Social Network Analysis for Computer Scientists

Vincent Traag

CWTS, Leiden University

https://liacs.leidenuniv.nl/~takesfw/SNACS

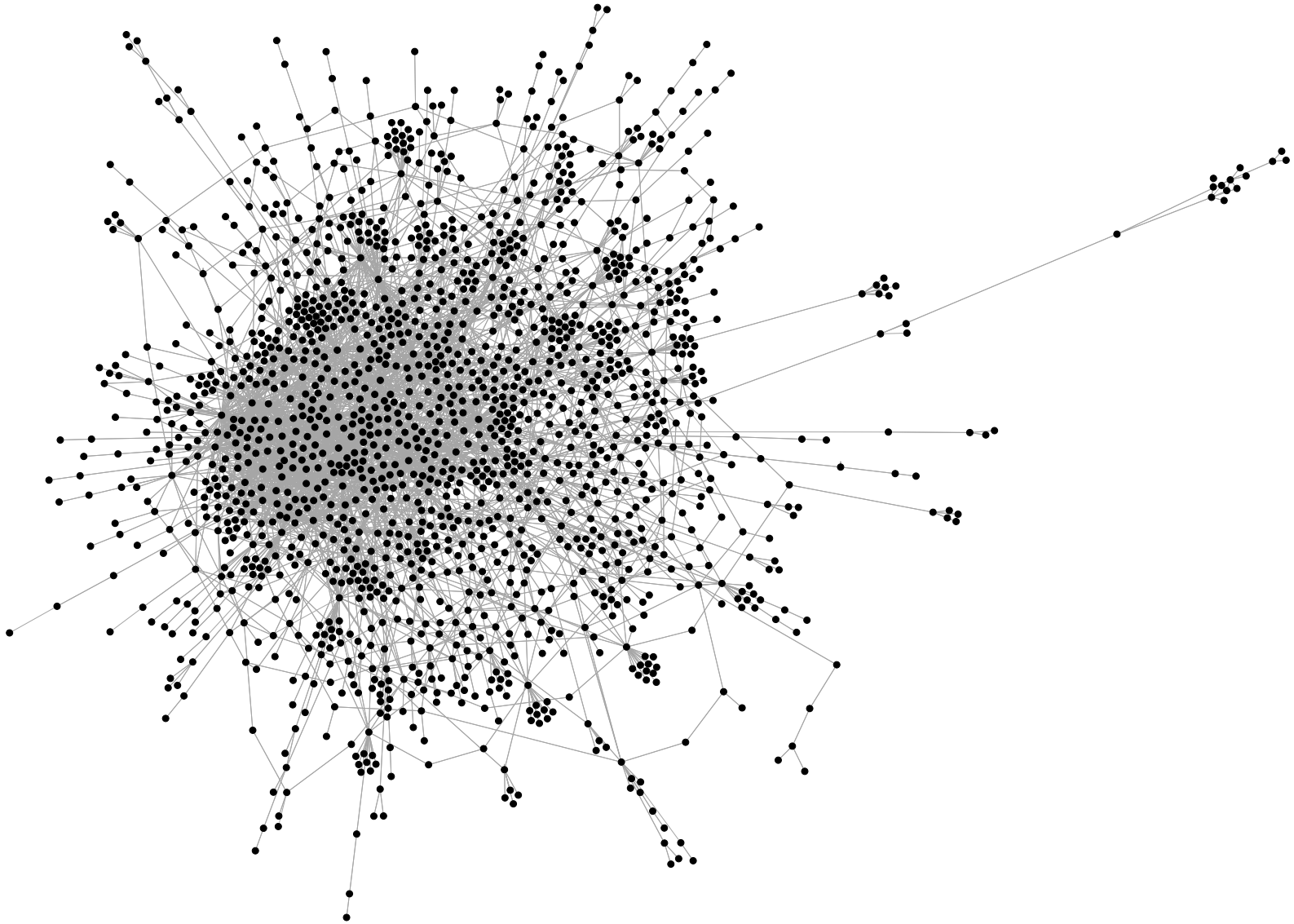
Lecture 4 — Community Detection

Overview today

- Recap
- Community detection
  - Cut-based perspective
  - Clustering perspective
  - Dynamical perspective
  - Inferential perspective

Recap

Networks



Notation

Concept	**Symbol**
Network (graph)	$G = (V, E)$
Nodes (objects, vertices, ...)	V
Links (ties, relationships, ...)	E
Directed — $E \subseteq V \times V$ — "links"	
Undirected — "edges"	
Number of nodes — $\lvert V \rvert$	n
Number of edges — $\lvert E \rvert$	m
Degree of node u	$deg(u)$
Distance from node u to v	$d(u, v)$

Real-world networks

1 Sparse networks	density
2 Fat-tailed power-law degree distribution	degree
3 Giant component	components
4 Low pairwise node-to-node distances	distance
5 Many triangles	clustering coefficient

Real-world networks

1. Sparse networks — density
2. Fat-tailed power-law degree distribution — degree
3. Giant component — components
4. Low pairwise node-to-node distances — distance
5. Many triangles — clustering coefficient

- Many examples: communication networks, citation networks, collaboration networks (Erdös, Kevin Bacon), protein interaction networks, information networks (Wikipedia), webgraphs, financial networks (Bitcoin) ...

Advanced concepts

- Assortativity, homophily
- Reciprocity
- Power law exponent
- Planar graphs
- Complete graphs
- Subgraphs
- Trees
- Spanning trees
- Diameter, eccentricity
- Bridges
- Graph traversal: DFS, BFS

Centrality measures

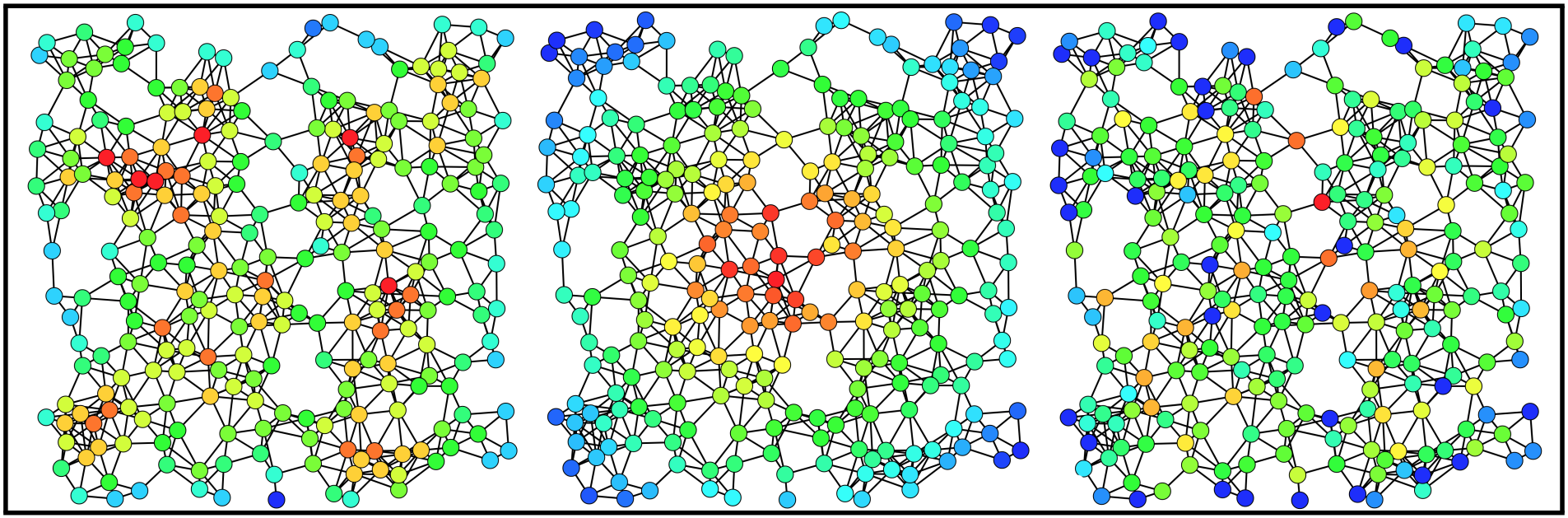
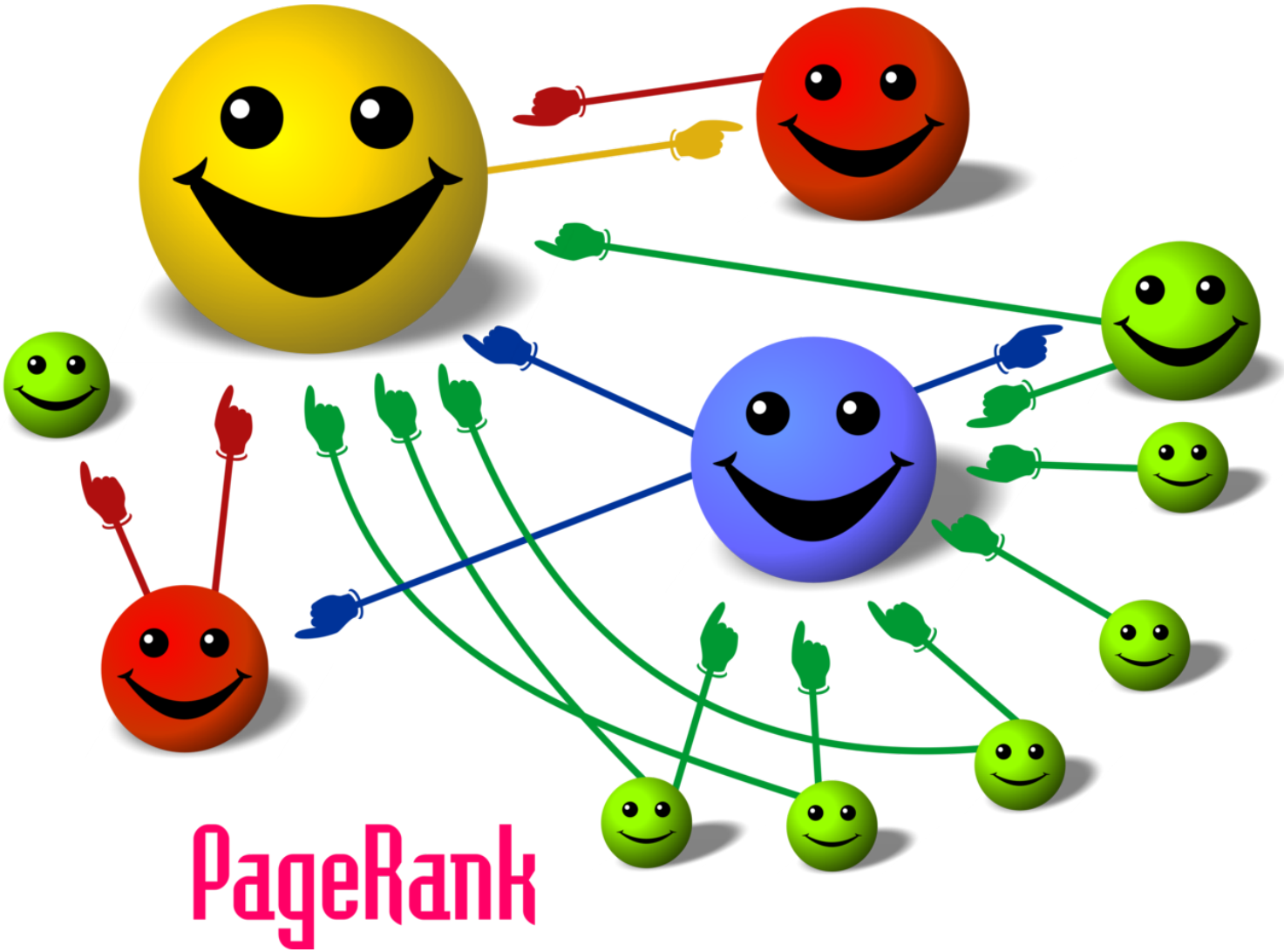


Figure: Degree, closeness and betweenness centrality

Source: "Centrality"' by Claudio Rocchini, Wikipedia `File:Centrality.svg`

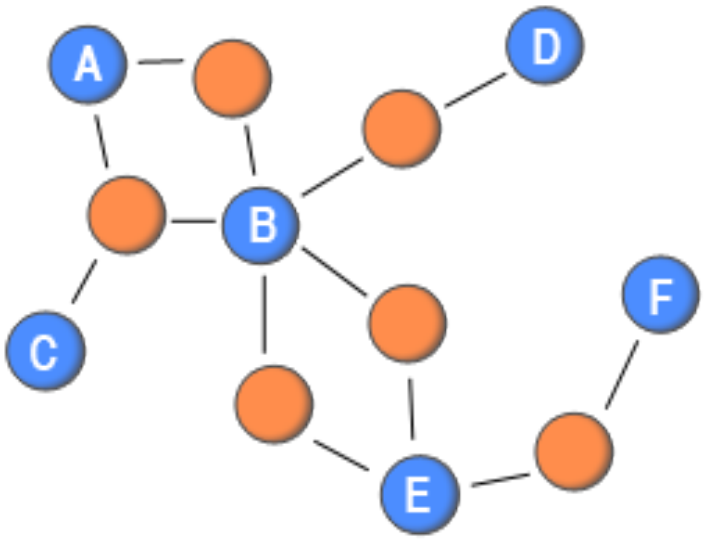
Centrality measures: PageRank



Centrality measures

- Distance/path-based measures:
  - Degree centrality $O(n)$
  - Closeness centrality $O(mn)$
  - Betweenness centrality $O(mn)$
  - Eccentricity centrality $O(mn)$
- **Propagation-based** measures:
  - Hyperlink Induced Topic Search (HITS) $O(m)$
  - PageRank $O(m)$

Network projection



$\rightarrow$

Network projection

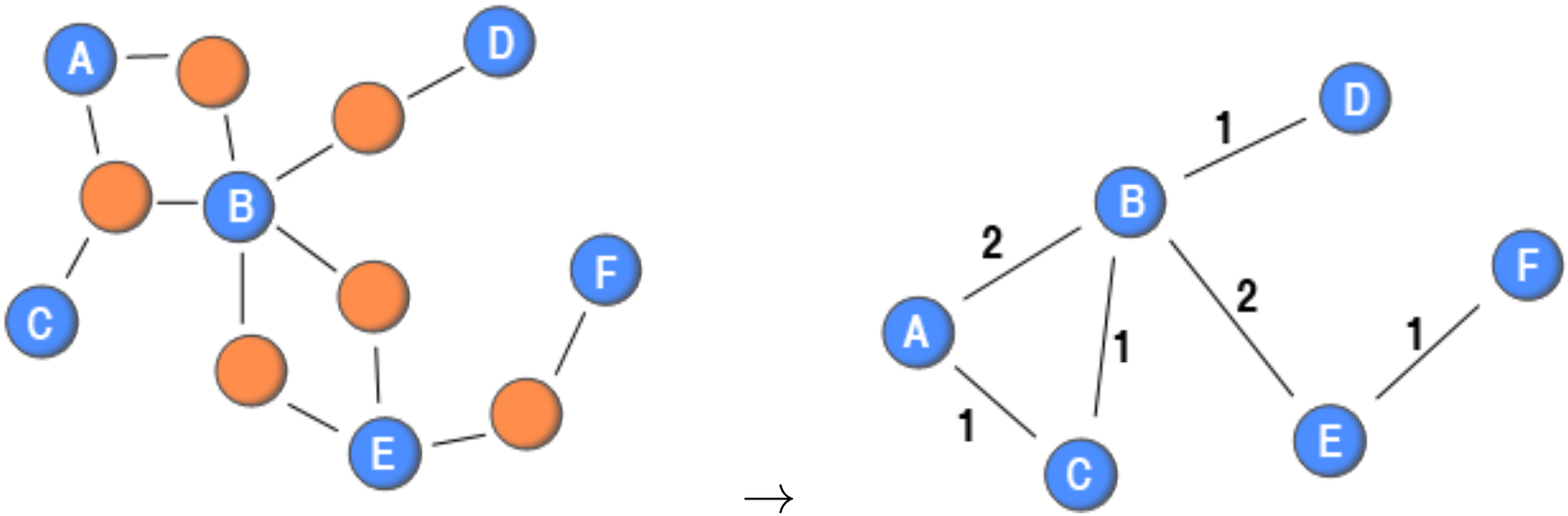


Image: `http://toreopsahl.com`

Community detection

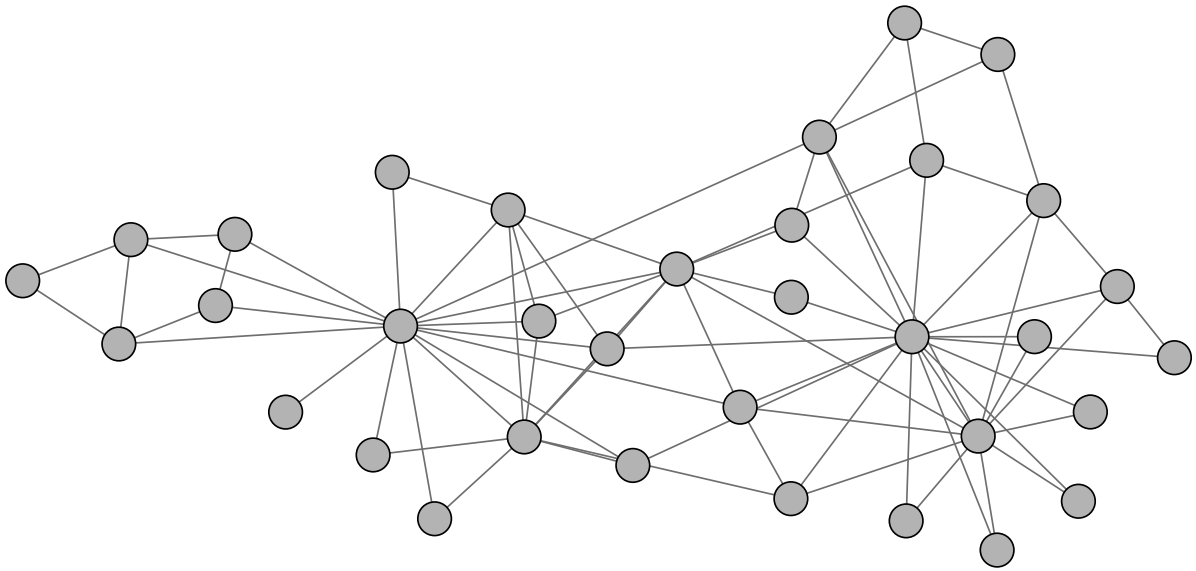


Figure: Communities: node subsets connected more strongly with each other

Community detection

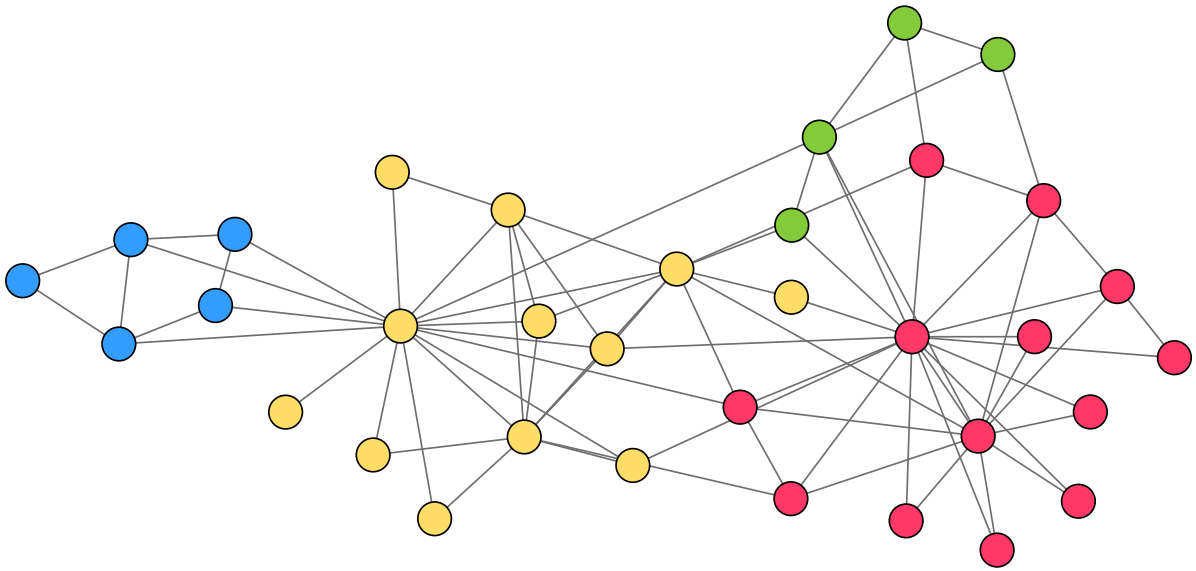
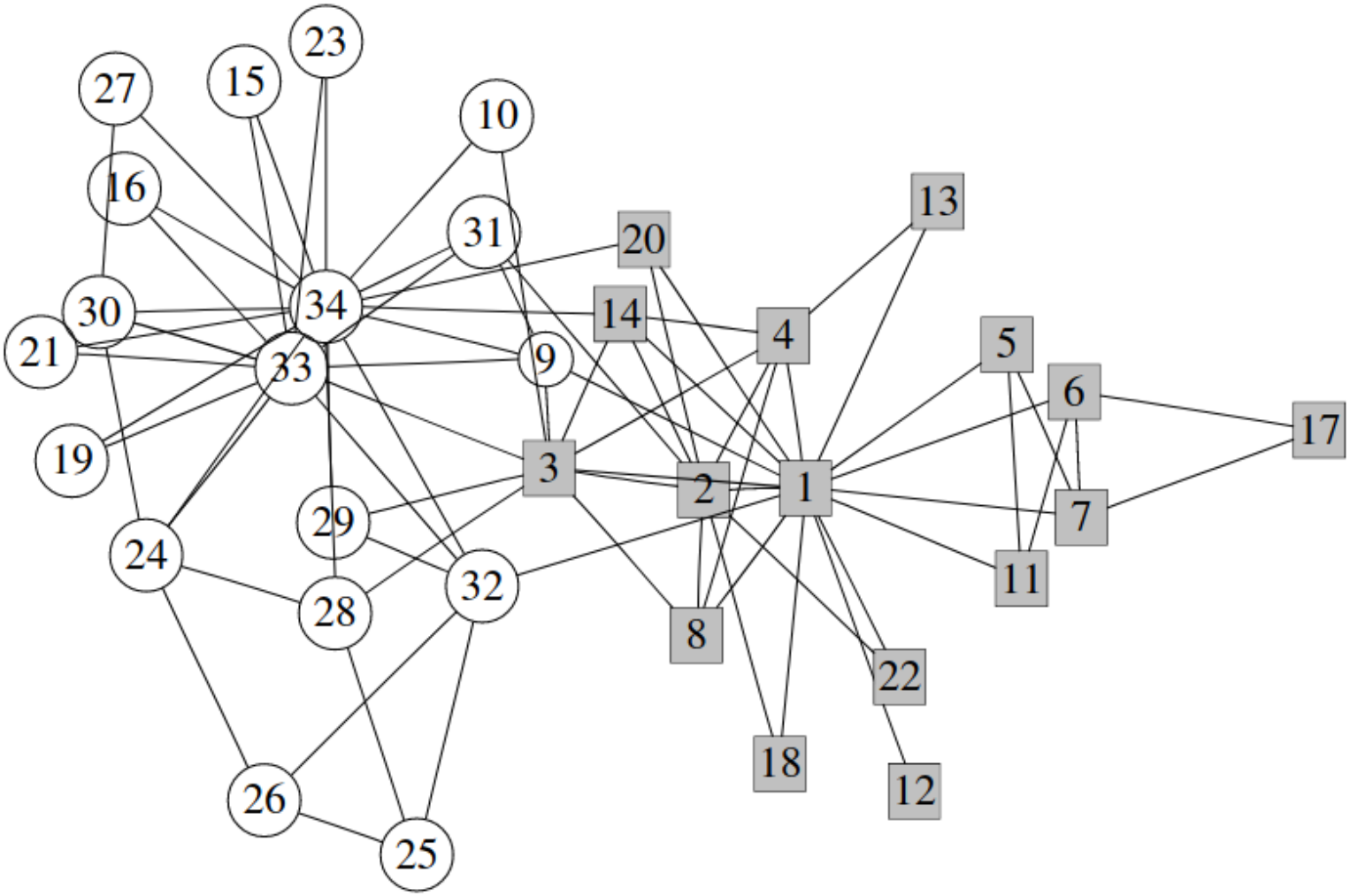


