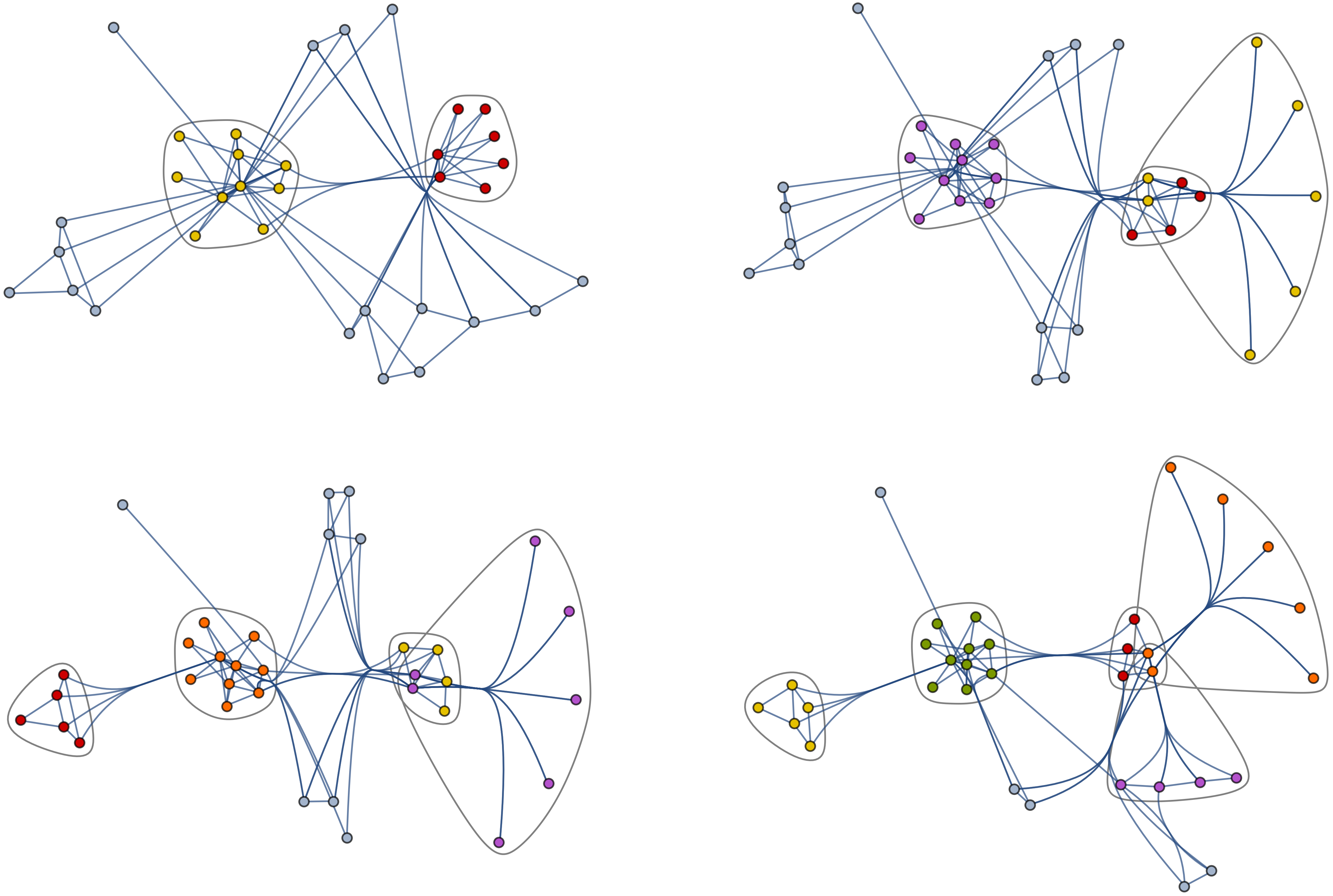


Comparison of the output of each algorithm with the metadata:

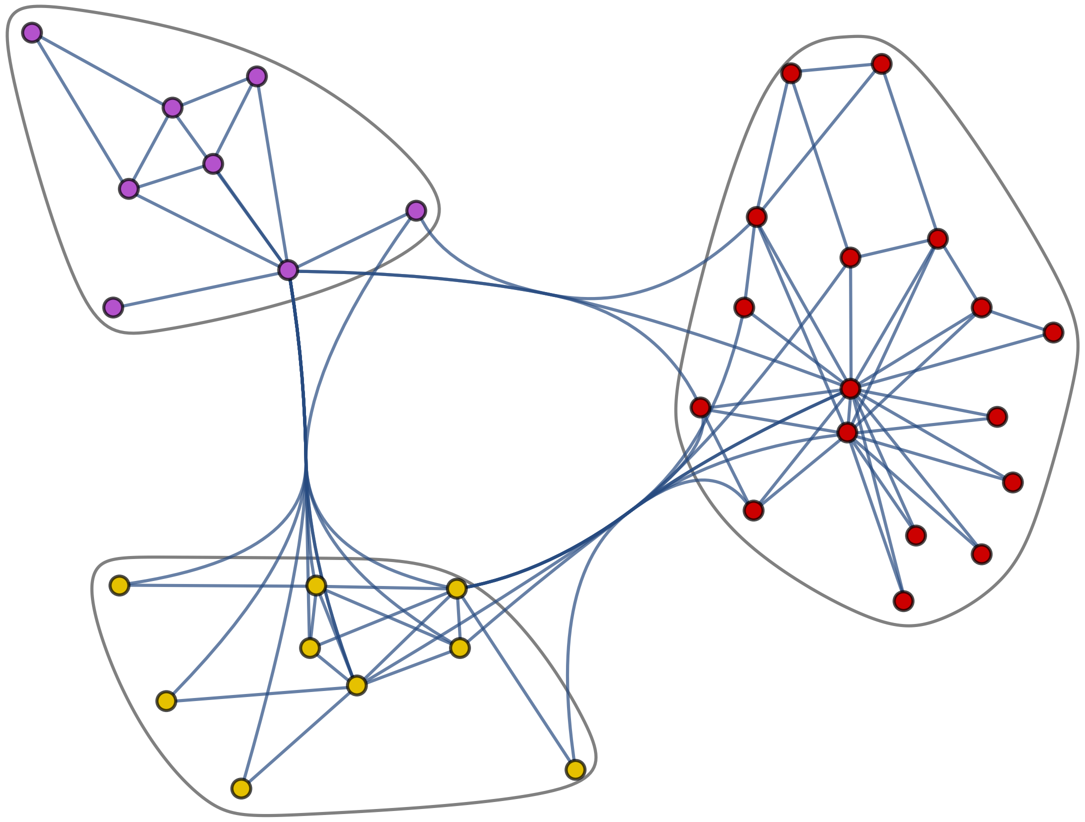
- Overlap coverage: how much overlap was discovered?
- Community coverage: how much of the network was classified?
- Overlap quality: how similar are overlapping nodes?
- Community quality: how similar are the nodes in a community?

Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann, “Link communities reveal multiscale complexity in networks,” Nature, vol. 466, no. 5, August, 2010.

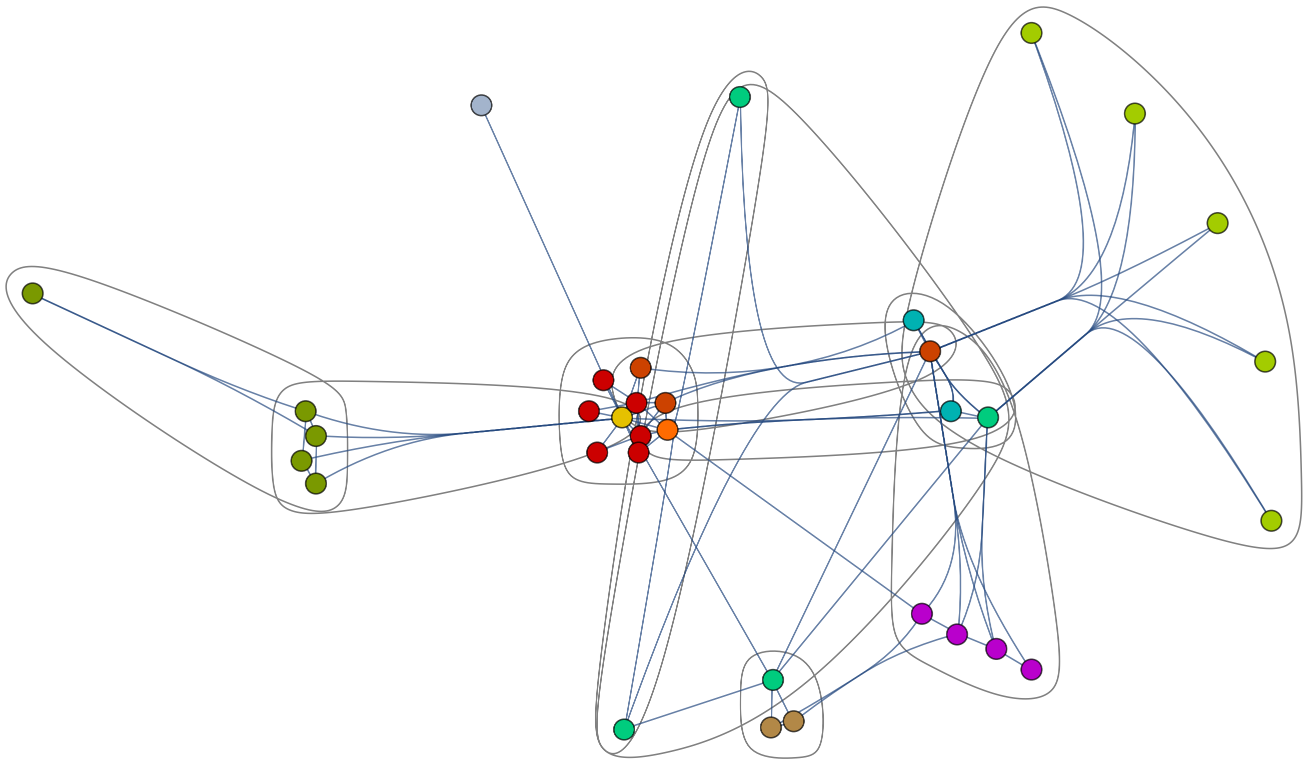
Non-overlapping communities



Non-overlapping vs. overlapping communities



Modularity



Hierarchical link clustering (11 communities)

Remarks

- Link partitioning
- How does it compare to:
    - Modularity based approximations
    - Non-modularity based methods
- Community: a set of closely interrelated links
- Nodes belong to multiples groups
- Links capture a well-defined relationship: two people...
    - are in the same family
    - work together
    - have common interests
- Incorporates overlap while revealing hierarchical organization

Remarks

- Link communities as fundamental blocks
    - Reveal overlapping
    - Hierarchical organization

Key features (summary)

- Number of communities
- Strength of communities
- Overlap coverage
- Community coverage

Next

- Degree analytics
  - Centrality measures
  - Centrality distributions (degree)