Figure: Communities: node subsets connected more strongly with each other

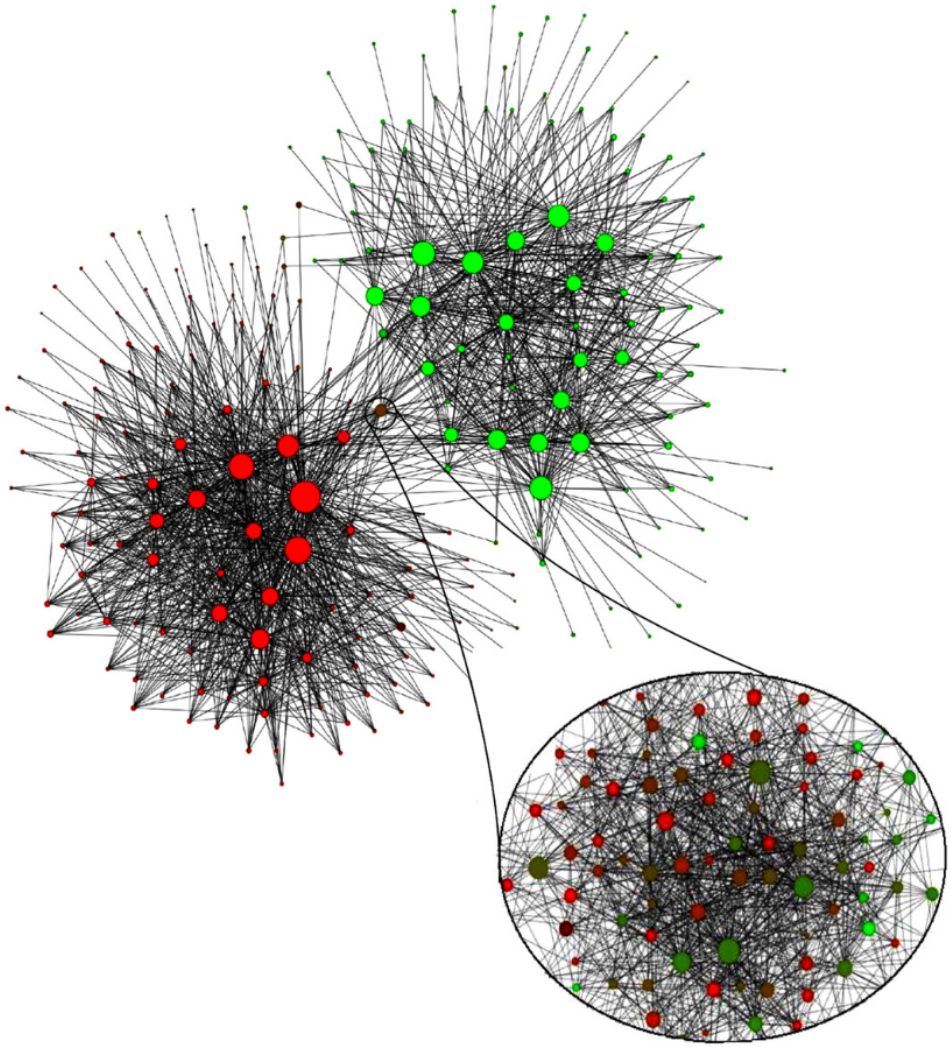
Introduction

Karate Club



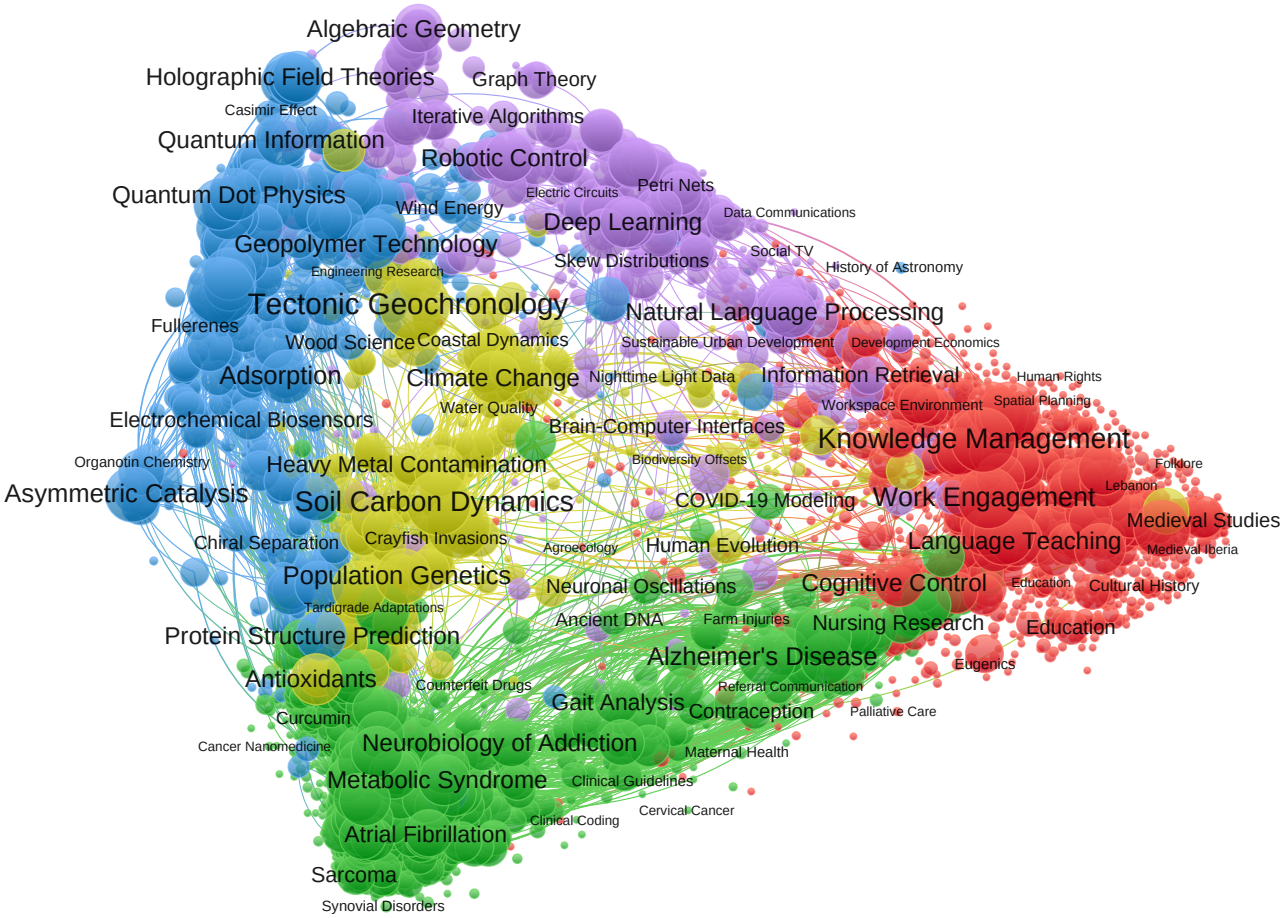
WW Zachary, *J Anthropol Res* **33**, 452–473, DOI 10.2307/3629752 (1977), M Newman, M Girvan, *Phys Rev E* **69**, 026113, DOI 10.1103/PhysRevE.69.026113 (2004).

Mobile phone network



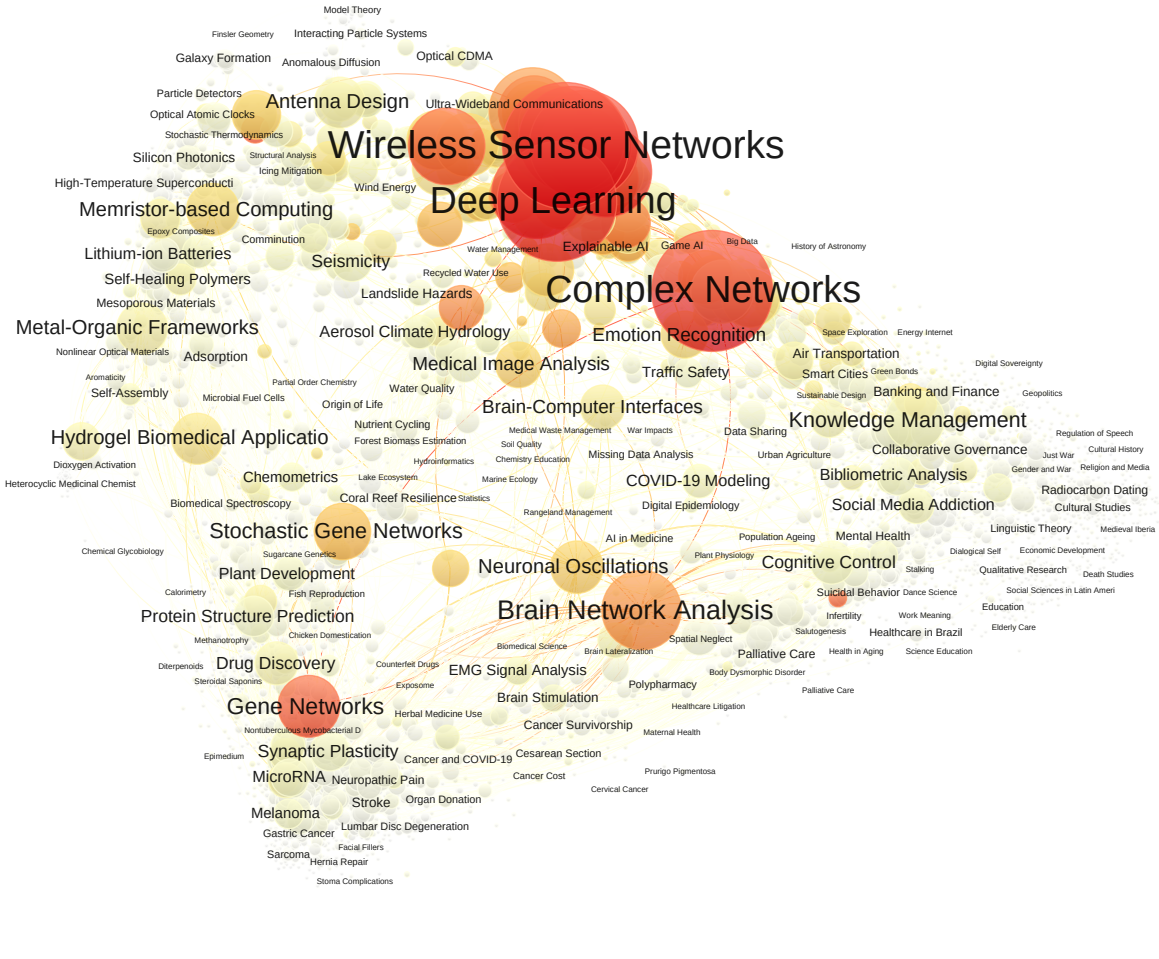
VD Blondel *et al.*, *J Stat Mech* **2008**, P10008, DOI 10.1088/1742-5468/2008/10/P10008 (2008).

Citation networks



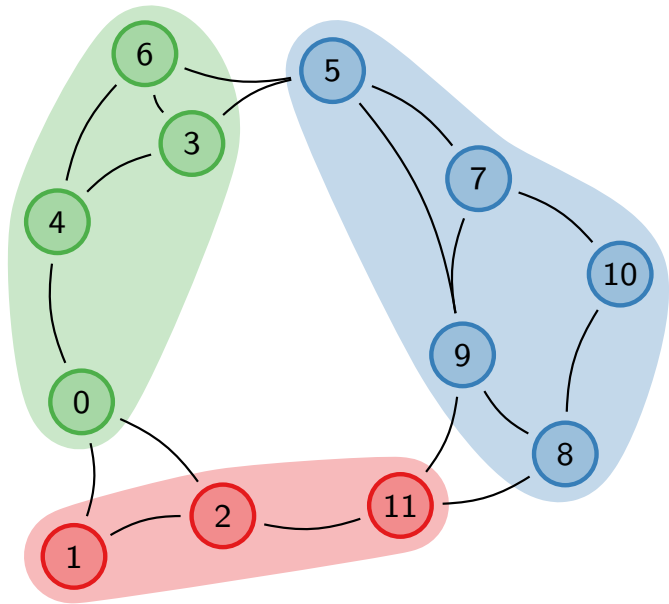
L Waltman, NJ van Eck, *Communications in Information Literacy* **63**, 2378–2392, DOI 10.1002/asi.22748 (2012), VA Traag, L Waltman, NJ van Eck, *Sci Rep* **9**, 5233, DOI 10.1038/s41598-019-41695-z (2019).

Citation networks



L Waltman, NJ van Eck, *Communications in Information Literacy* **63**, 2378–2392, DOI 10.1002/asi.22748 (2012), VA Traag, L Waltman, NJ van Eck, *Sci Rep* **9**, 5233, DOI 10.1038/s41598-019-41695-z (2019).

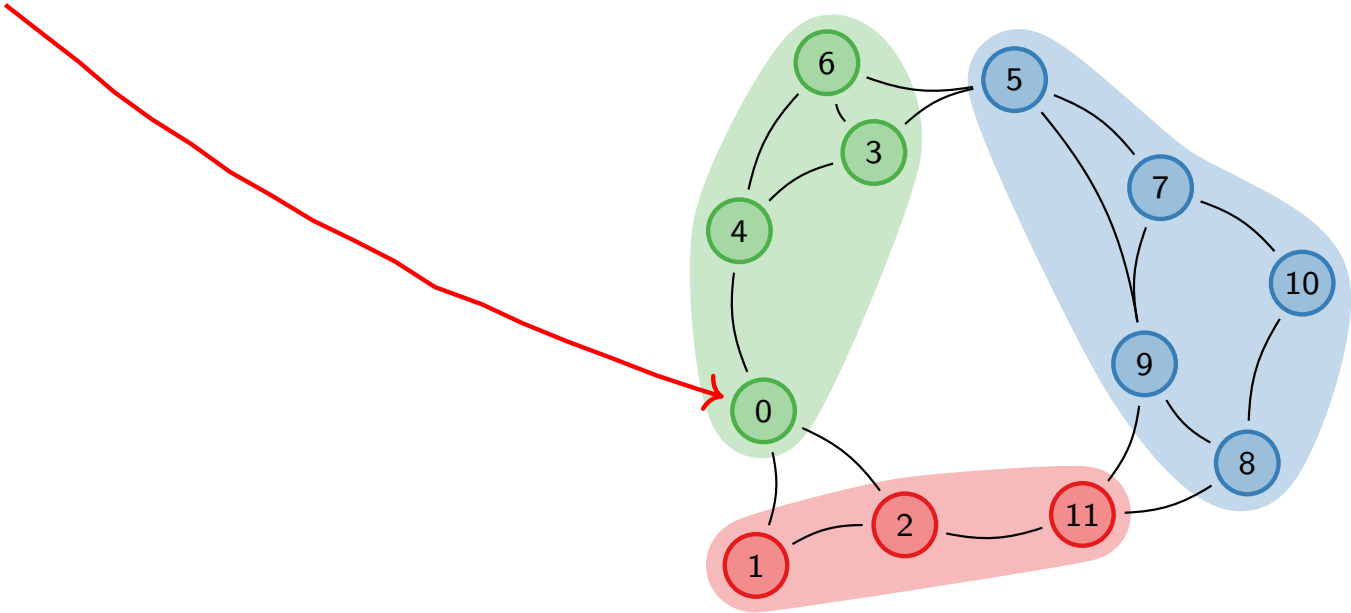
Label propagation algorithm



UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

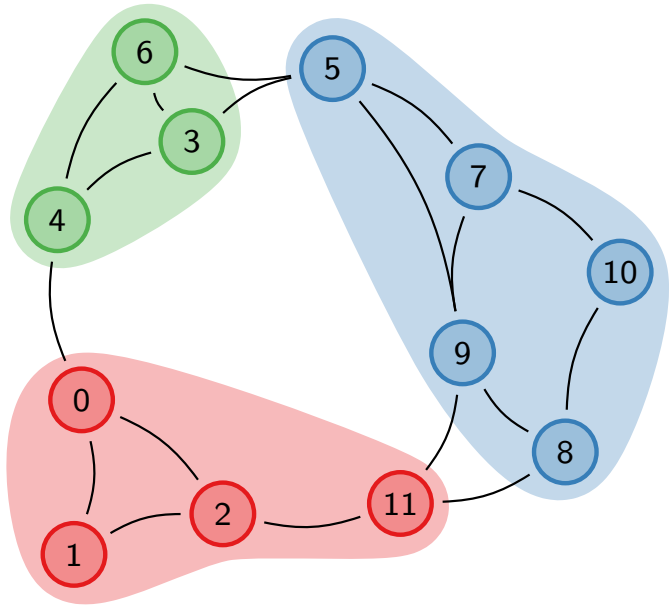
Change label 0?



UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

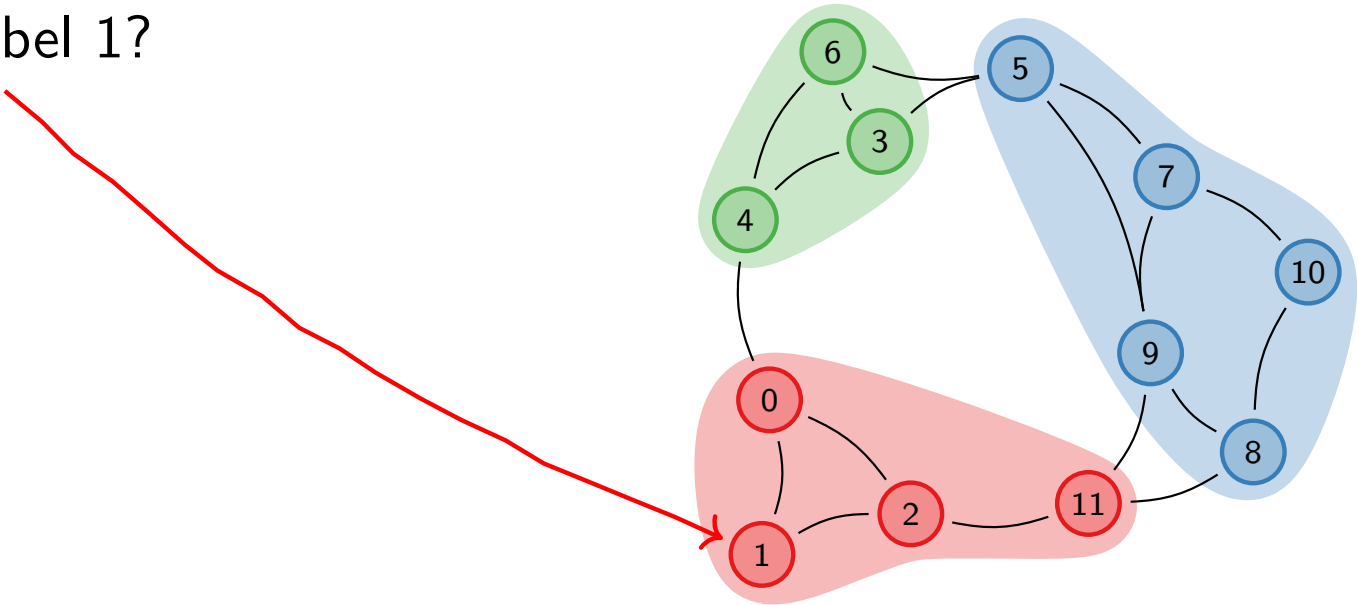


UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 1?

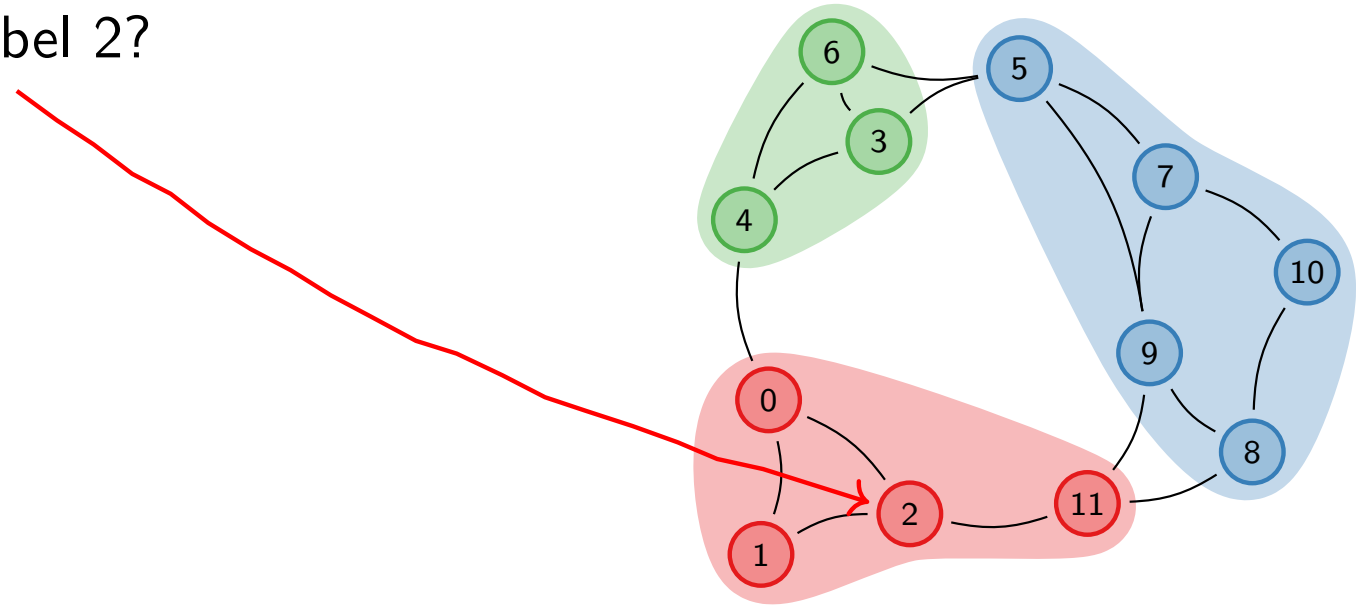


UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 2?

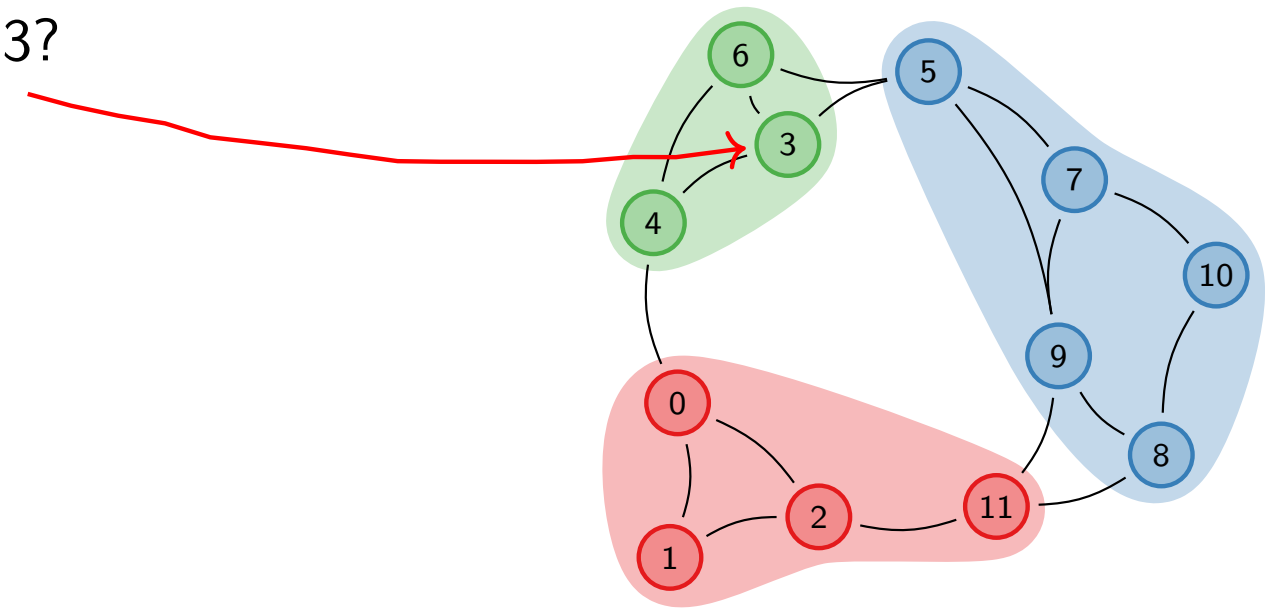


UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 3?

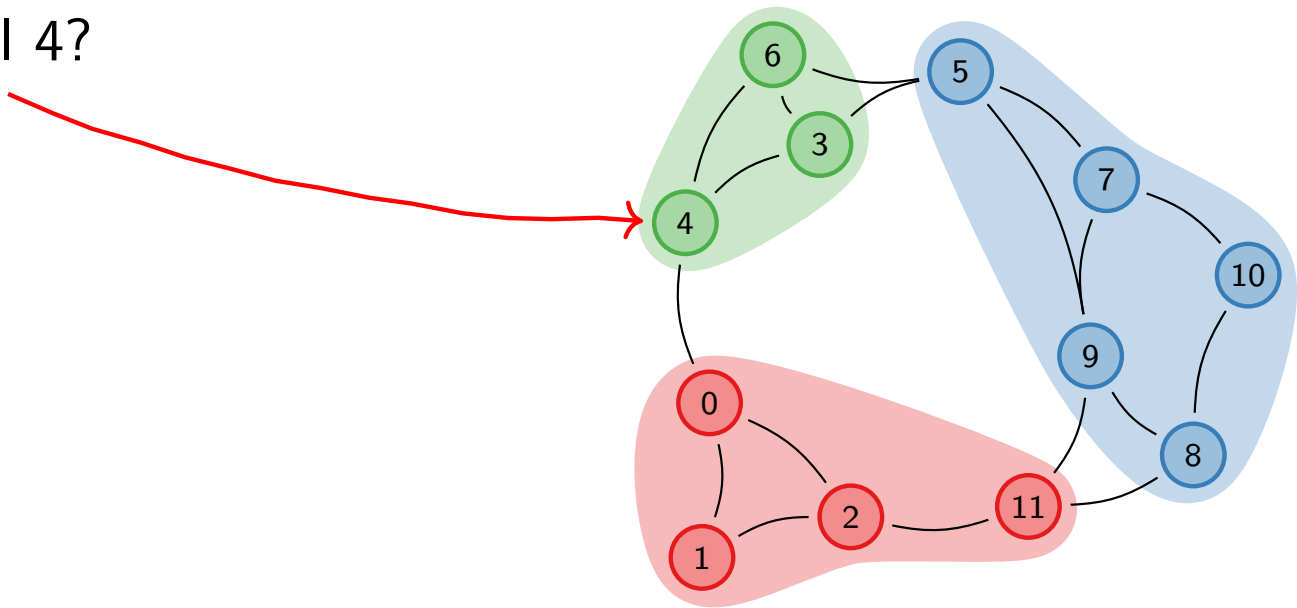


UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 4?

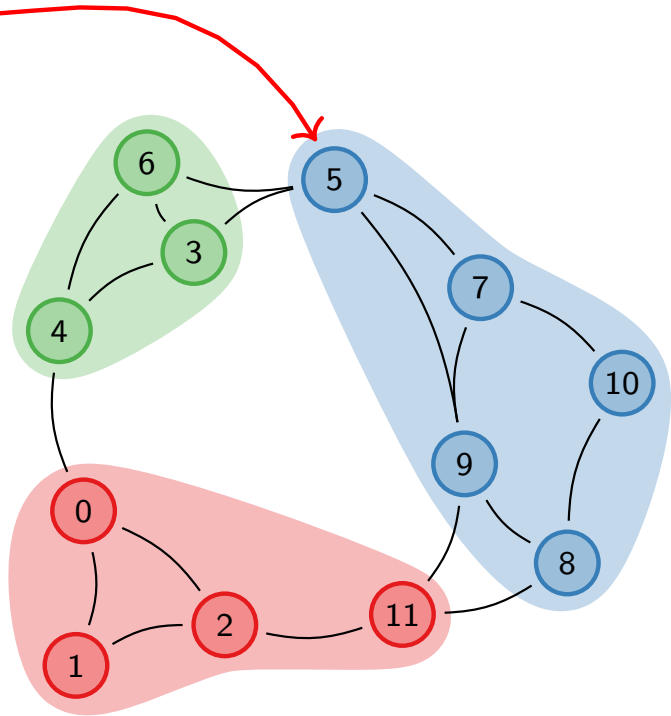


UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5?

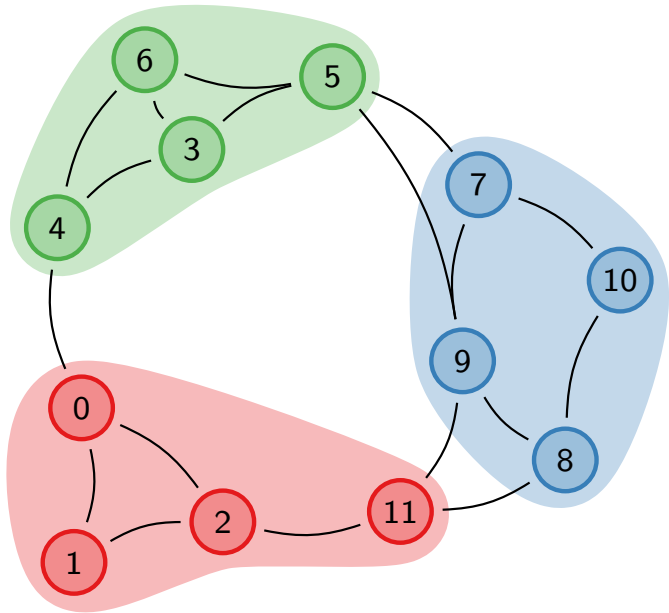


UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5



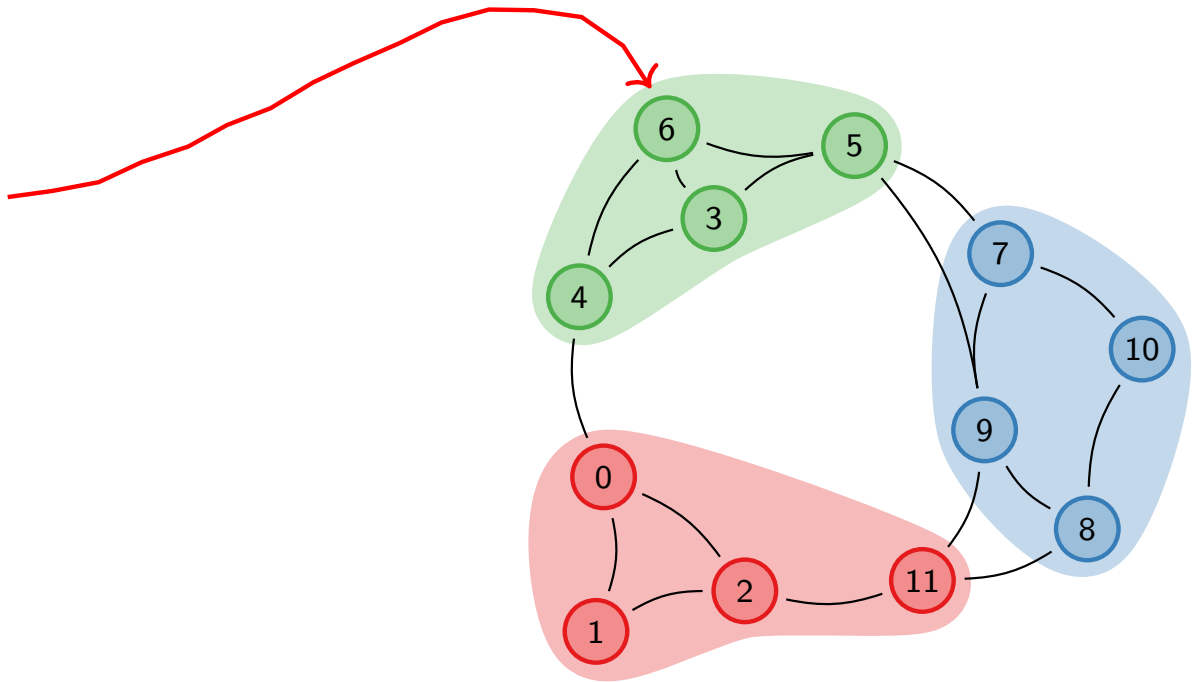
UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5

Change label 6?



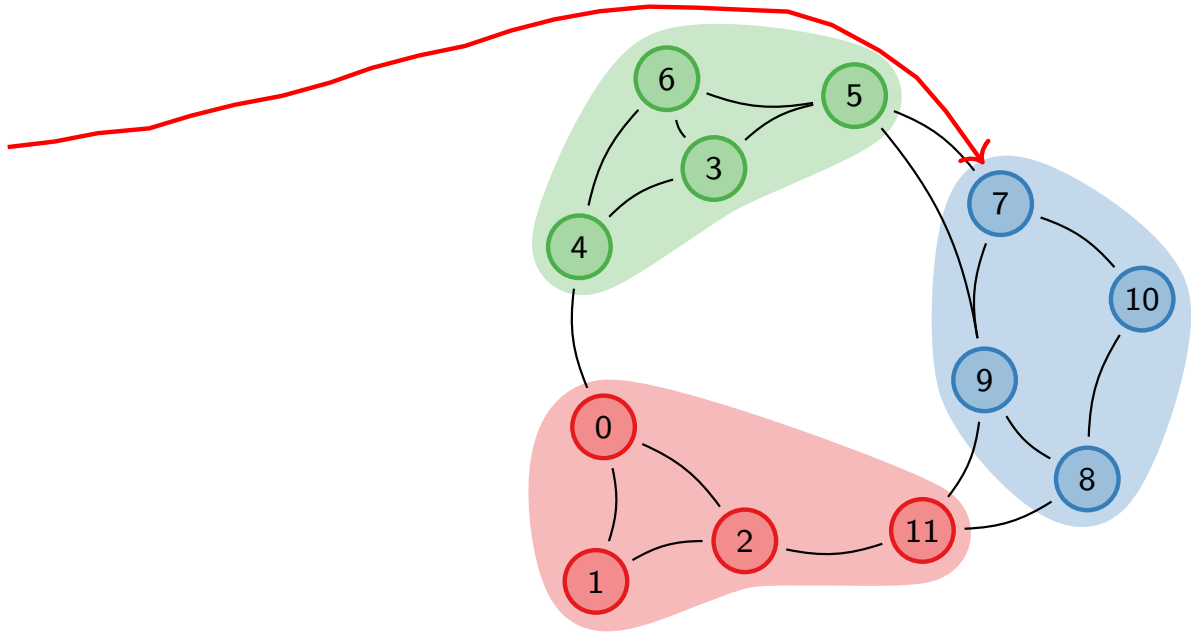
UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5

Change label 7?



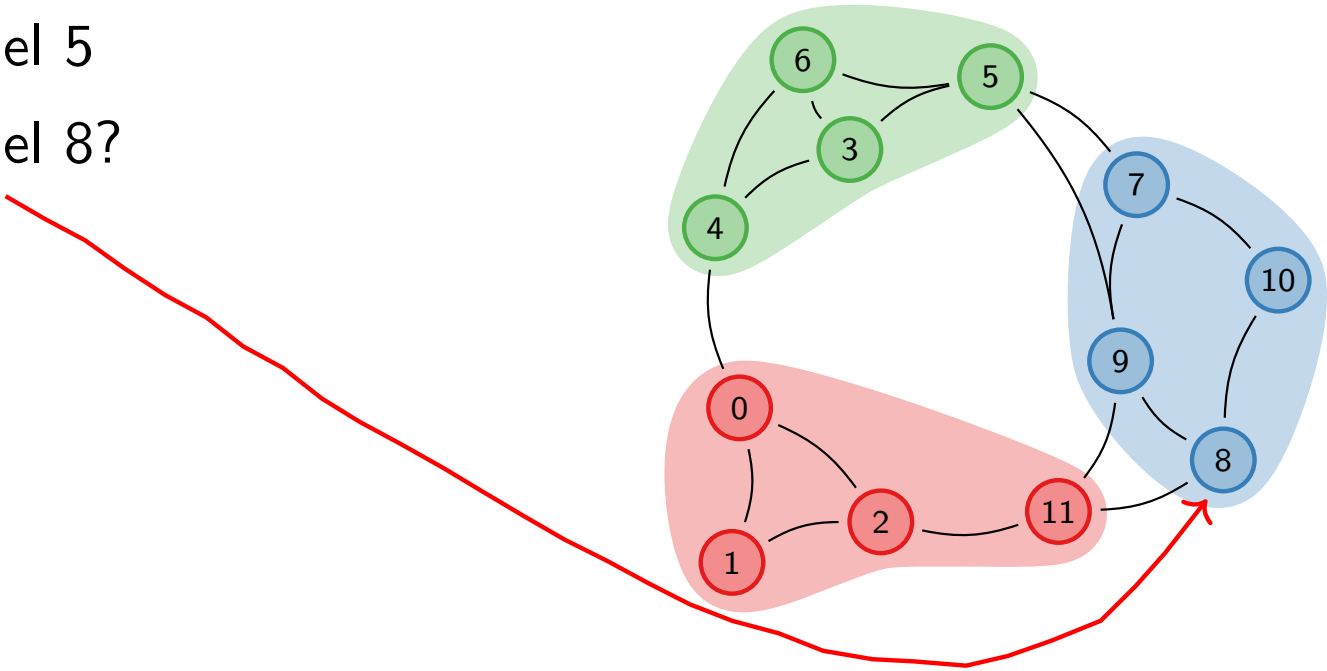
UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5

Change label 8?



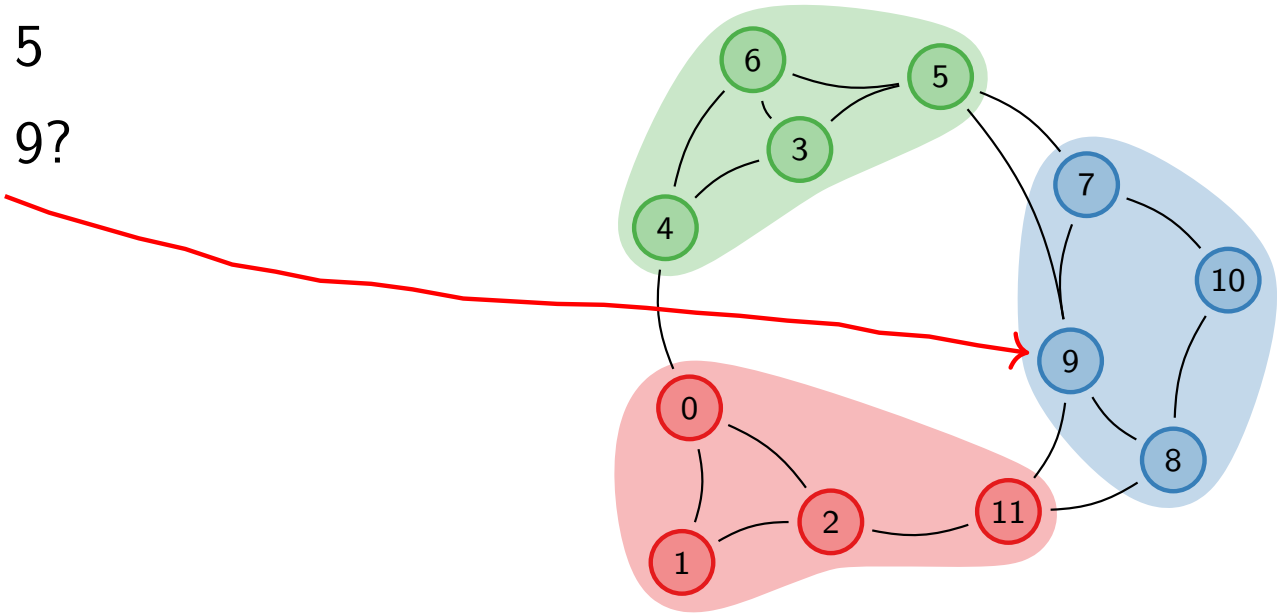
UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5

Change label 9?



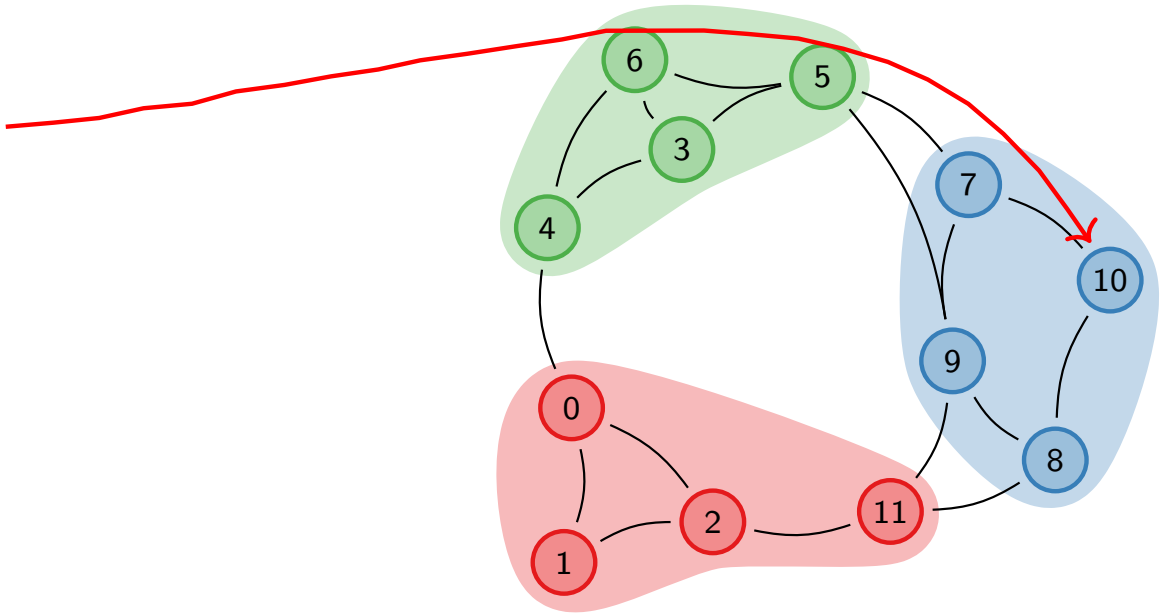
UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5

Change label 10?



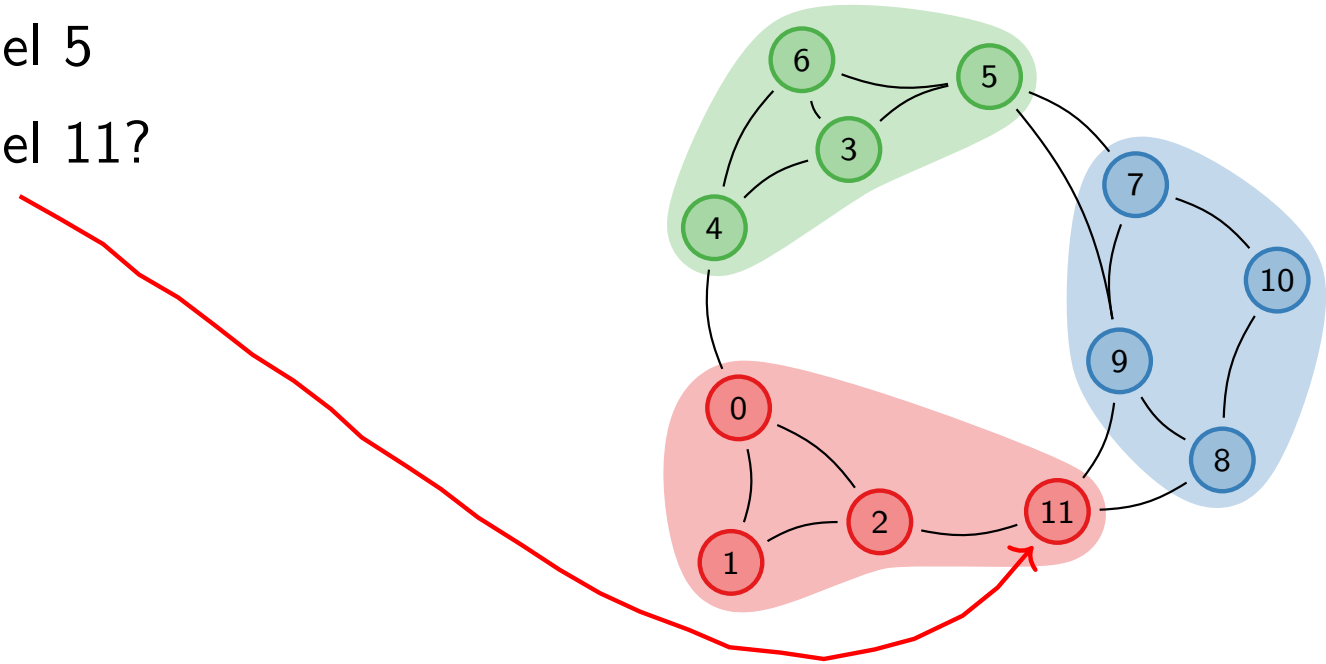
UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5

Change label 11?



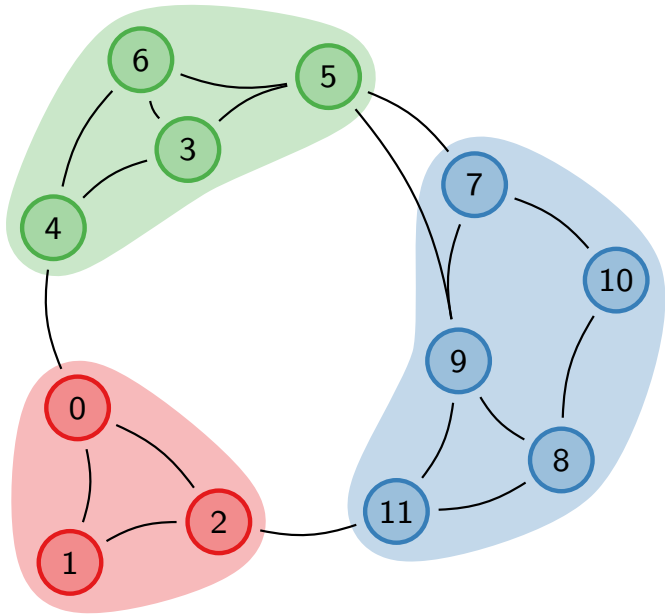
UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Label propagation algorithm

Change label 0

Change label 5

Change label 11



UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

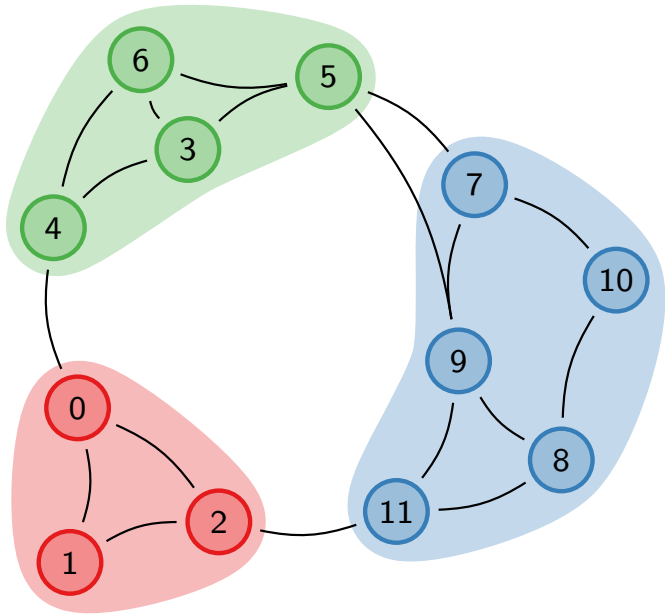
Label propagation algorithm

Change label 0

Change label 5

Change label 11

Repeat everything.



UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

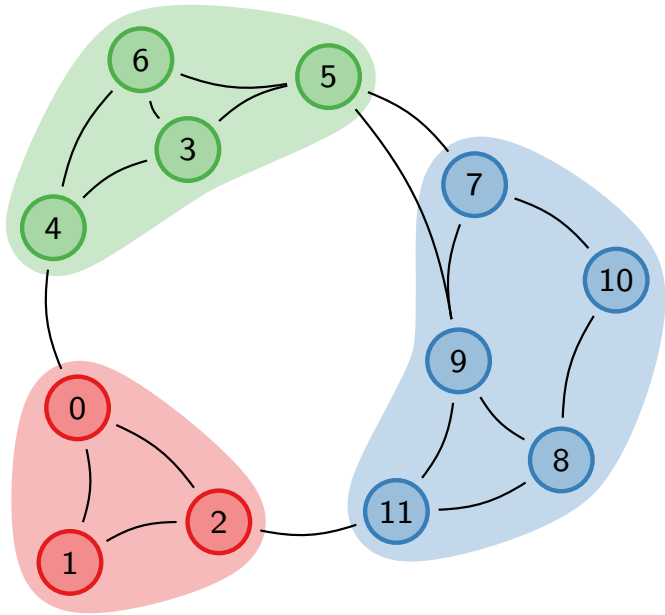
Label propagation algorithm

Change label 0

Change label 5

Change label 11

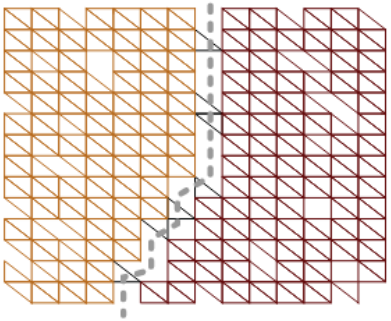
Everything is maximal



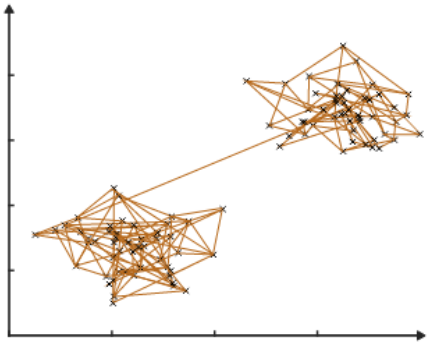
UN Raghavan, R Albert, S Kumara, *Phys. Rev. E* **76**, 036106, DOI 10.1103/PhysRevE.76.036106 (2007).

Community detection approaches

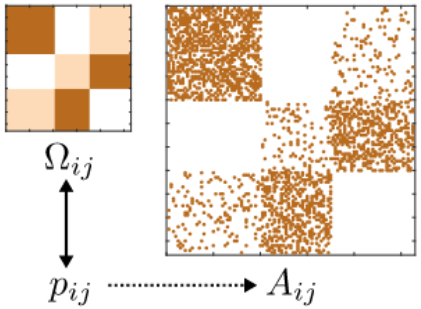
(i) Cut-based perspective



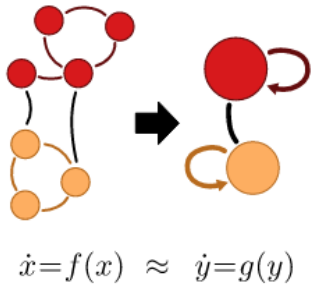
(ii) Clustering perspective



(iii) Stochastically equivalent nodes



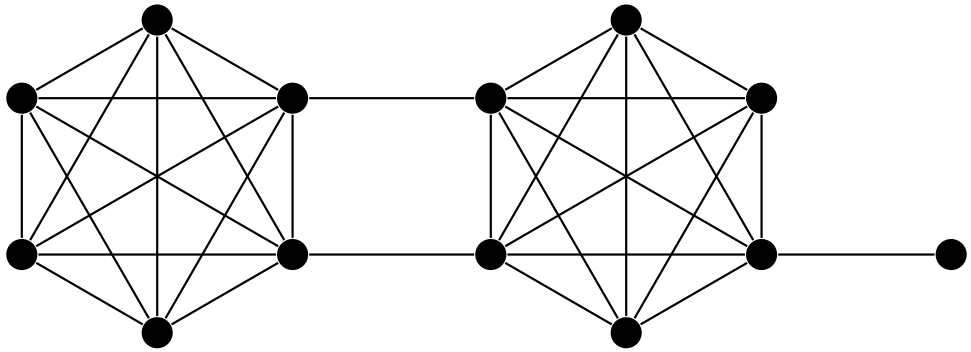
(iv) Dynamical perspective



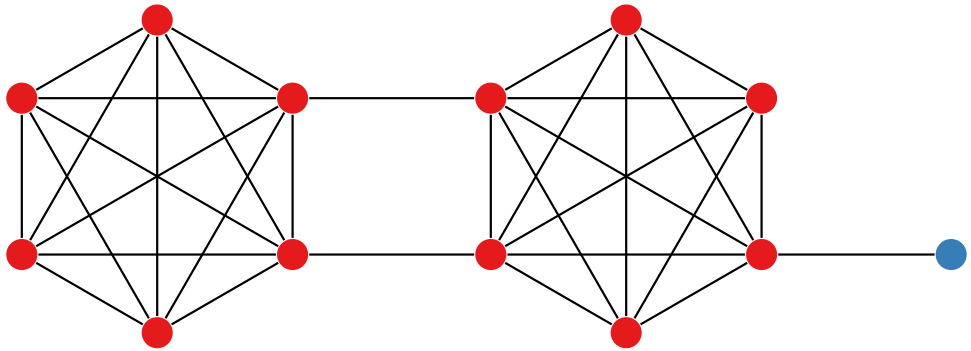
M Rosvall *et al.*, in *Advances in Network Clustering and Blockmodeling* (John Wiley & Sons, Ltd, 2019), pp. 105–119, DOI 10.1002/9781119483298.ch4.

Cut-based perspective

Minimum cut



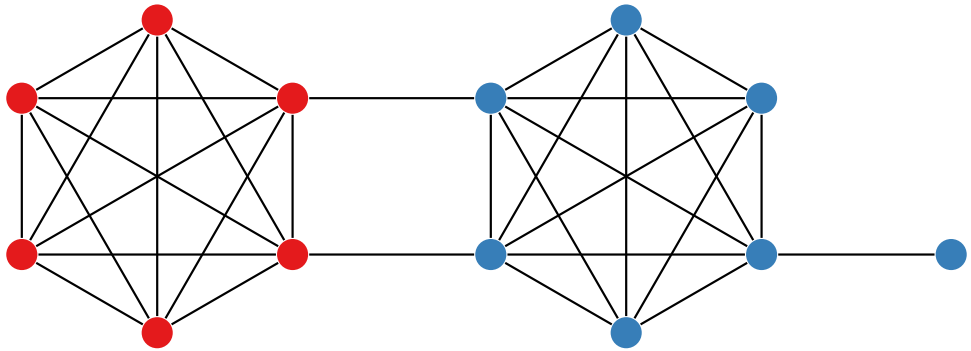
Minimum cut



$$\min_{\{V_i\}} \frac{1}{2} \sum_i E(V_i, V - V_i)$$

- Easy to solve, min-cut, max-flow.

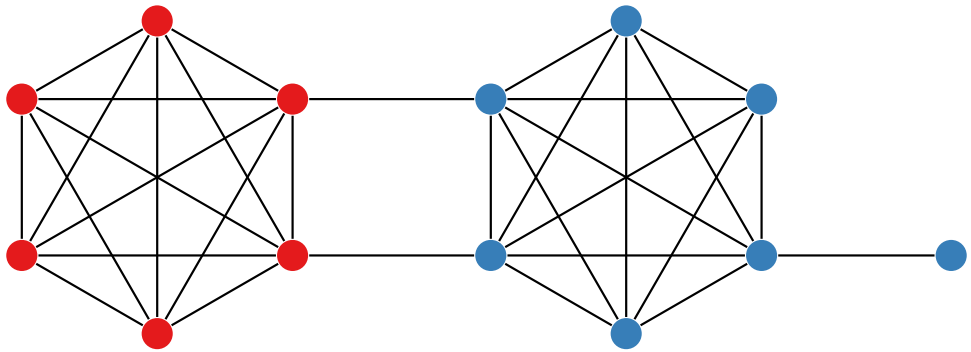
Ratio cut



$$\min_{\{V_i\}} \frac{1}{2} \sum_i \frac{E(V_i, V - V_i)}{|V_i|}$$

- NP hard

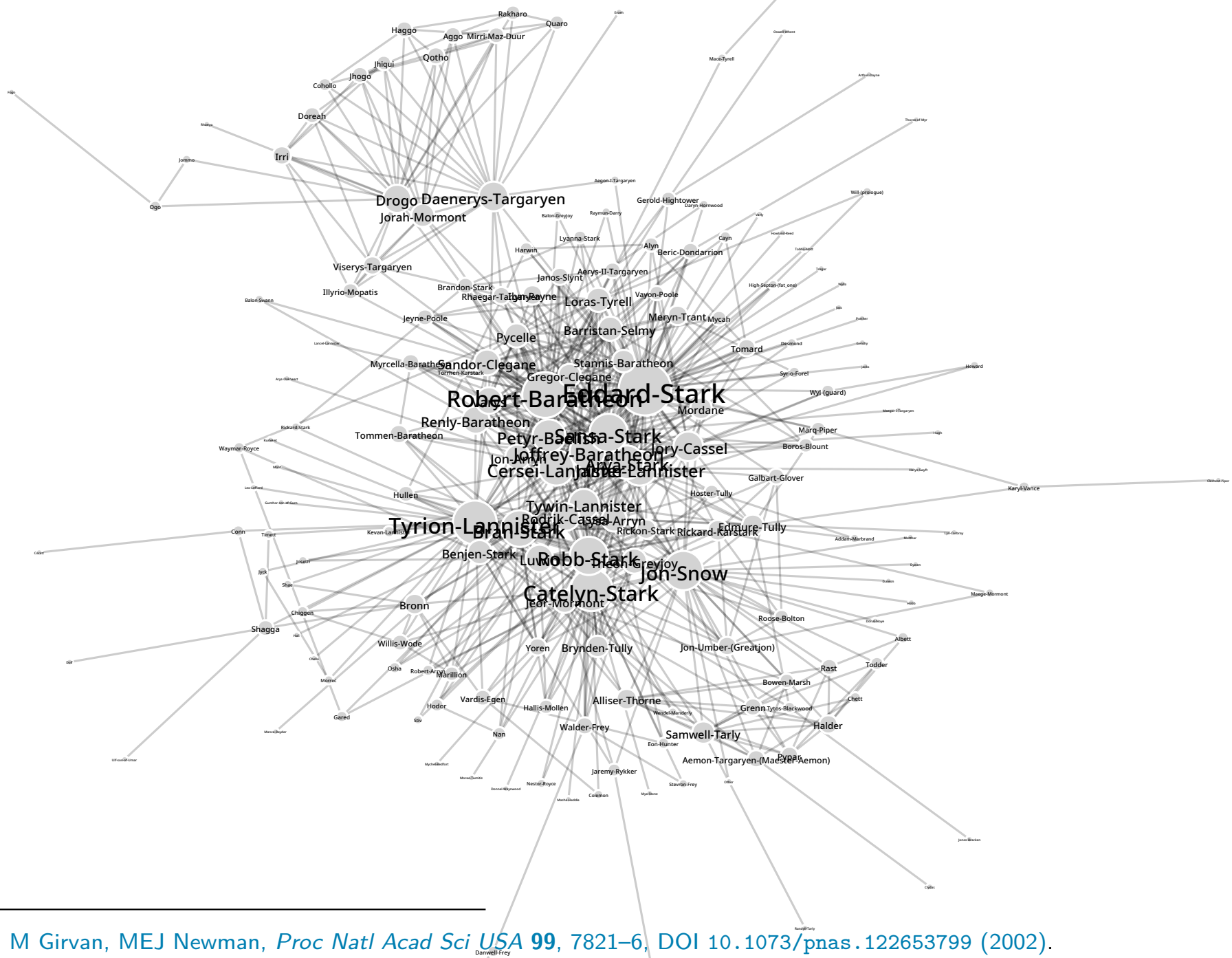
Ratio cut approximation



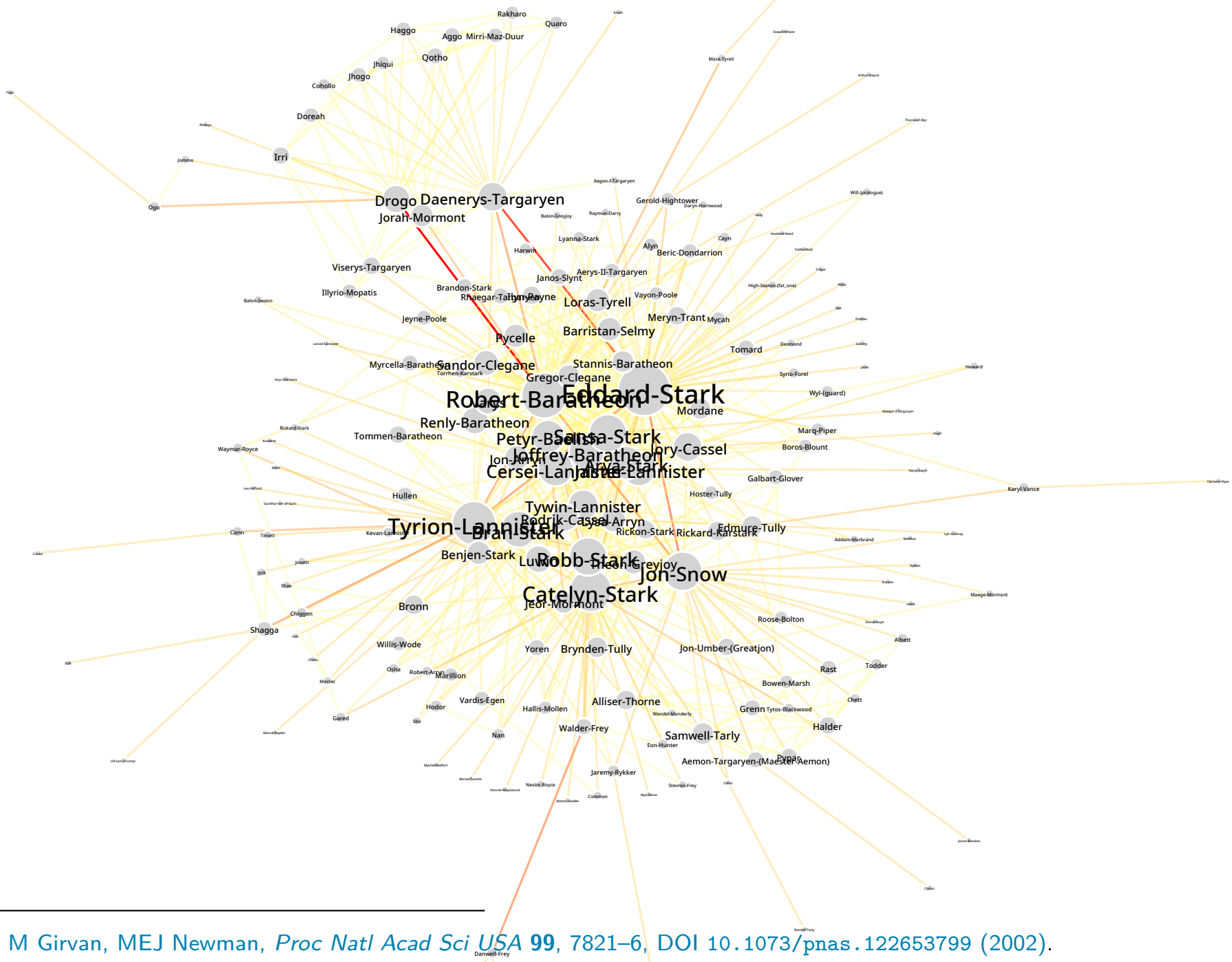
$$\min_{\{V_i\}} \frac{1}{2} \sum_i \frac{E(V_i, V - V_i)}{|V_i|} \sim \min_x x^\intercal \mathcal{L} x$$

- Laplacian $\mathcal{L} = D - A$.
- x_i specific values, exact optimisation (NP Hard).
- $x_i \in \mathbb{R}$, approximation, second eigenvector (Fiedler vector).

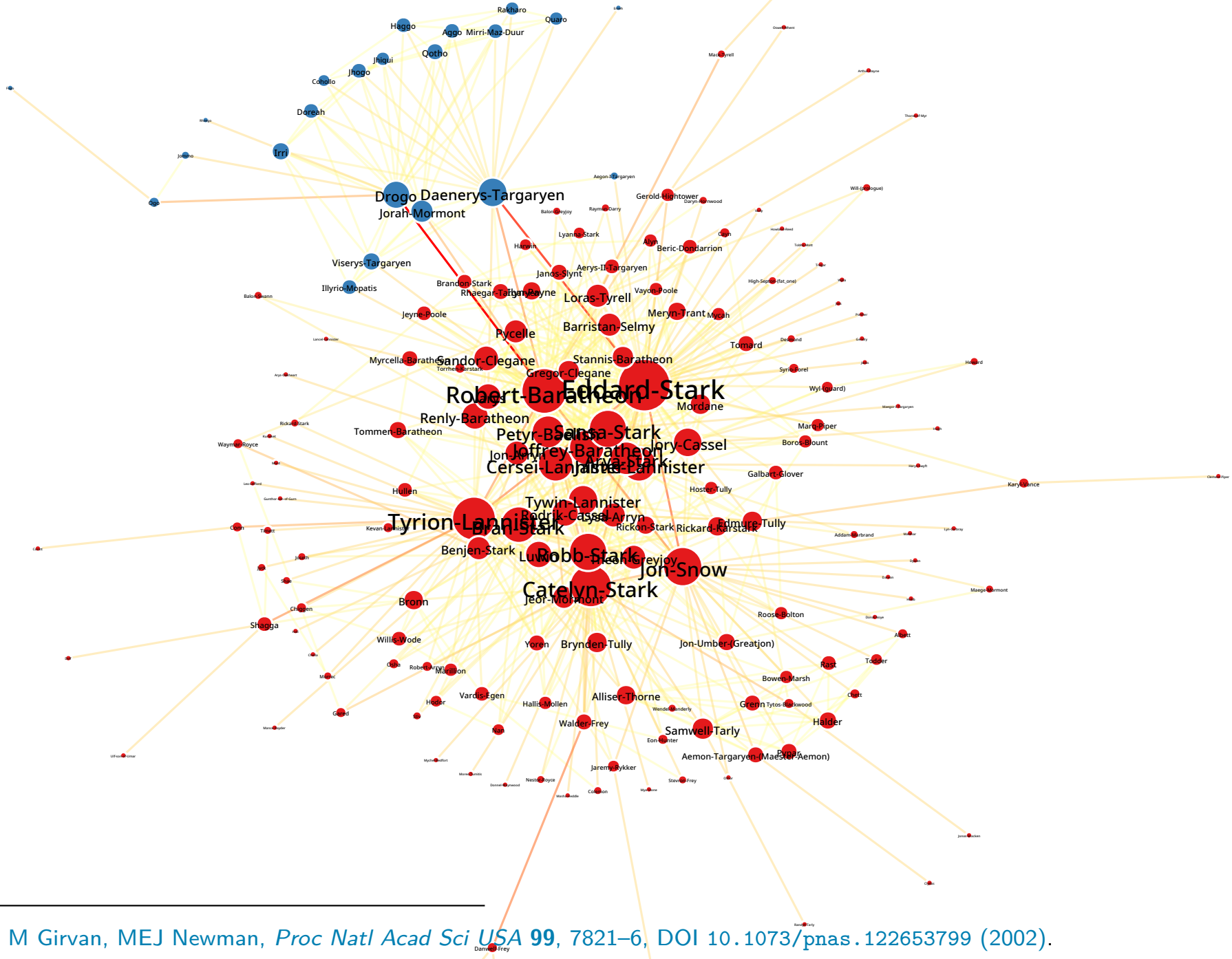
U von Luxburg, *Stat Comput* **17**, 395–416, DOI 10.1007/s11222-007-9033-z (2007).



M Girvan, MEJ Newman, *Proc Natl Acad Sci USA* **99**, 7821–6, DOI 10.1073/pnas.122653799 (2002).

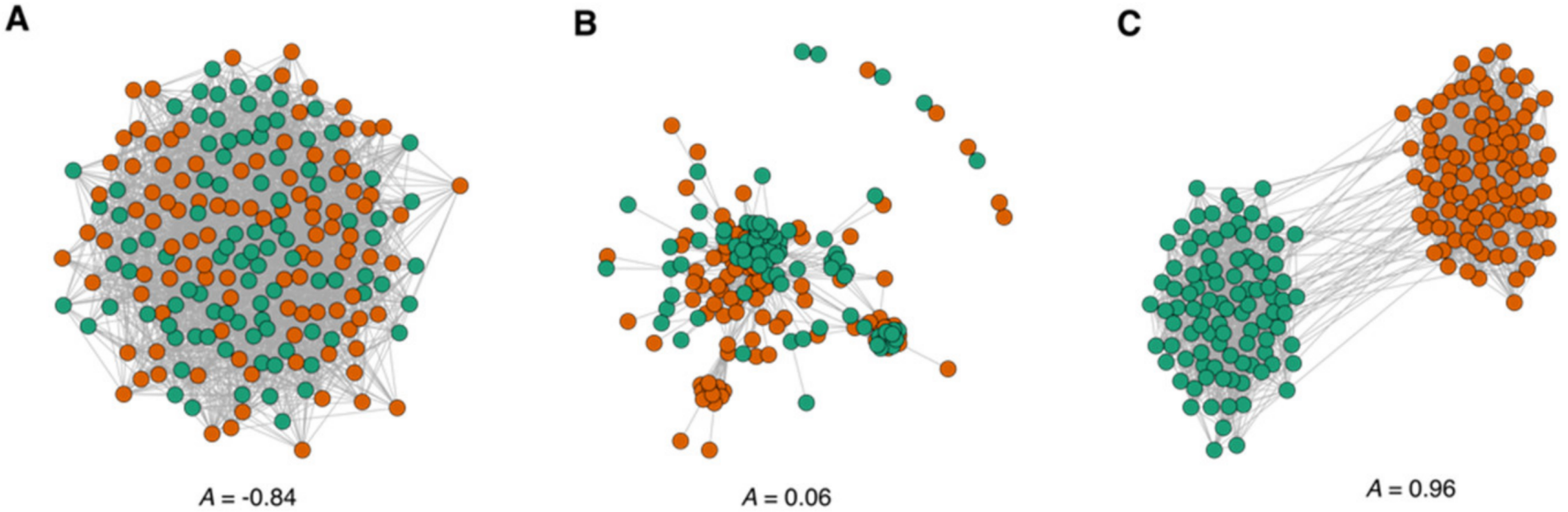


M Girvan, MEJ Newman, *Proc Natl Acad Sci USA* **99**, 7821–6, DOI 10.1073/pnas.122653799 (2002).



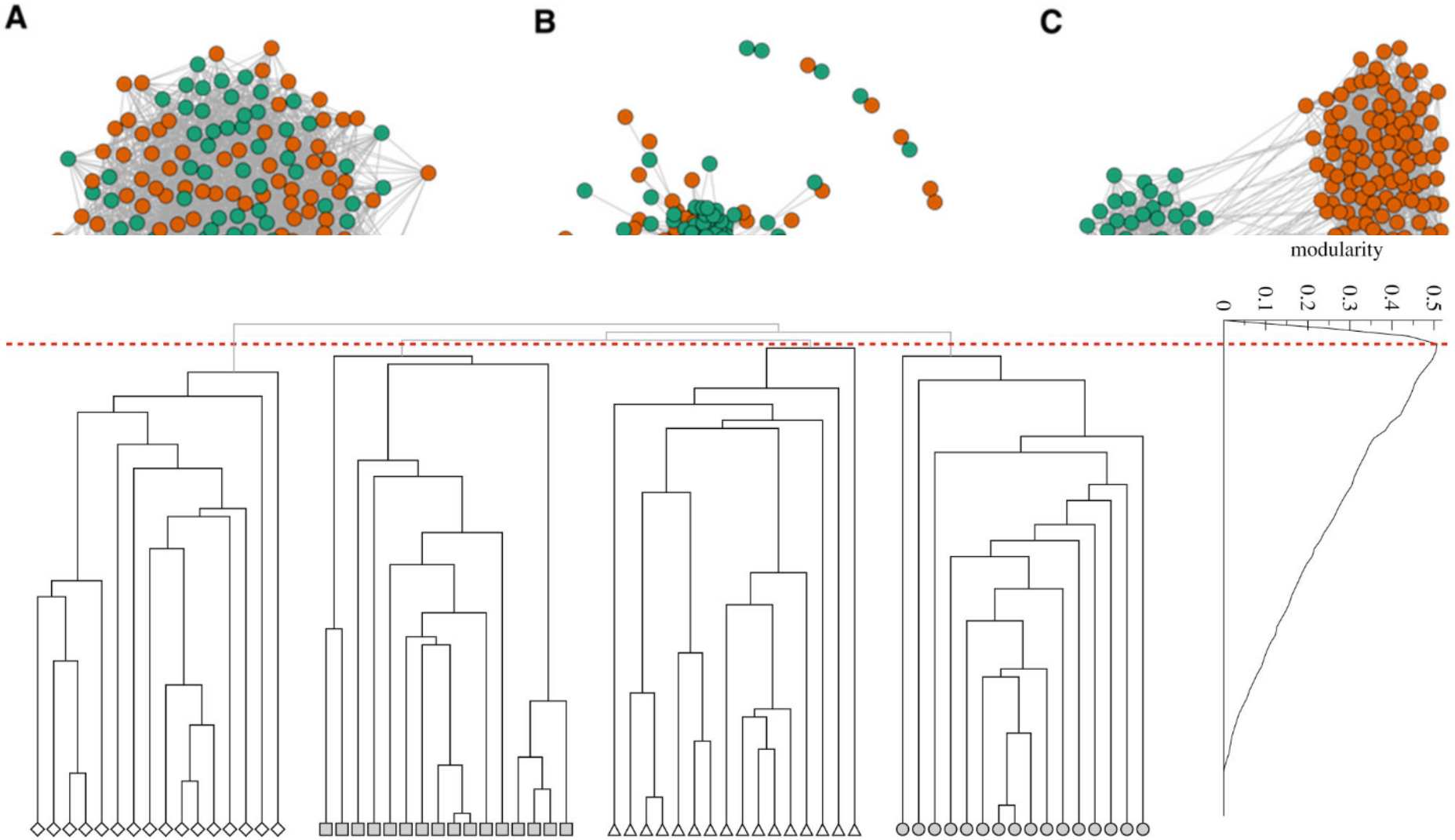
M Girvan, MEJ Newman, *Proc Natl Acad Sci USA* **99**, 7821–6, DOI 10.1073/pnas.122653799 (2002).

Dendrogram cutting



AA Moya-García *et al.*, *Genes* **13**, 1292, DOI 10.3390/genes13071292 (2022).

Dendrogram cutting



M Newman, M Girvan, *Phys Rev E* **69**, 026113, DOI 10.1103/PhysRevE.69.026113 (2004).

Clustering perspective

Modularity

Find partition that maximises

$$\mathcal{Q} = \frac{1}{2m} \sum_c \Big(e_c \; - \; \langle e_c \rangle \Big).$$

M Newman, M Girvan, *Phys Rev E* **69**, 026113, DOI 10.1103/PhysRevE.69.026113 (2004).

Modularity

Find partition that maximises

$$\mathcal{Q} = \frac{1}{2m} \sum_c \Big(e_c \; - \; \langle e_c \rangle \Big).$$

- e_c: Actual number of edges in community c,

M Newman, M Girvan, *Phys Rev E* **69**, 026113, DOI 10.1103/PhysRevE.69.026113 (2004).

Modularity

Find partition that maximises

$$\mathcal{Q} = \frac{1}{2m} \sum_c \Big(e_c \; - \; \langle e_c \rangle \Big).$$

- e_c: Actual number of edges in community c,
- $\langle e_c \rangle$: Expected number of edges in community c.

M Newman, M Girvan, *Phys Rev E* **69**, 026113, DOI 10.1103/PhysRevE.69.026113 (2004).

Modularity

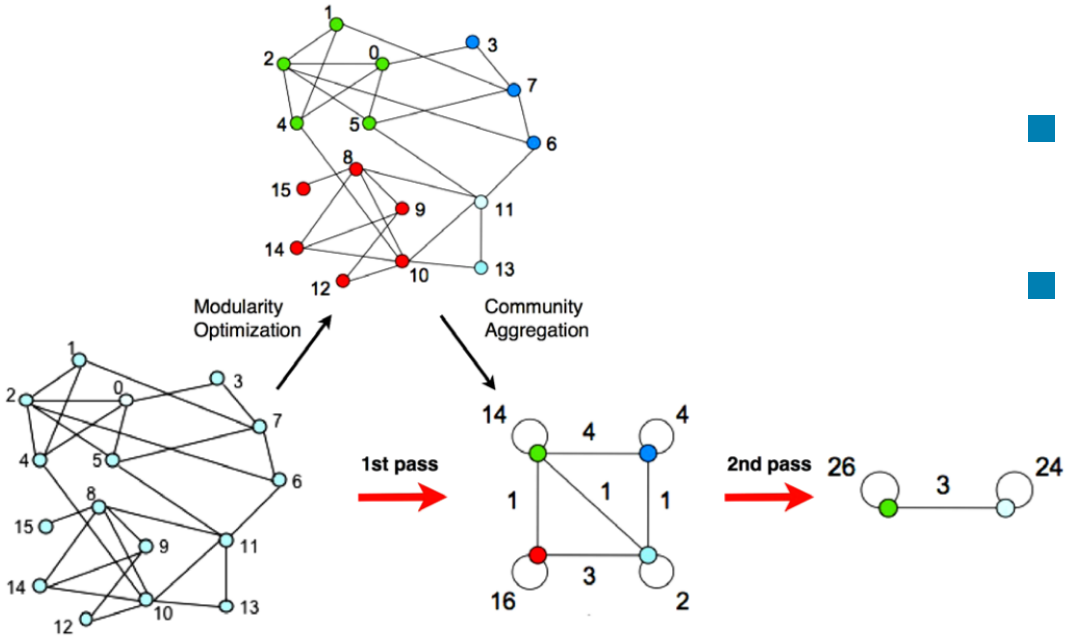
Find partition that maximises

$$\mathcal{Q} = \frac{1}{2m} \sum_c \Big(e_c \; - \; \langle e_c \rangle \Big).$$

- e_c: Actual number of edges in community c,
- $\langle e_c \rangle$: Expected number of edges in community c.
- Different "null models" possible:
  - Erdős-Rényi model: $\langle e_c \rangle = p\binom{n_c}{2}$.
  - Configuration (Chung-Lu) model: $\langle e_c \rangle = \frac{K_c^2}{2m}$.
  - ...

M Newman, M Girvan, *Phys Rev E* **69**, 026113, DOI 10.1103/PhysRevE.69.026113 (2004).

Louvain algorithm

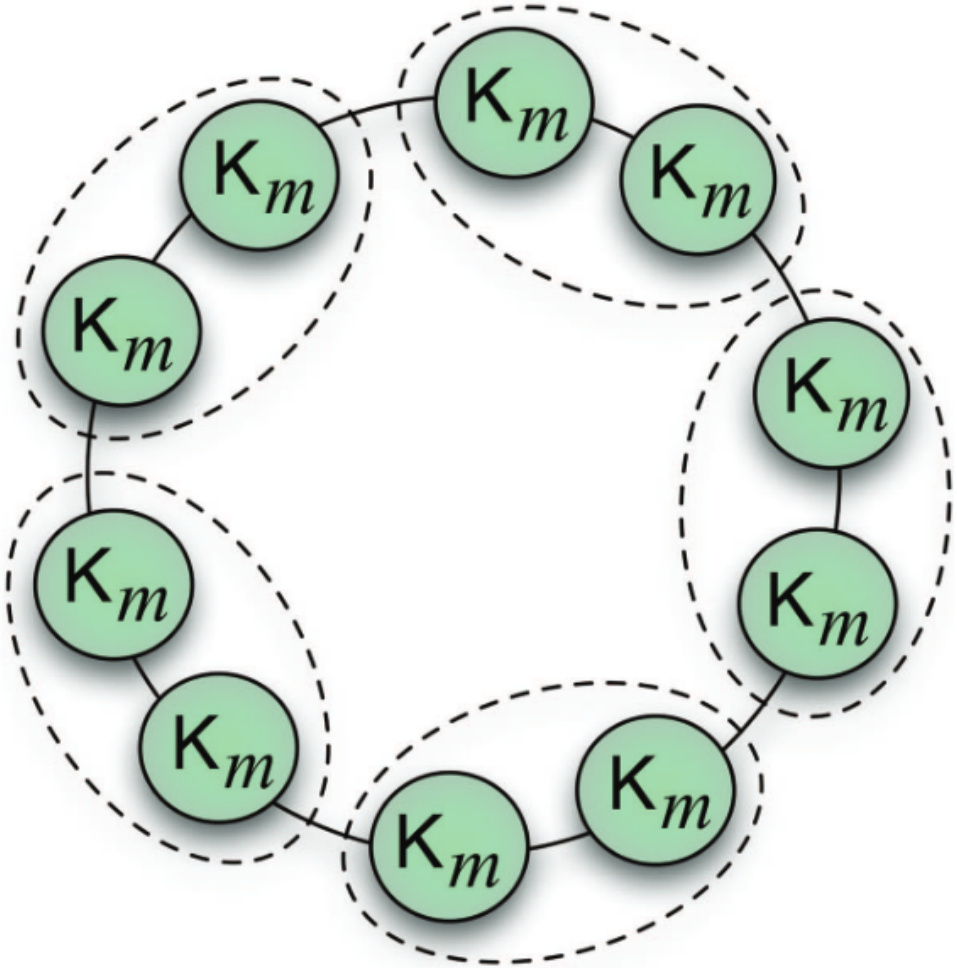


- Start with partition $c_i = i$ of graph G
- Repeat until convergence
  - For each node $v \in V(G)$
    - Move node $c_v \mapsto \arg\max_c \Delta\mathcal{Q}(c)$
  - Aggregate G based on c

- Complexity inner loop $O(m)$
- Number of loops unknown, estimated complexity $O(m \log m)$.

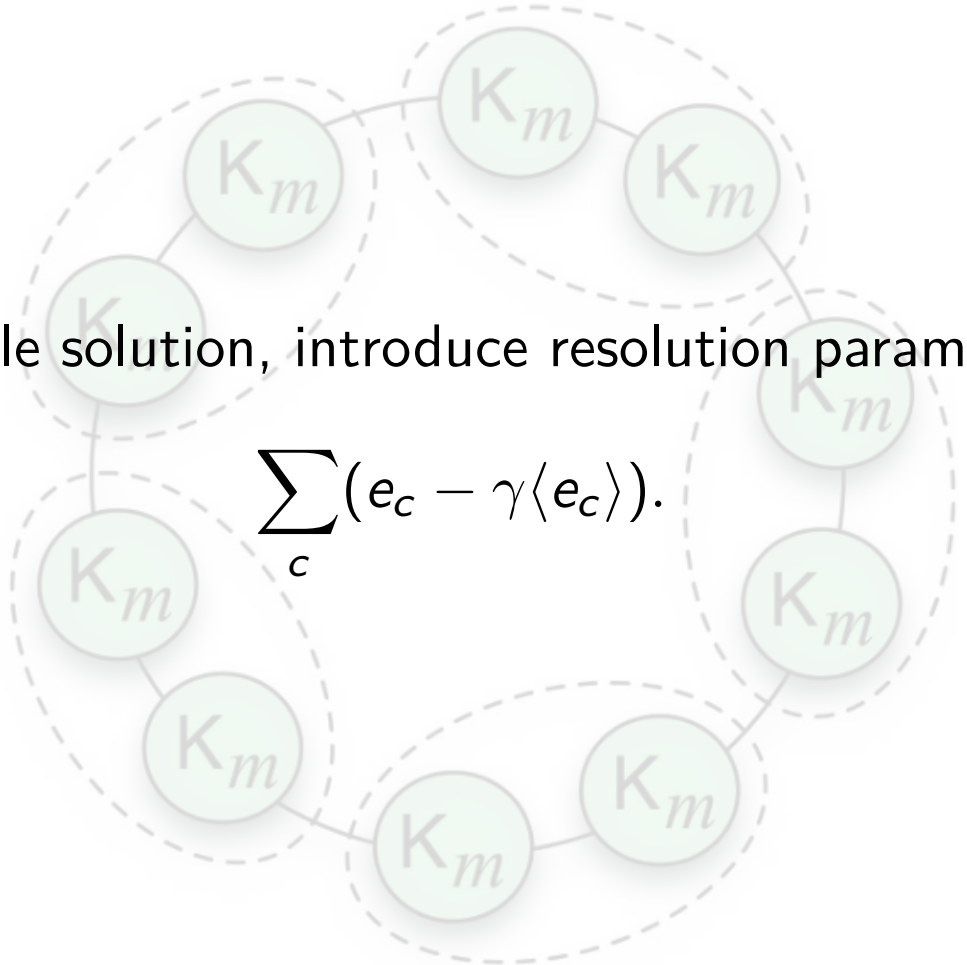
VD Blondel *et al.*, *J Stat Mech* **2008**, P10008, DOI 10.1088/1742-5468/2008/10/P10008 (2008).

Resolution limit



S Fortunato, M Barthélemy, *Proc Natl Acad Sci USA* **104**, 36–41, DOI 10.1073/pnas.0605965104 (2007).

Resolution limit

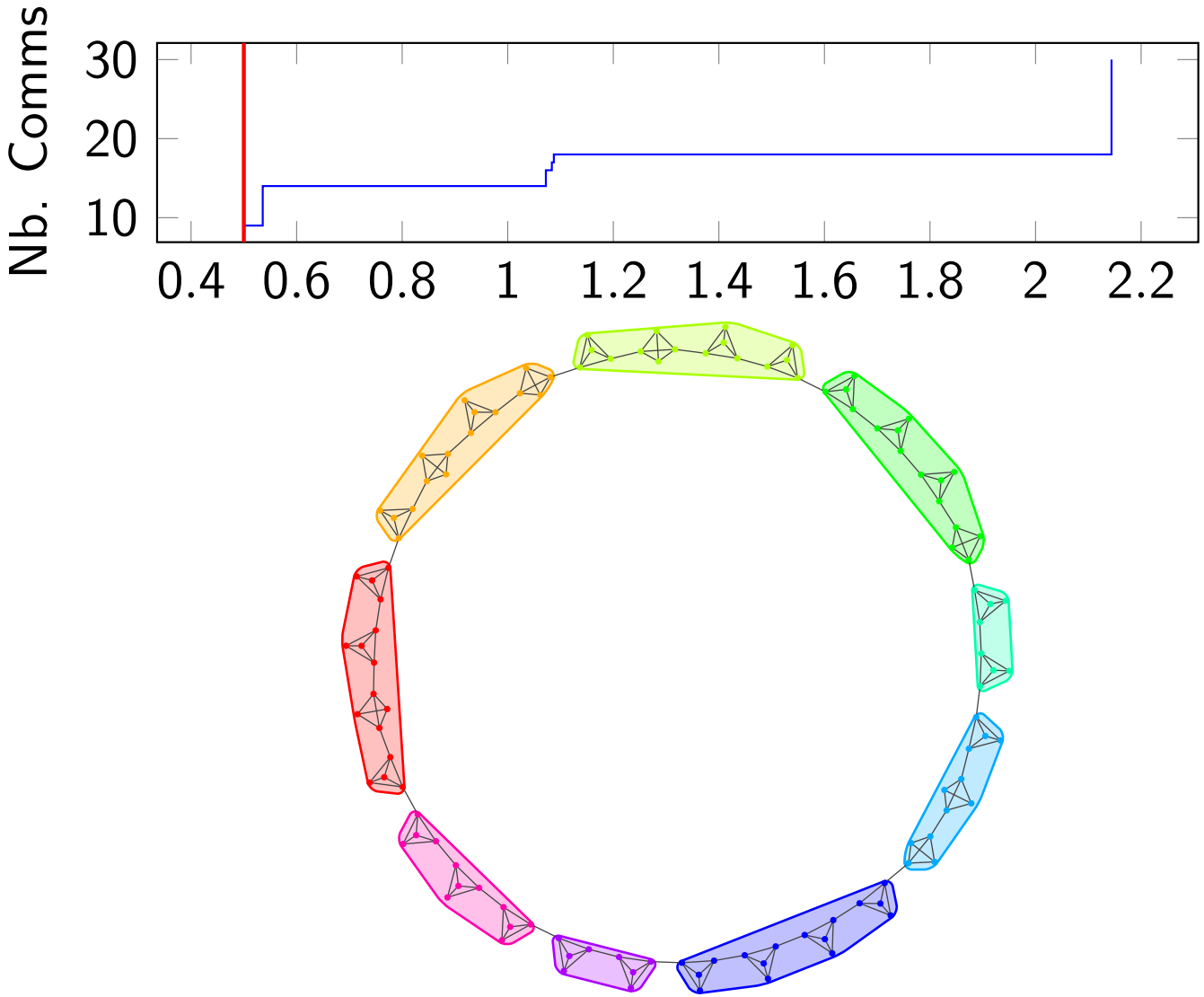


Possible solution, introduce resolution parameter:

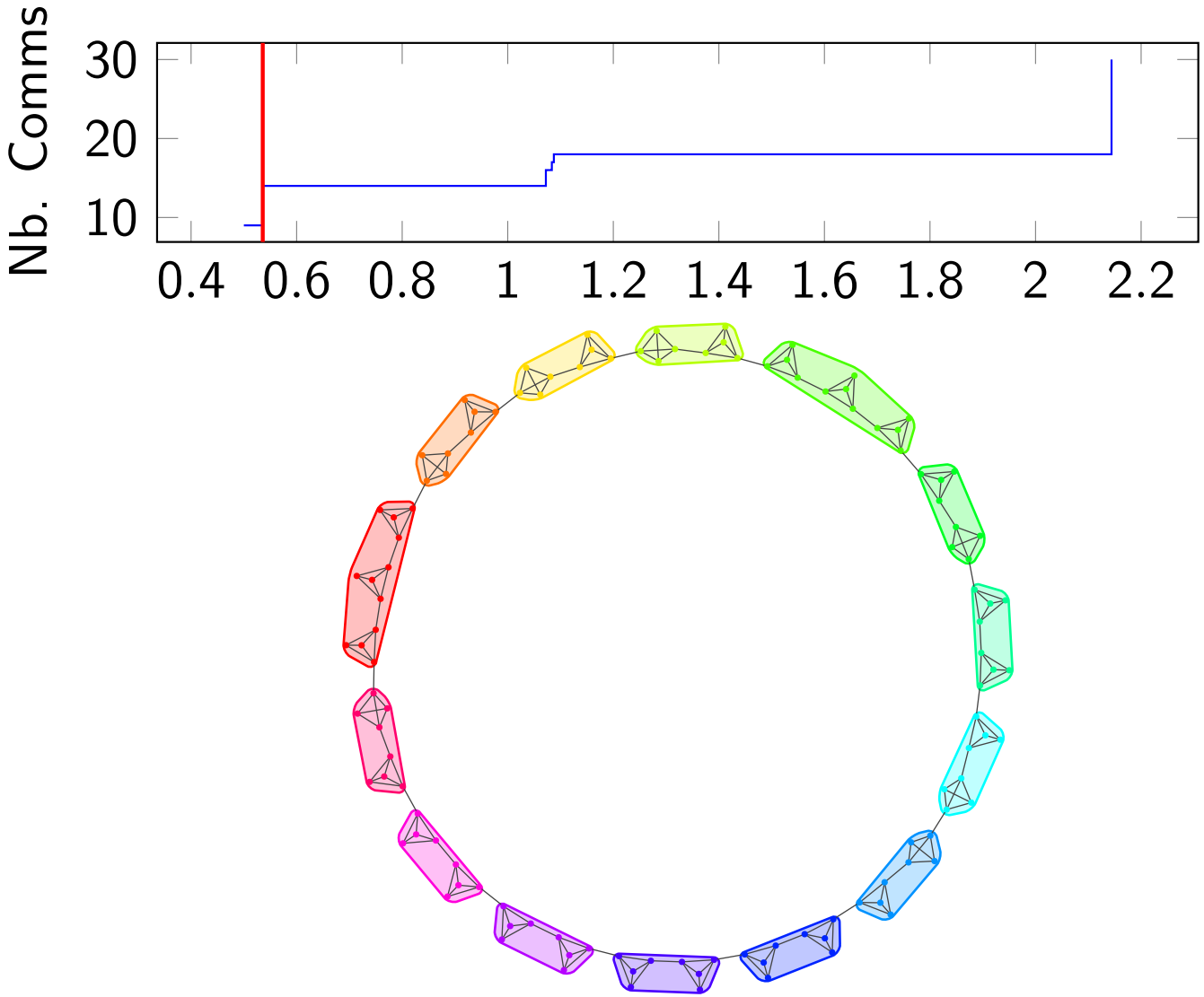
$$\sum_c (e_c - \gamma \langle e_c \rangle).$$

S Fortunato, M Barthélemy, *Proc Natl Acad Sci USA* **104**, 36–41, DOI 10.1073/pnas.0605965104 (2007).

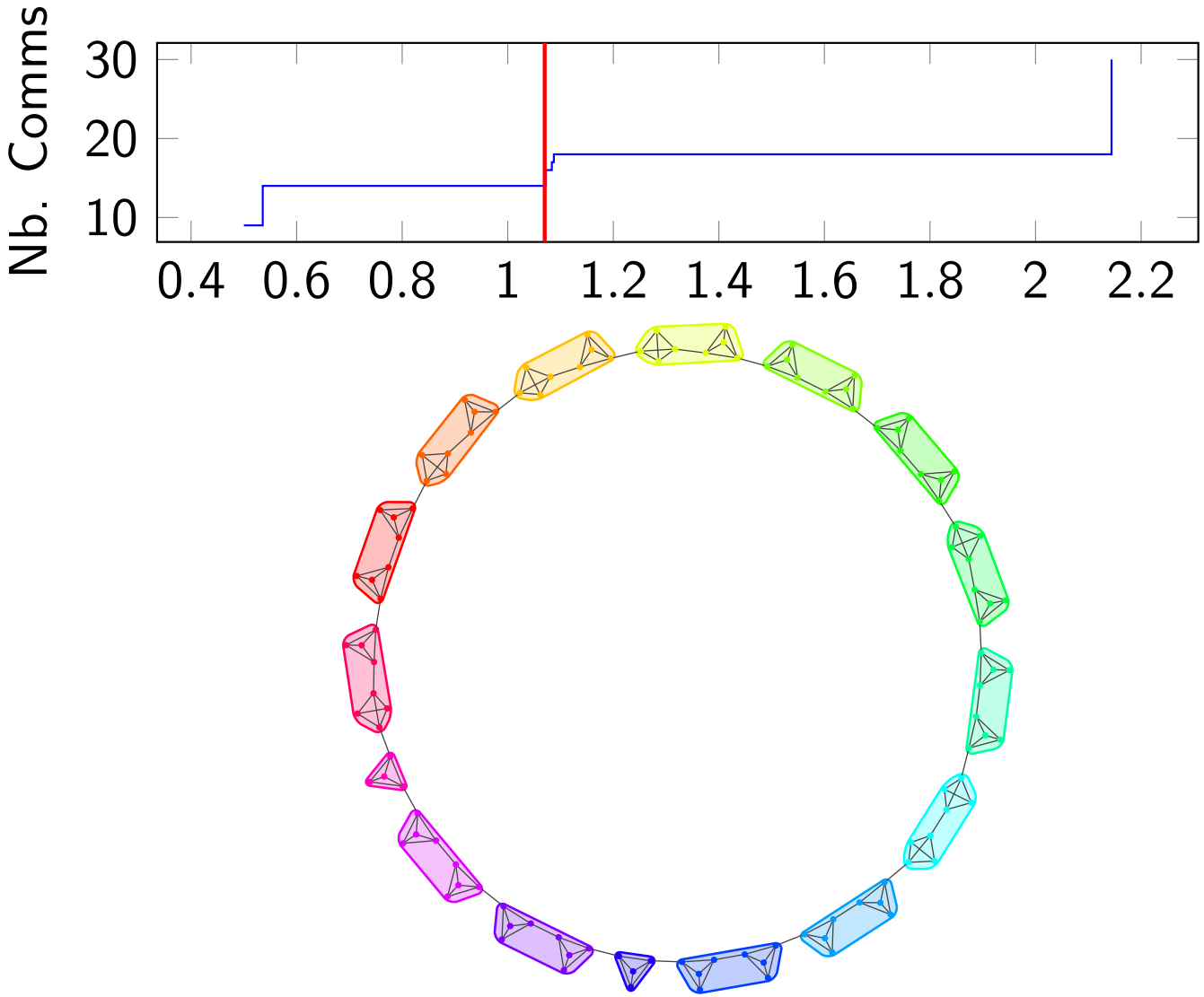
Resolution parameter



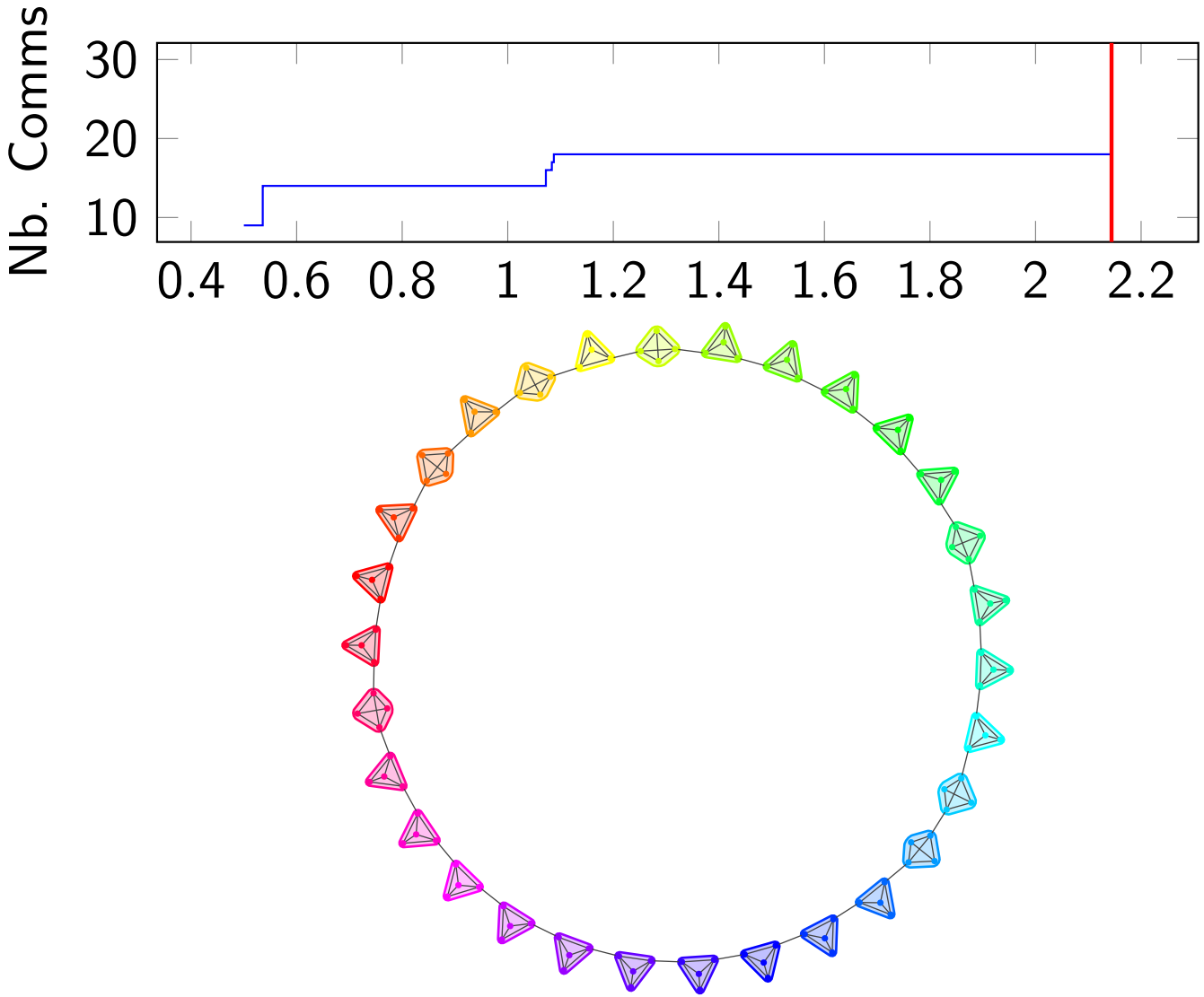
Resolution parameter



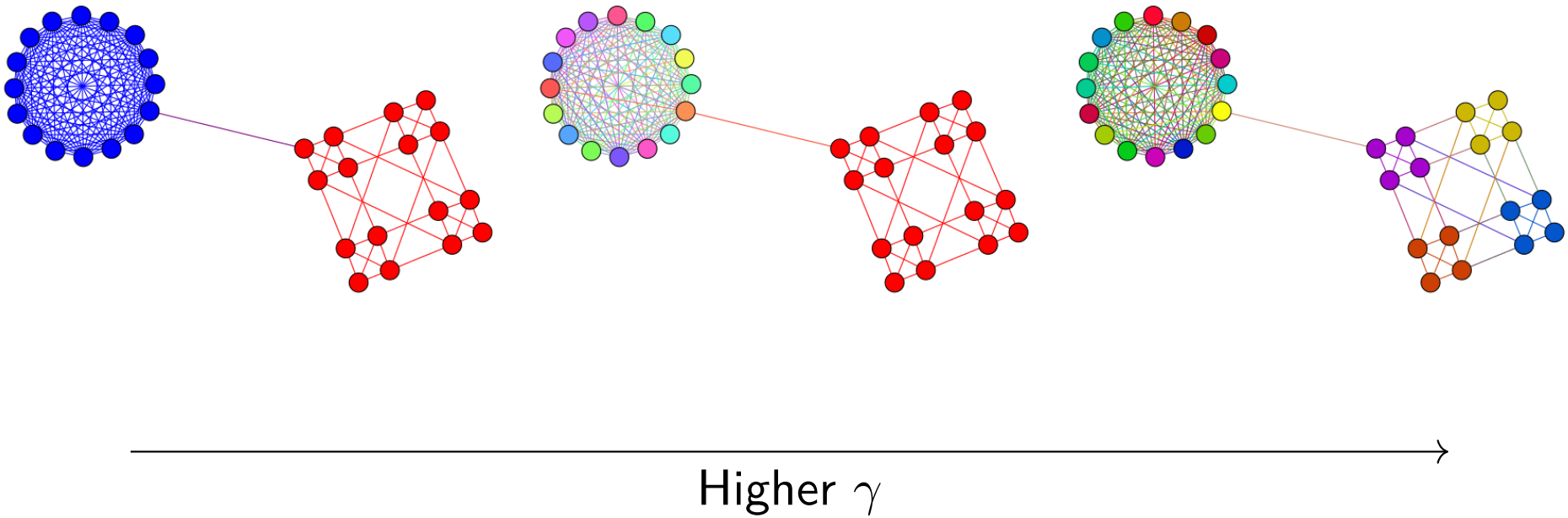
Resolution parameter



Resolution parameter



Upper resolution limit



G Krings, VD Blondel (2011).

Modularity in non-modular graphs

Modularity as sign of community structure

- Modularity $-1 \leq \mathcal{Q} \leq 1$.
- High modularity $\Rightarrow$ community structure?
- Modularity higher than 0.3 sometimes perceived as "significant".

Modularity in non-modular graphs

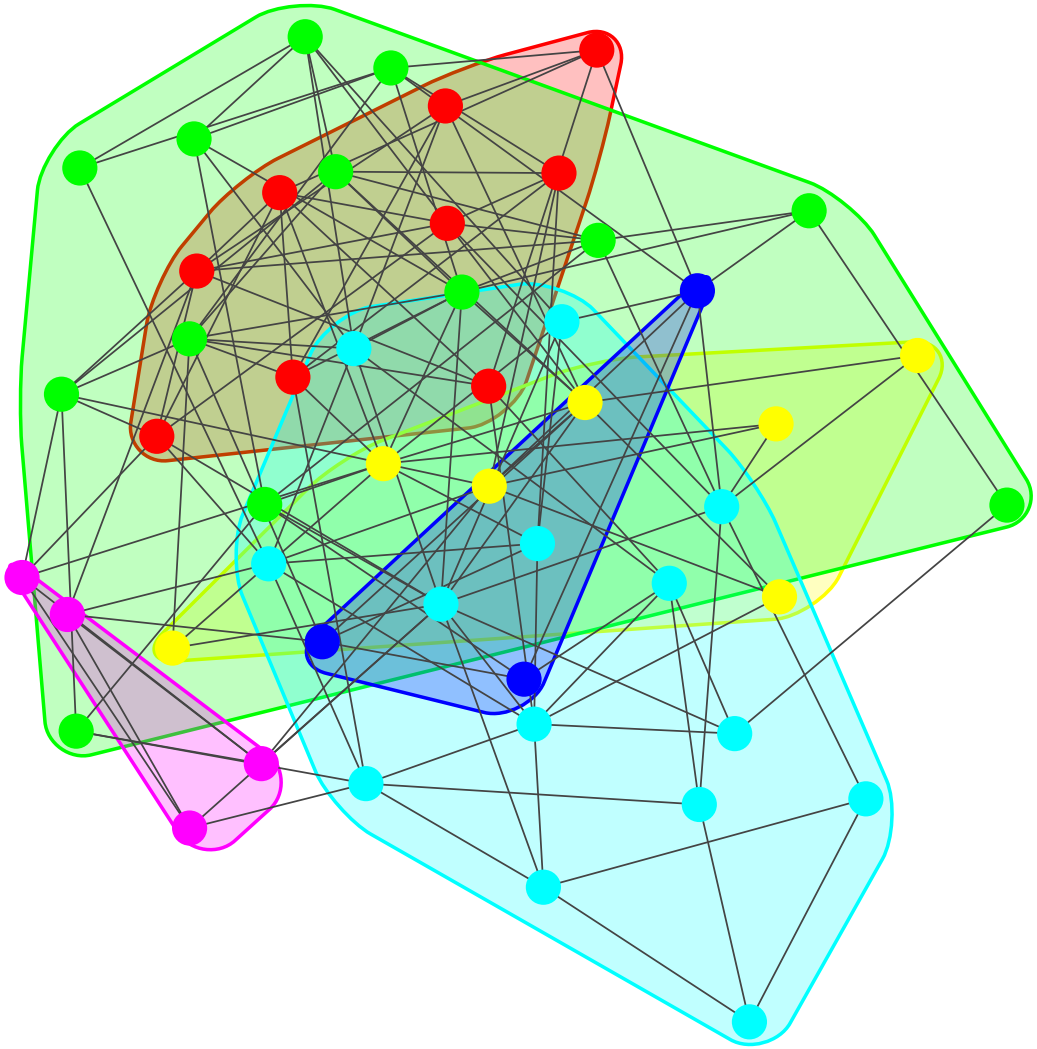
Modularity as sign of community structure

- Modularity $-1 < Q < 1$
- High modularity $\Rightarrow$ community structure?
- Modularity higher than expected can be viewed as "significant".

Many graphs have high modularity, but no community structure.

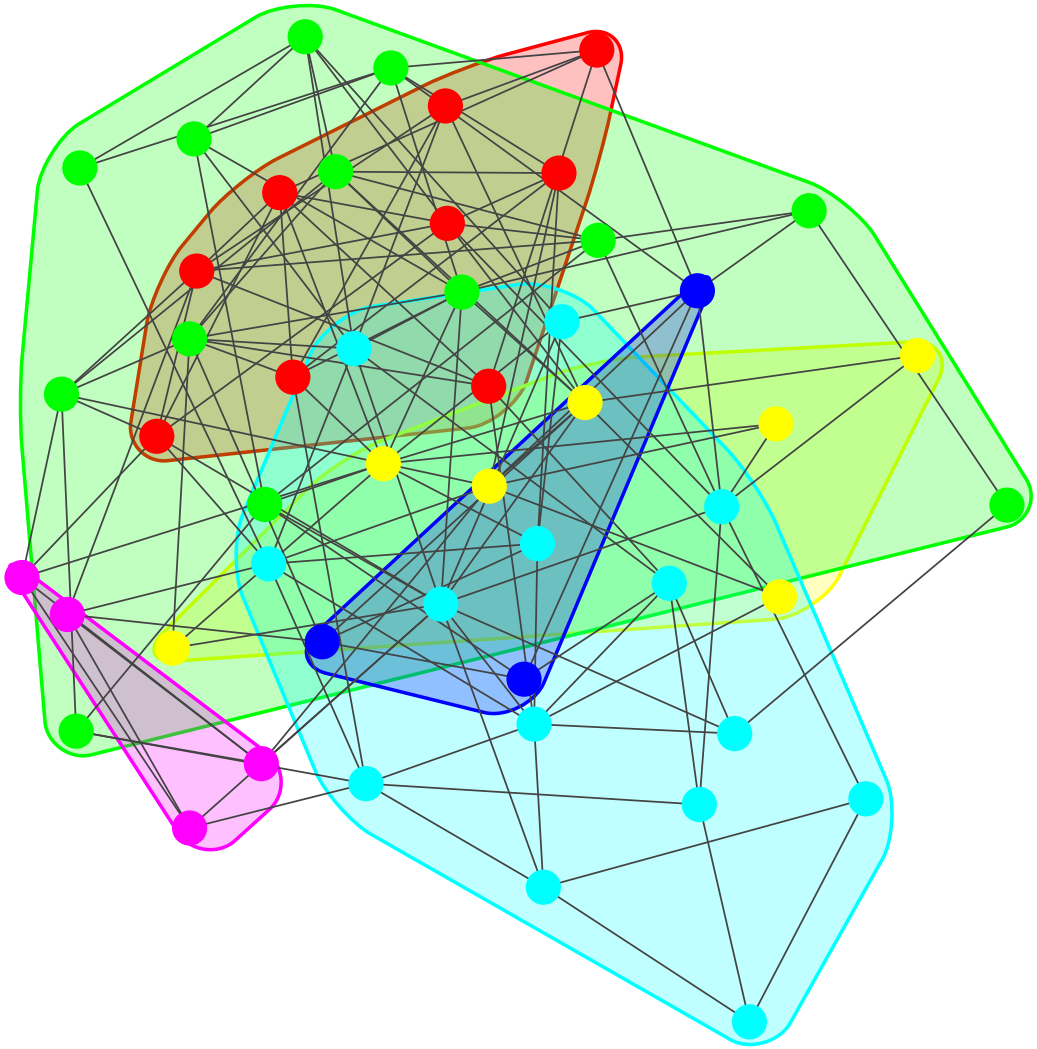
J Reichardt, S Bornholdt, *Physica D: Nonlinear Phenomena* **224**, 20–26, DOI 10.1016/j.physd.2006.09.009 (2006).

Modularity without community structure

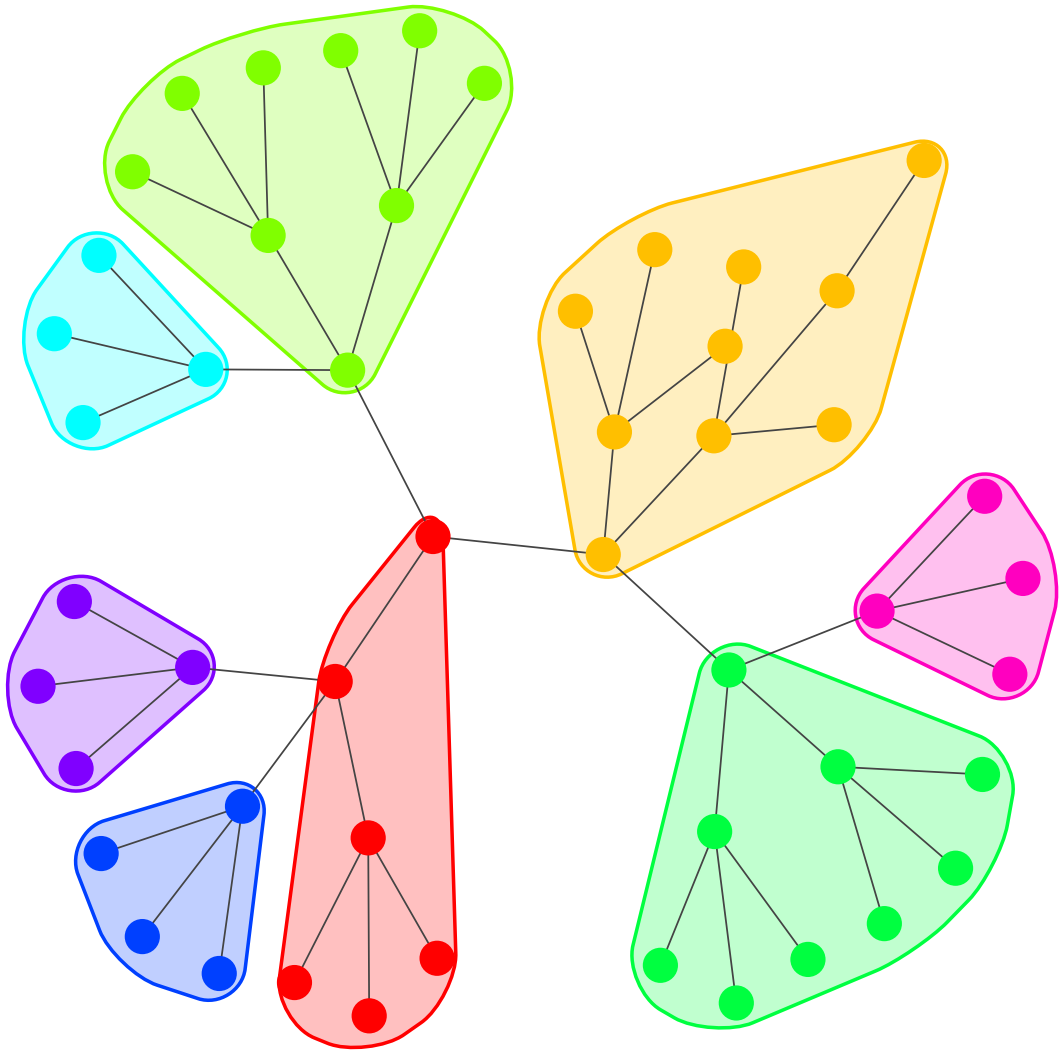


Modularity without community structure

$\mathcal{Q} = 0.31$

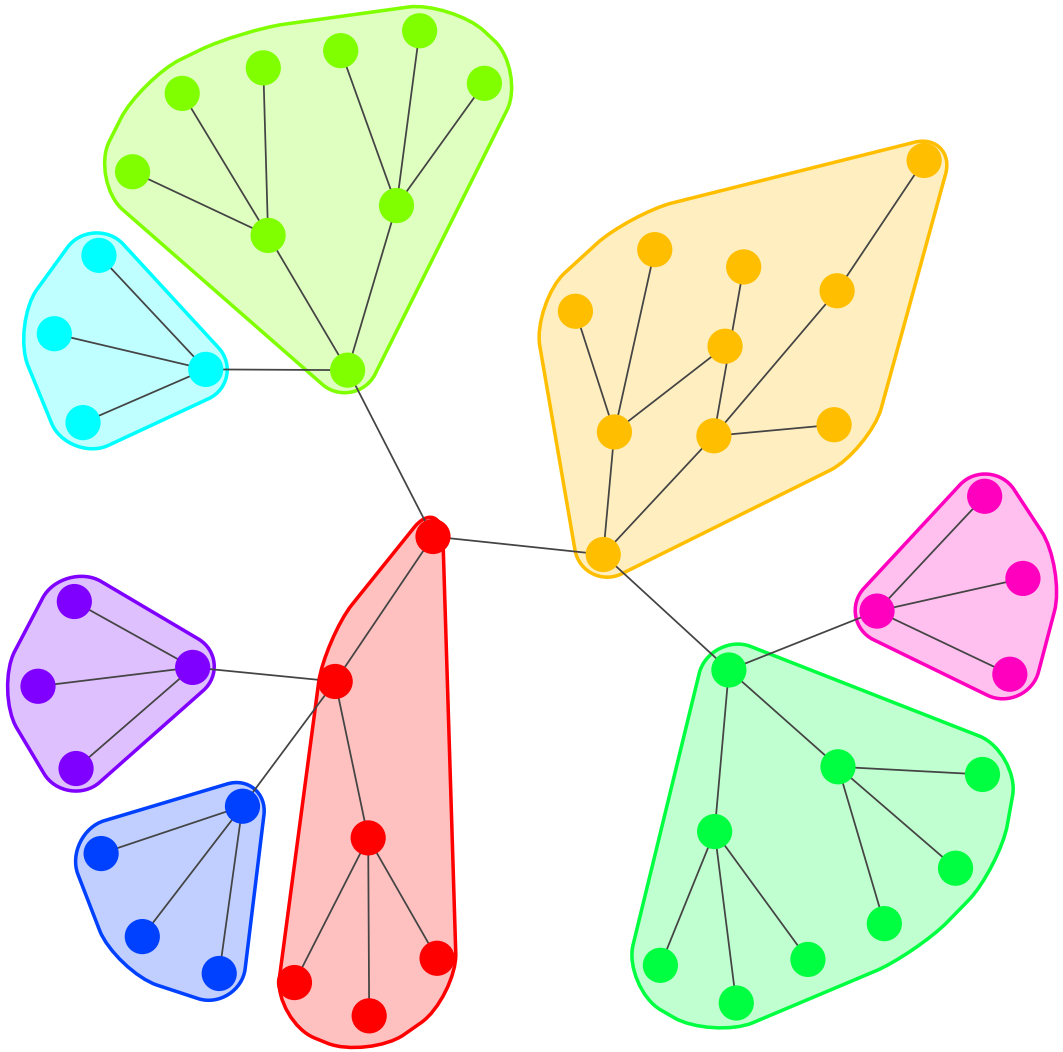


Modularity without community structure

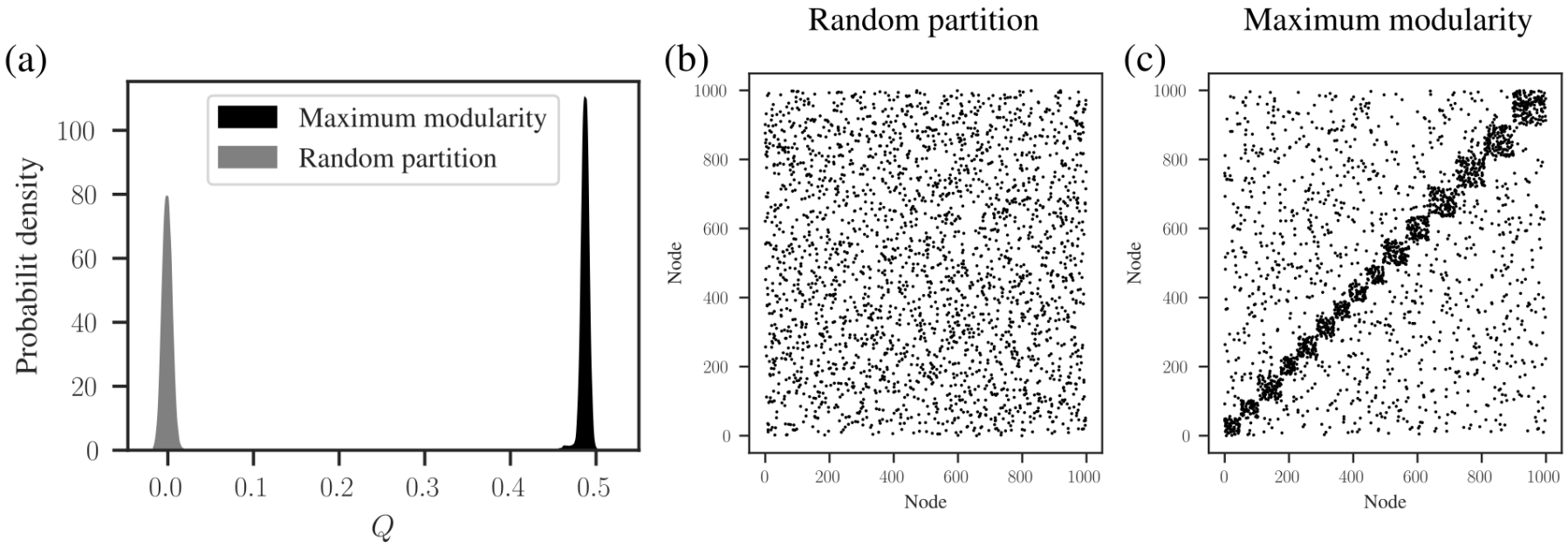


Modularity without community structure

$\mathcal{Q} = 0.71$



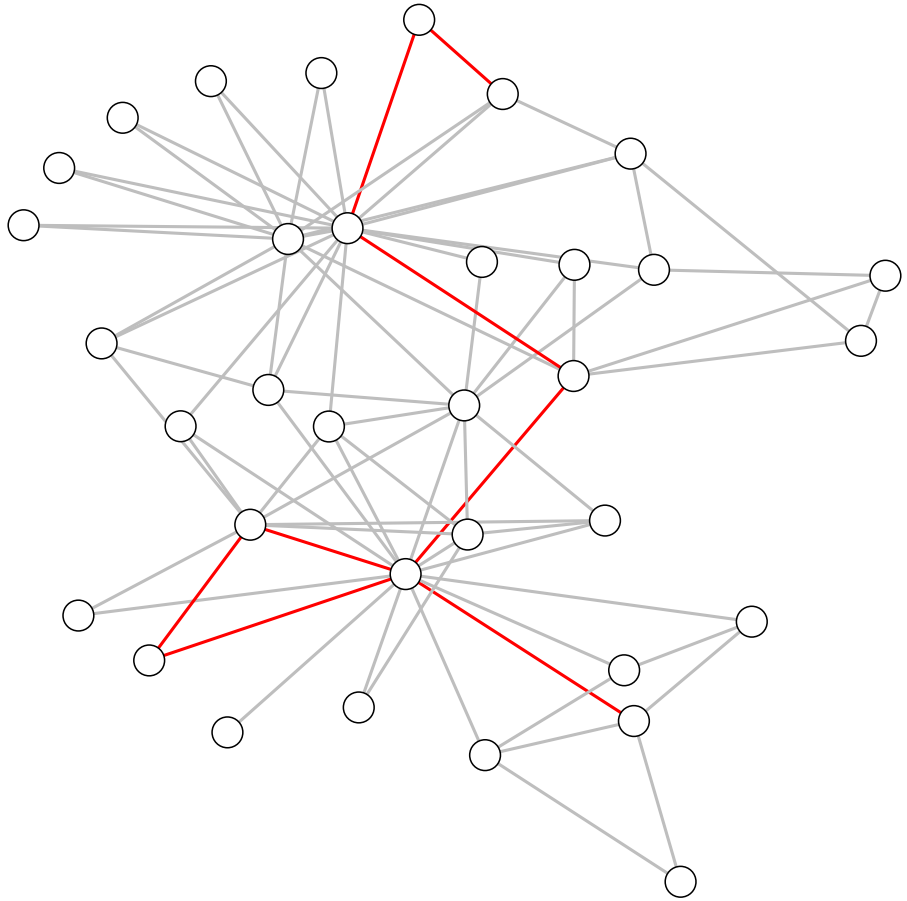
Modularity without community structure



TP Peixoto, *Elements in the Structure and Dynamics of Complex Networks*, DOI 10.1017/9781009118897 (2023).

Dynamical perspective

Random walk on network



$$p_{t+1} = p_t D^{-1} A$$

Two approaches

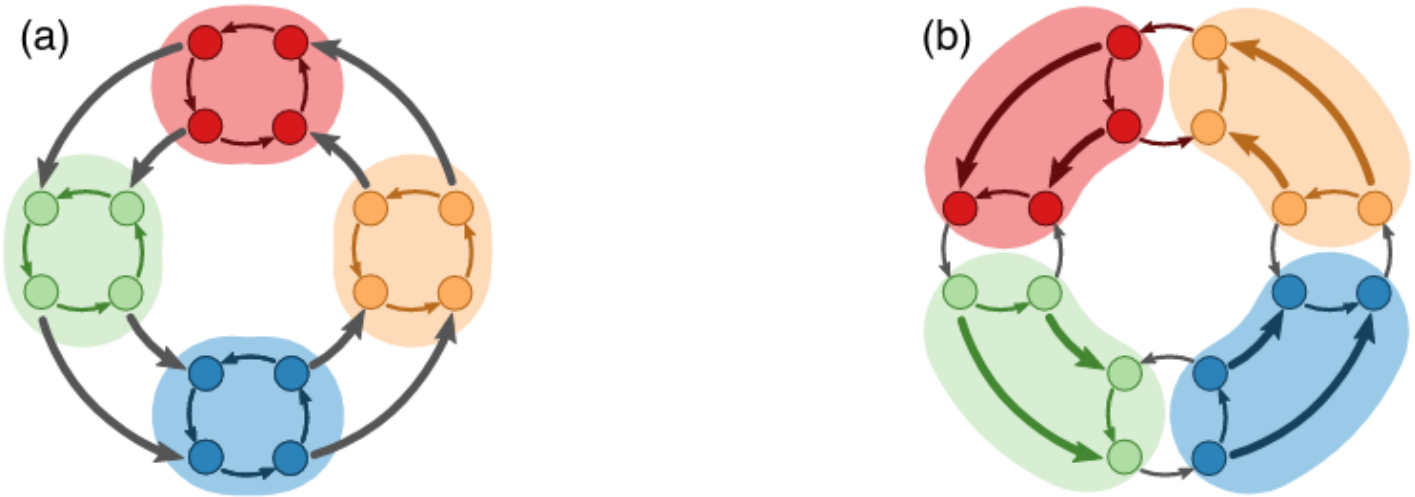
Markov stability

- Clustered autocovariance of random walk.
- Probability of random walk to stay within a cluster.

Infomap

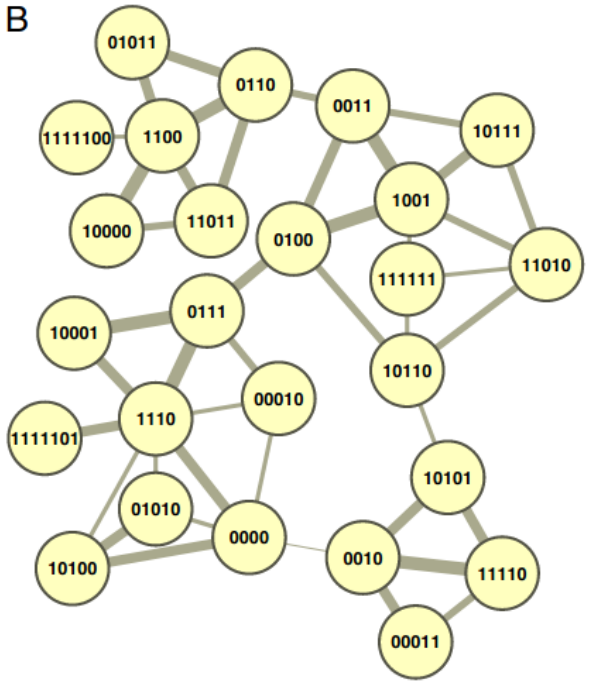
- Compression of random walk.
- Good clusters compress better.

Dynamics vs structure



M Rosvall *et al.*, in *Advances in Network Clustering and Blockmodeling* (John Wiley & Sons, Ltd, 2019), pp. 105–119, DOI 10.1002/9781119483298.ch4.

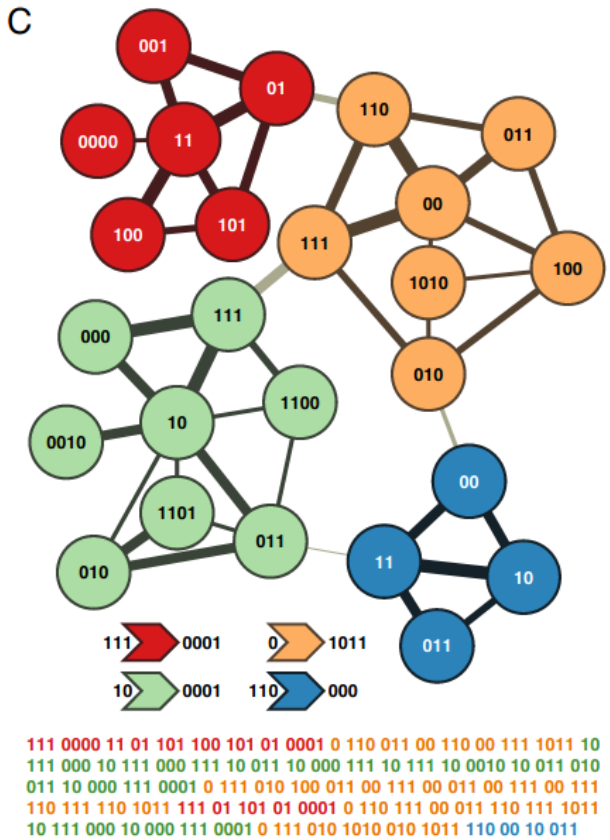
Infomap compression



1111100 1100 0110 11011 10000 11011 0110 0011 10111 1001 0011 1001 0100 0111 10001 1110 0111 10001 0111 1110 0000 1110 10001 0111 1110 0111 1110 1111101 1110 0000 10100 0000 1110 10001 0111 0100 10110 11010 10111 1001 0100 1001 10111 1001 0100 1001 0100 0011 0100 0011 0110 11011 0110 0011 0100 1001 10111 0011 0100 0111 10001 1110 10001 0111 0100 10110 111111 10110 10101 11110 00011

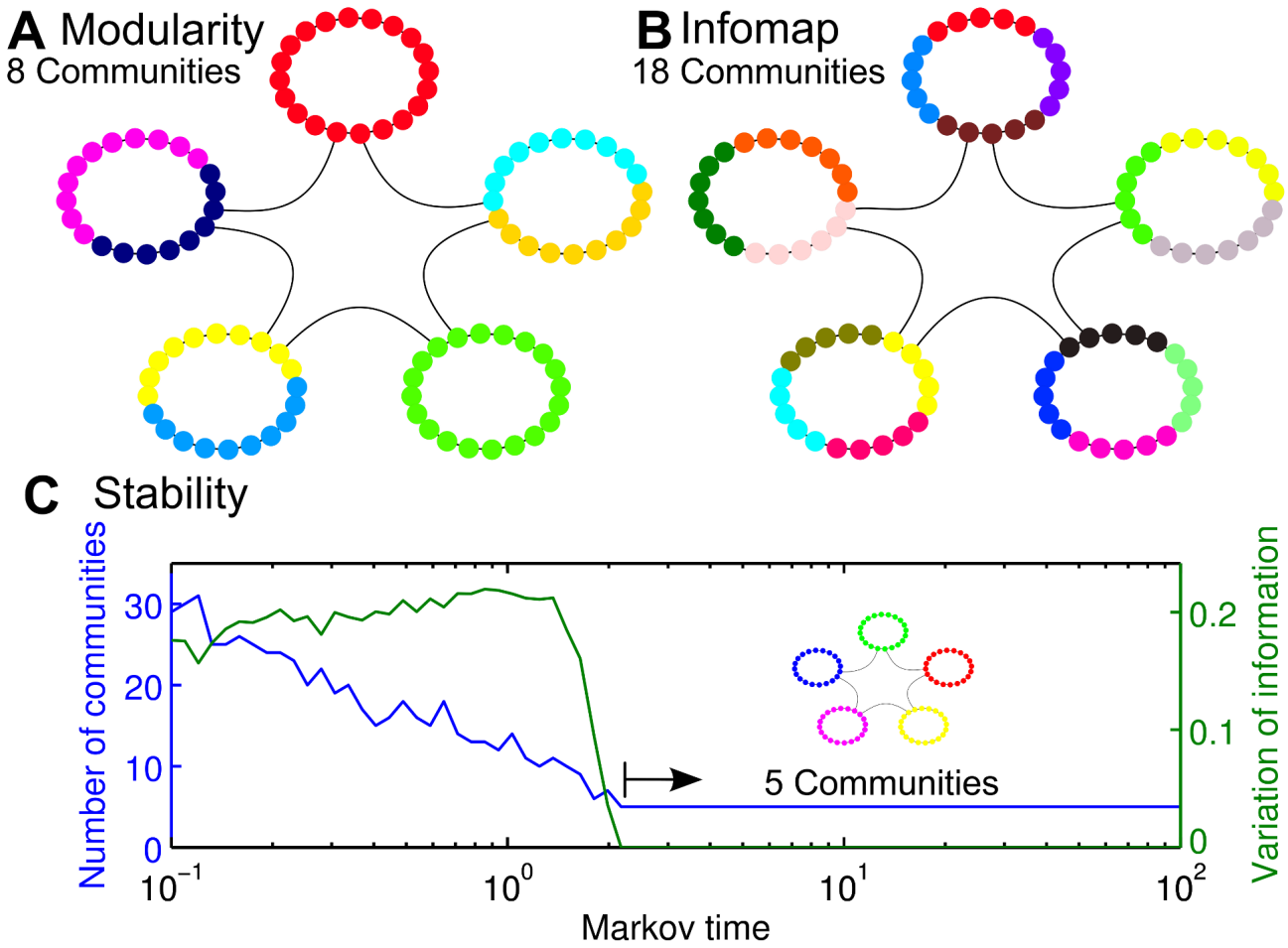
M Rosvall *et al.*, in *Advances in Network Clustering and Blockmodeling* (John Wiley & Sons, Ltd, 2019), pp. 105–119, DOI 10.1002/9781119483298.ch4.

Infomap compression



M Rosvall *et al.*, in *Advances in Network Clustering and Blockmodeling* (John Wiley & Sons, Ltd, 2019), pp. 105–119, DOI 10.1002/9781119483298.ch4.

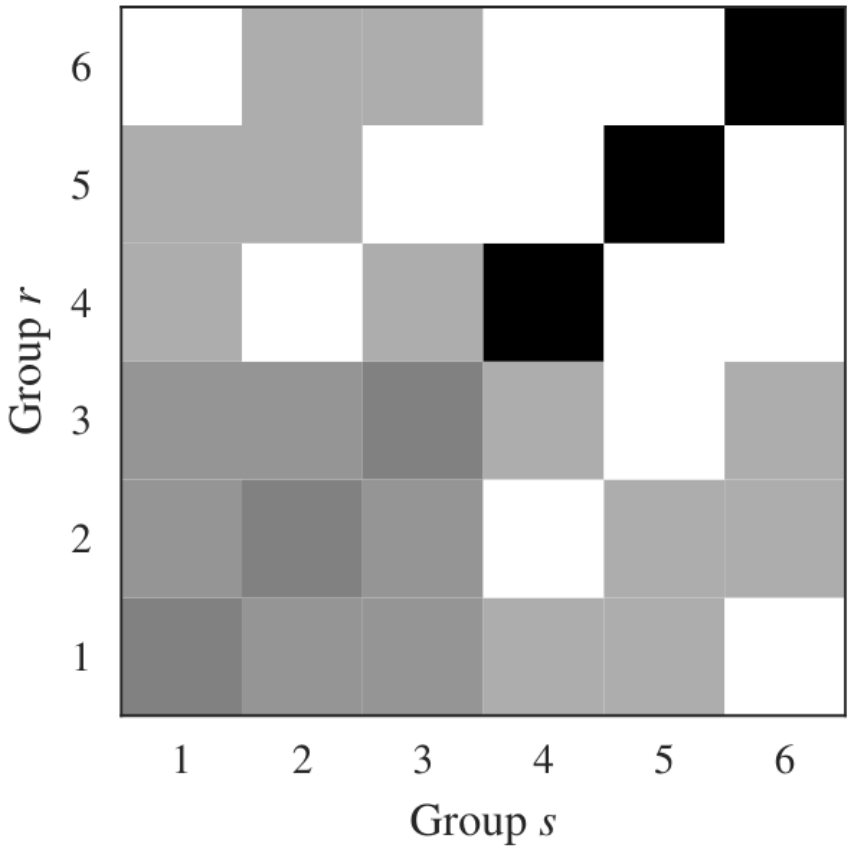
Field of view limit

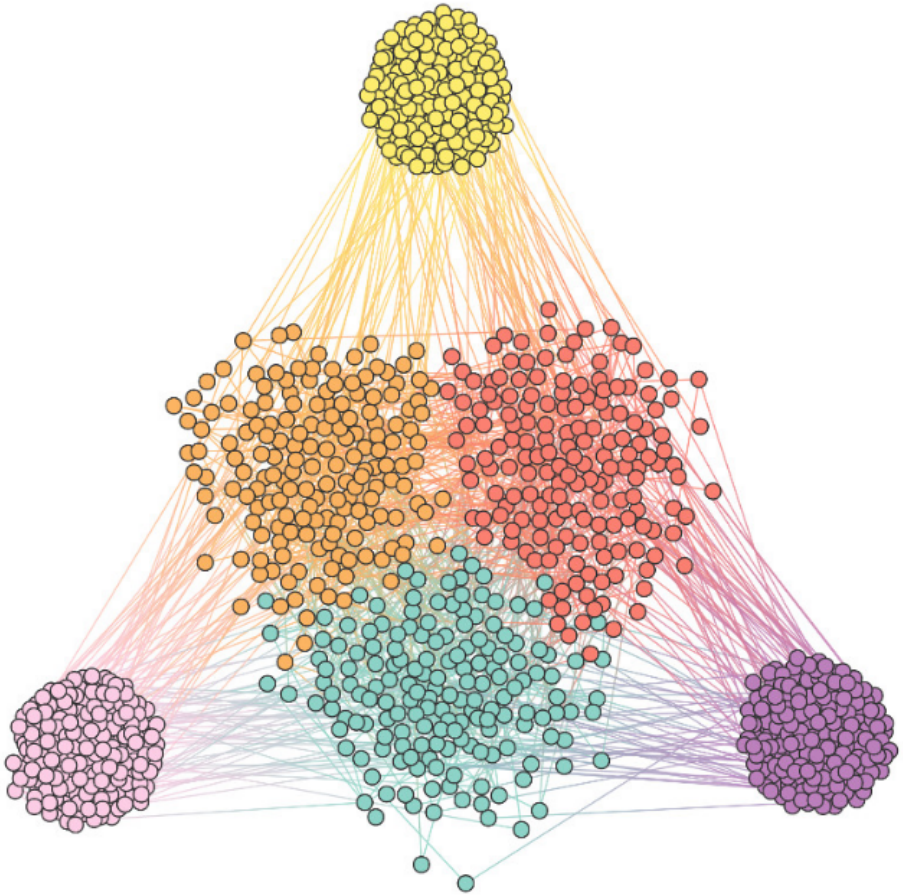


MT Schaub *et al.*, *PLoS ONE* **7**, e32210, DOI 10.1371/journal.pone.0032210 (2012).

Inferential perspective

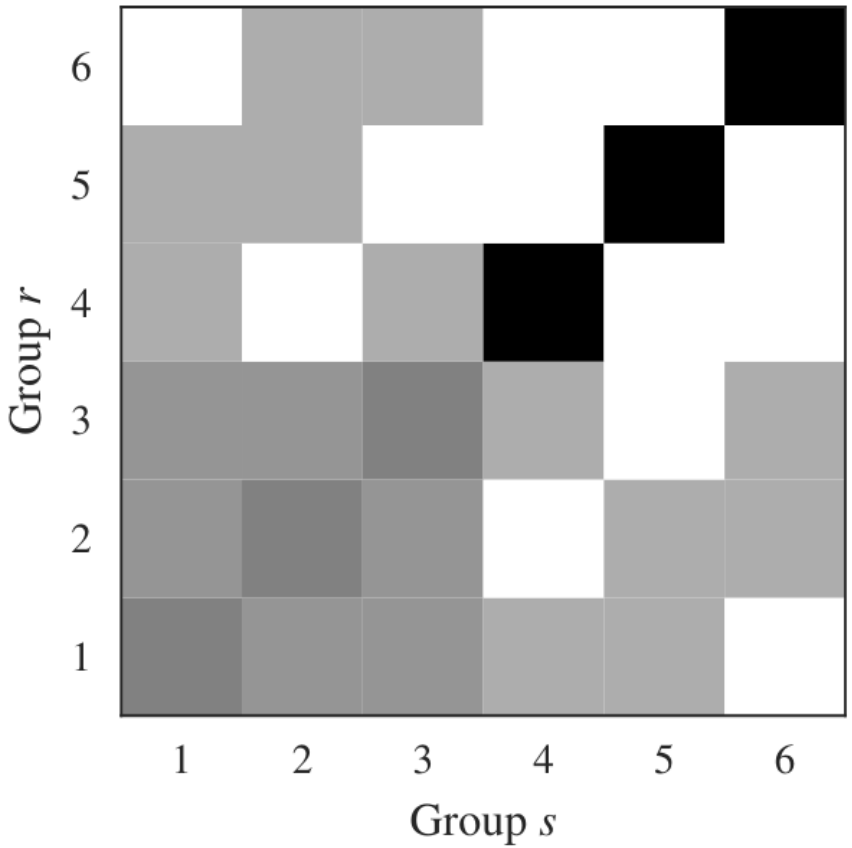
Stochastic Block Model

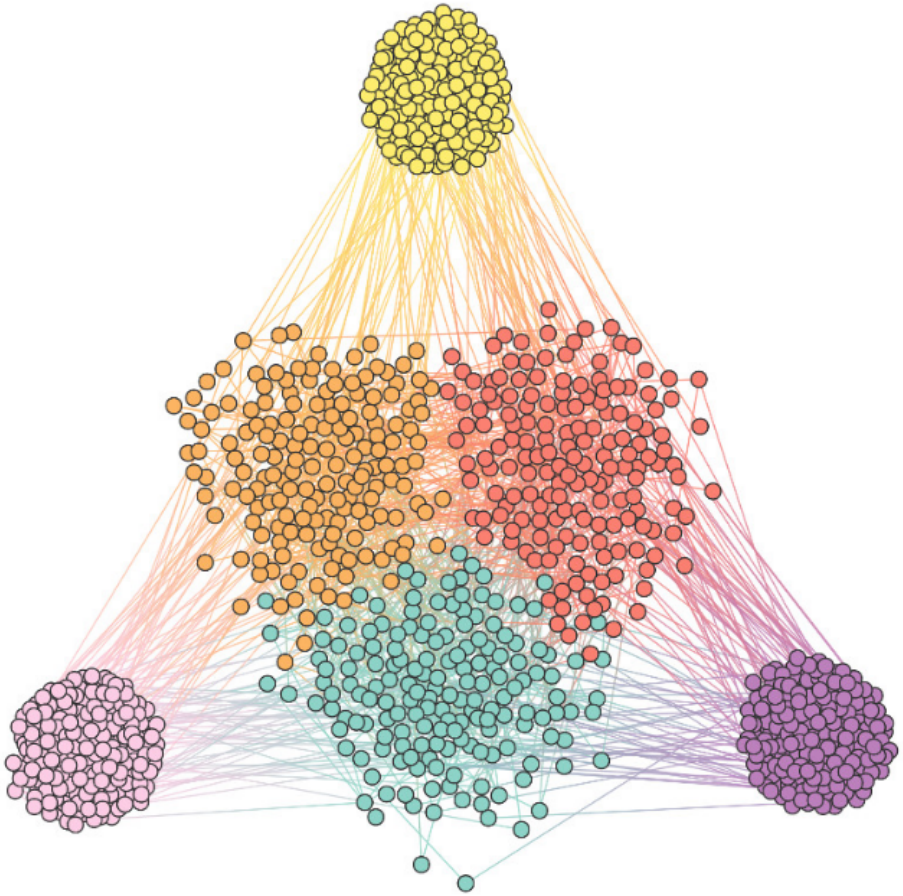




$$A_{ij} \sim \text{Bernoulli}(p_{b_i b_j})$$

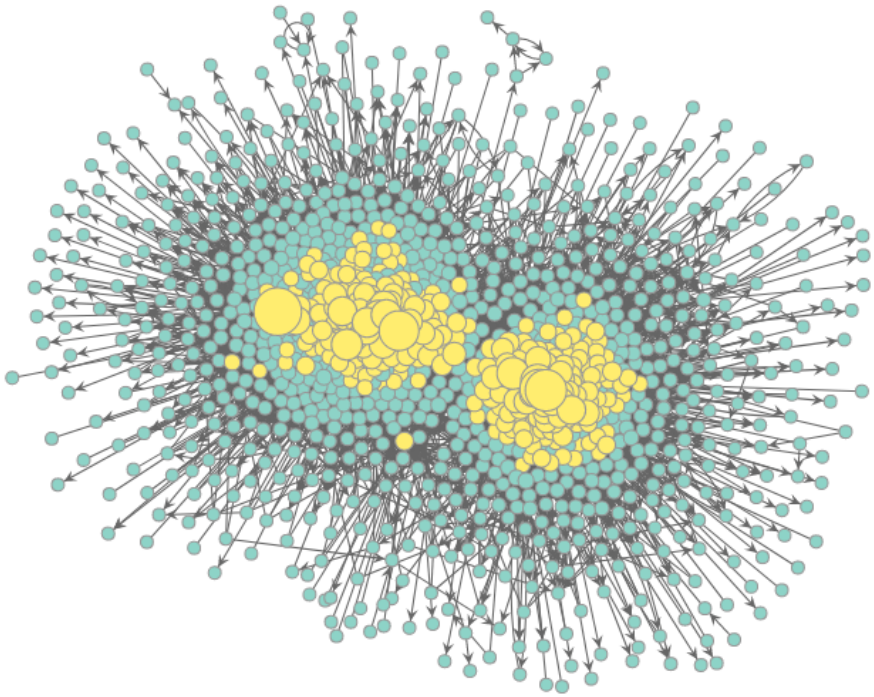
Stochastic Block Model

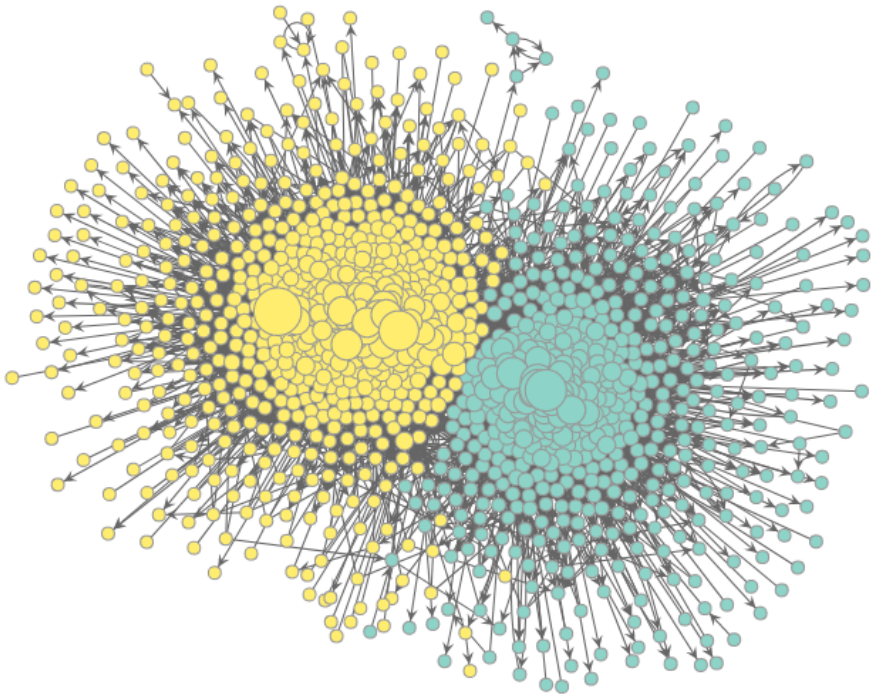




$$A_{ij} \sim \text{Poisson}(\omega_{b_i b_j})$$

Stochastic Block Model - Degree Correction





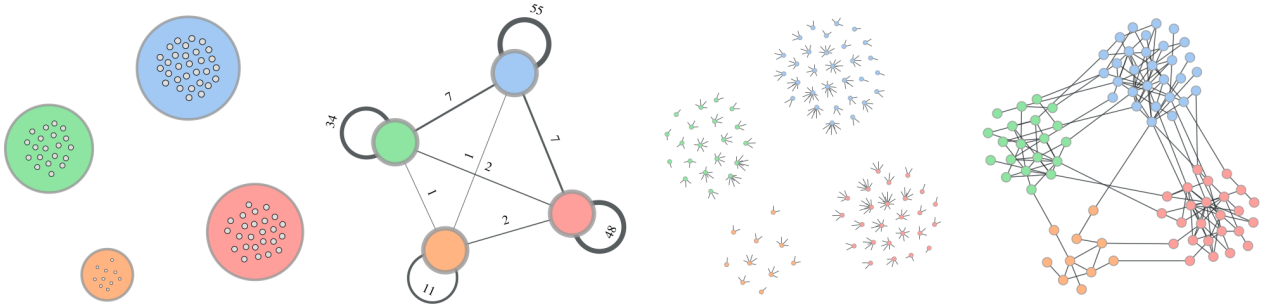
$$A_{ij} \sim \text{Poisson}(\omega_{b_i b_j})$$

$$A_{ij} \sim \text{Poisson}(\theta_i \theta_j \omega_{b_i b_j})$$

TP Peixoto, in *Advances in Network Clustering and Blockmodeling* (John Wiley & Sons, Ltd, 2019), pp. 289–332, DOI 10.1002/9781119483298.ch11.

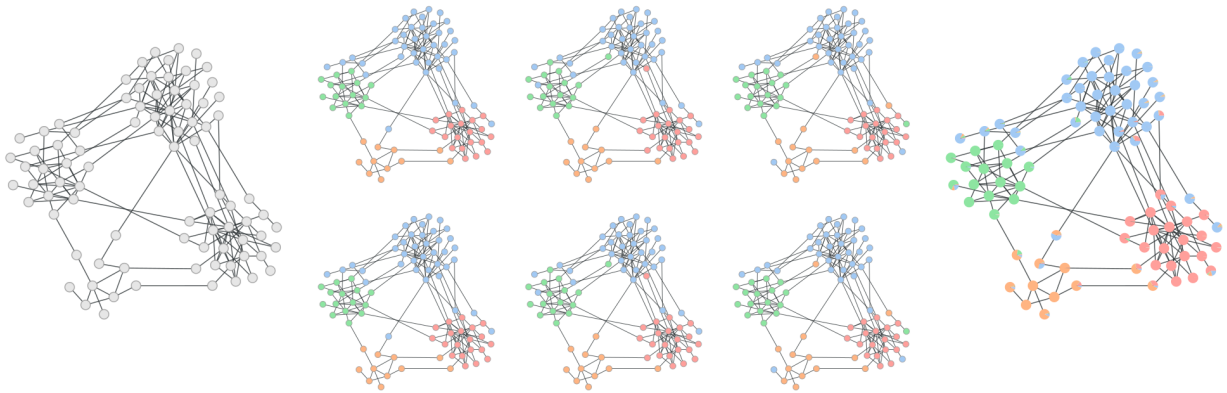
Generative approach

(a) Generative process



Node partition, $P(\boldsymbol{b})$ | Edges between groups, $P(\boldsymbol{e}|\boldsymbol{b})$ | Degrees, $P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})$ | Network, $P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$

(b) Inference procedure



Observed network $\boldsymbol{A}$ | Posterior distribution $P(\boldsymbol{b}|\boldsymbol{A})$ | Marginal probabilities

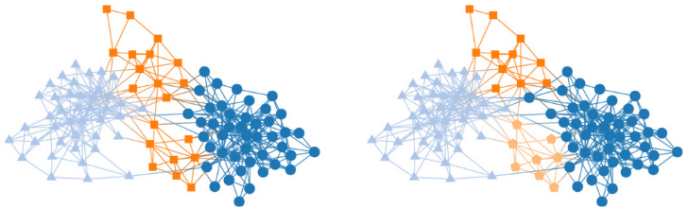
TP Peixoto, *Elements in the Structure and Dynamics of Complex Networks*, DOI 10.1017/9781009118897 (2023).

Planted Partition

(a) Nested DC-SBM
$\Sigma = 1780.58$ (nats)

(b) PP (uniform)
$\Sigma = 1761.50$ (nats)

Political books



(c) Nested DC-SBM
$\Sigma = 1343.44$ (nats)

(d) PP (non-uniform)
$\Sigma = 1337.69$ (nats)

$$A_{ij} \sim \text{Poisson}(\theta_i \theta_j \omega_{b_i b_j})$$

$$\omega_{rs} = \omega_{\text{in}} \delta_{rs} + \omega_{\text{out}}(1 - \delta_{rs})$$

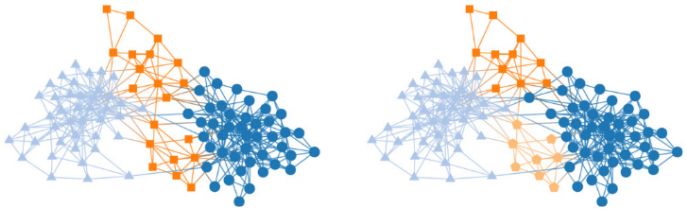
L Zhang, TP Peixoto, *Phys Rev Res* **2**, 043271, DOI 10.1103/PhysRevResearch.2.043271 (2020).

Planted Partition

(a) Nested DC-SBM
$\Sigma = 1780.58$ (nats)

(b) PP (uniform)
$\Sigma = 1761.50$ (nats)

Political books



(c) Nested DC-SBM
$\Sigma = 1343.44$ (nats)

(d) PP (non-uniform)
$\Sigma = 1337.69$ (nats)

$$A_{ij} \sim \text{Poisson}(\theta_i \theta_j \omega_{b_i b_j})$$

$$\omega_{rs} = \omega_r \delta_{rs} + \omega_{\text{out}}(1 - \delta_{rs})$$

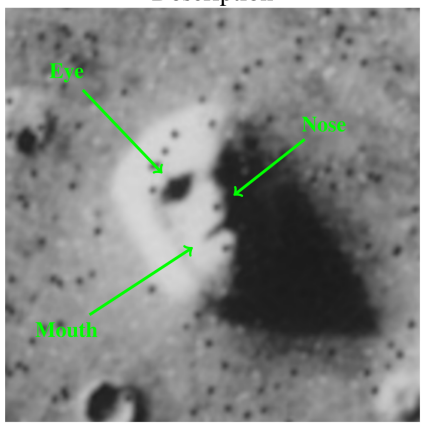
L Zhang, TP Peixoto, *Phys Rev Res* **2**, 043271, DOI 10.1103/PhysRevResearch.2.043271 (2020).

Descriptive vs inferential

TP Peixoto, *Elements in the Structure and Dynamics of Complex Networks*, DOI 10.1017/9781009118897 (2023).

Descriptive vs inferential

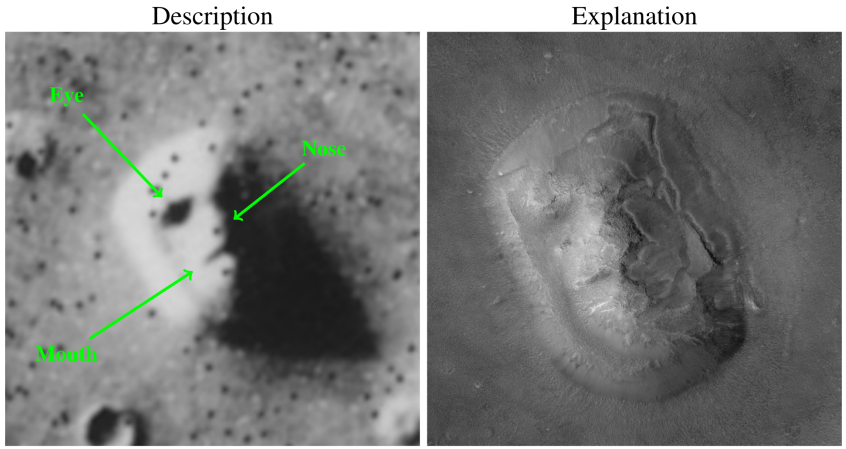
Description



A face

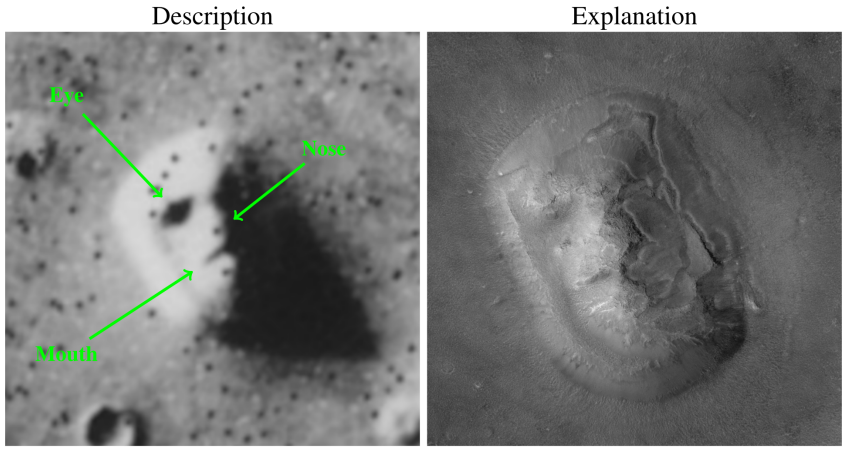
TP Peixoto, *Elements in the Structure and Dynamics of Complex Networks*, DOI 10.1017/9781009118897 (2023).

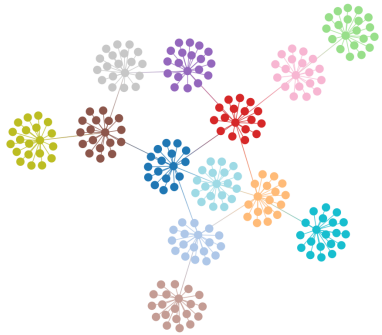
Descriptive vs inferential



TP Peixoto, *Elements in the Structure and Dynamics of Complex Networks*, DOI 10.1017/9781009118897 (2023).

Descriptive vs inferential

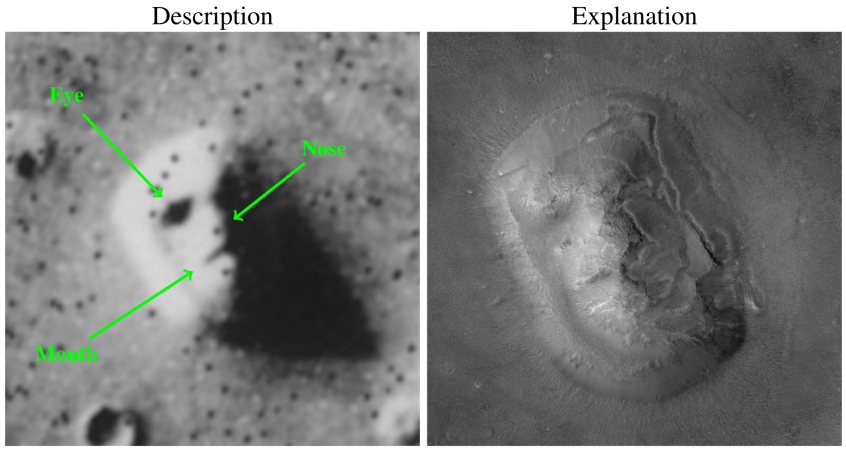




A network with 13 communities

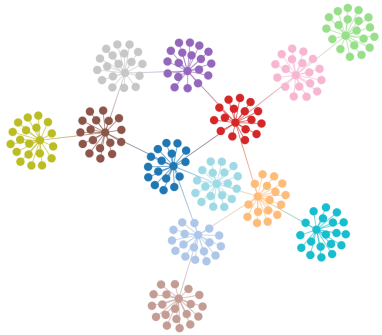
TP Peixoto, *Elements in the Structure and Dynamics of Complex Networks*, DOI 10.1017/9781009118897 (2023).

Descriptive vs inferential

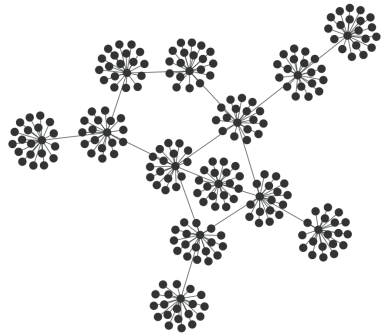


A face

A mountain



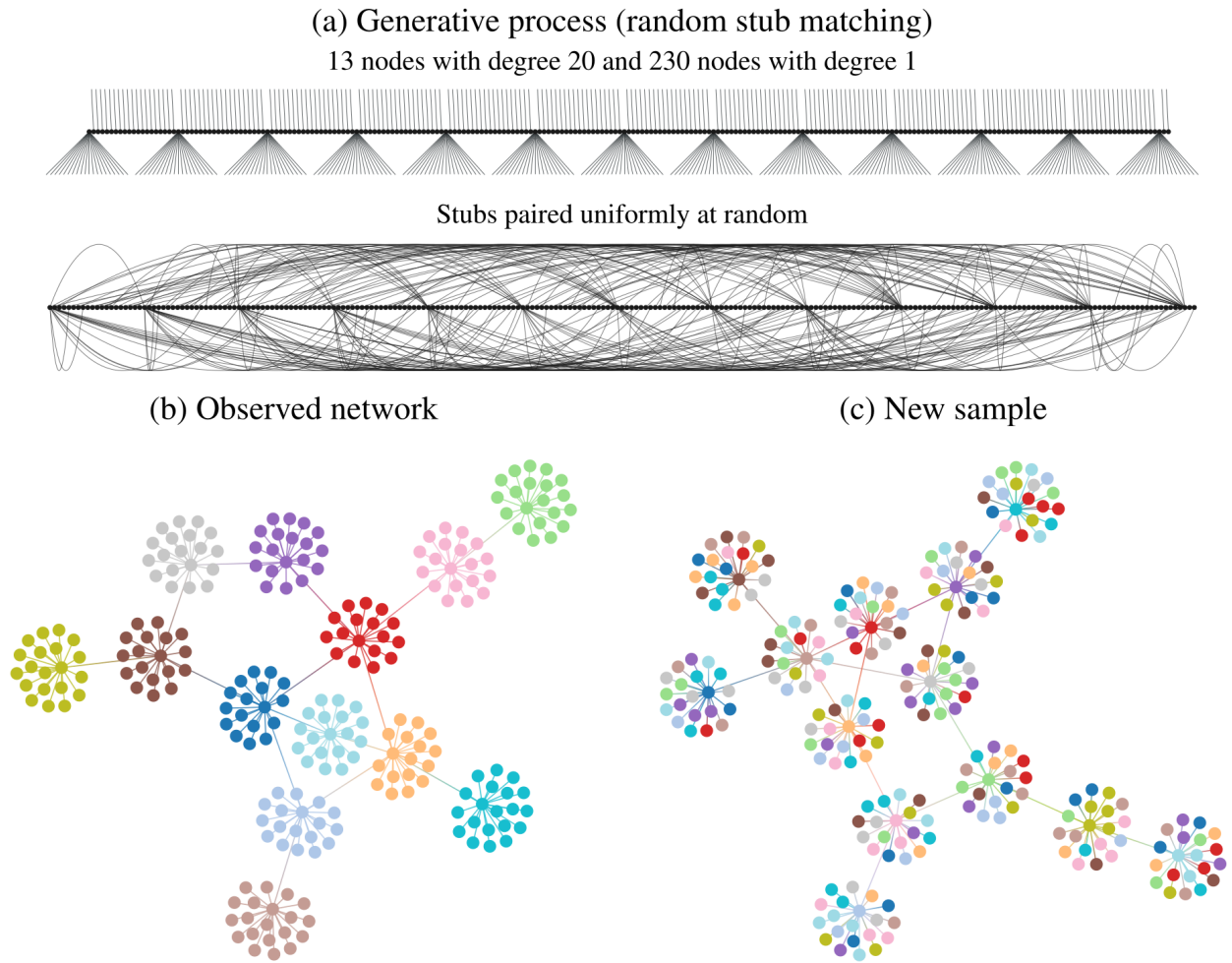
A network with 13 communities



A random network with a prescribed degree sequence, and no community structure.

TP Peixoto, *Elements in the Structure and Dynamics of Complex Networks*, DOI 10.1017/9781009118897 (2023).

Descriptive vs inferential



TP Peixoto, *Elements in the Structure and Dynamics of Complex Networks*, DOI 10.1017/9781009118897 (2023).

Practicalities

Projects

- Project team match making, stick around!
- Choose your topic
  - Choice opens today, 12:00 at Brightspace
  - Choice closes Oct 4, 12:00
- Presentations are run in parallel tracks
- Assignment 1 due this Monday

Projects

- Project team match making, stick around!
- Choose your topic
  - Choice opens today, 12:00 at Brightspace
  - Choice closes Oct 4, 12:00
- Presentations are run in parallel tracks
- Assignment 1 due this Monday

No class next week.
Enjoy “Leidens ontzet”!