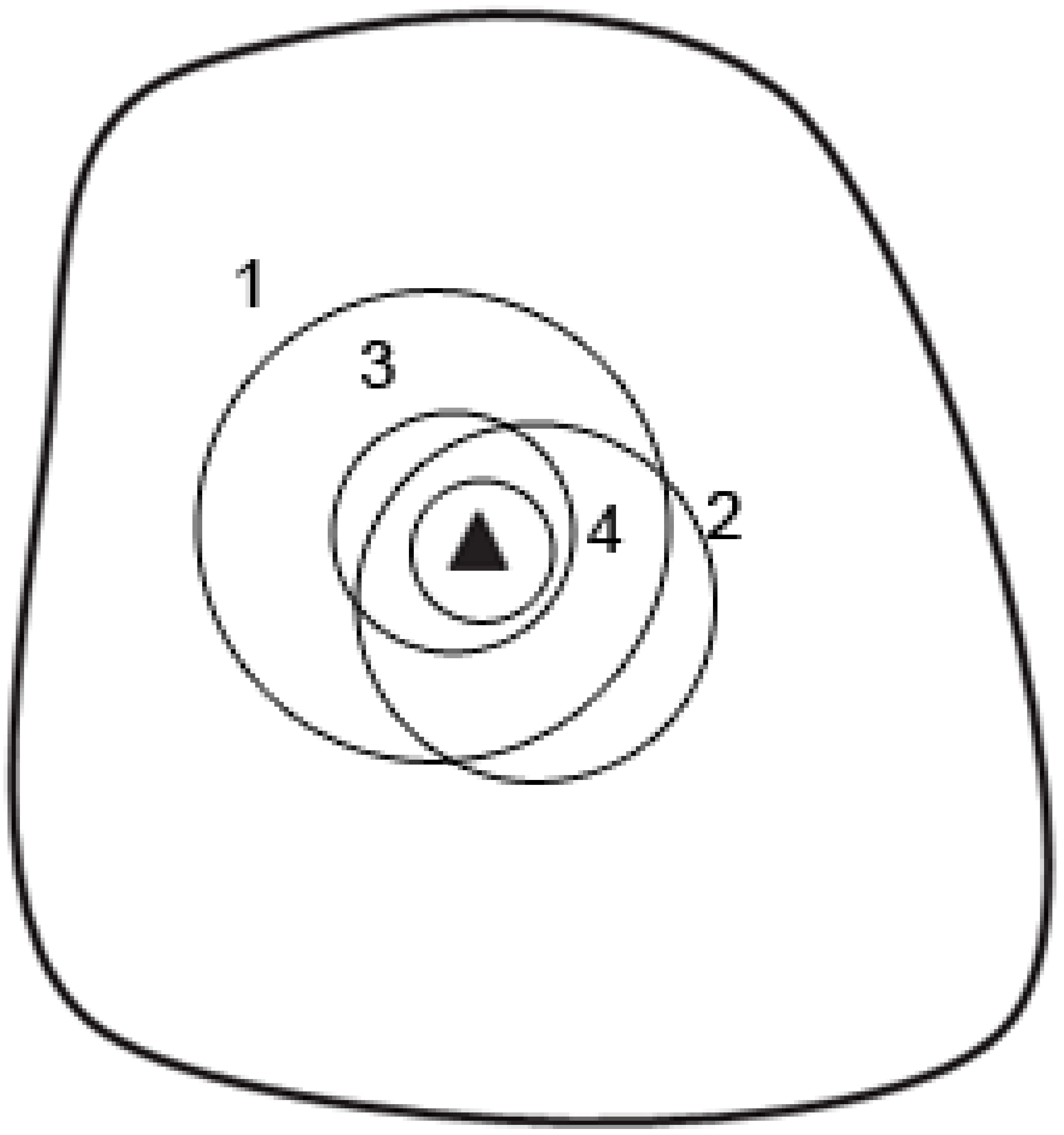
NATIONAL RESEARCH UNIVERSITY

CLUSTERING AND COMMUNITY DETECTION

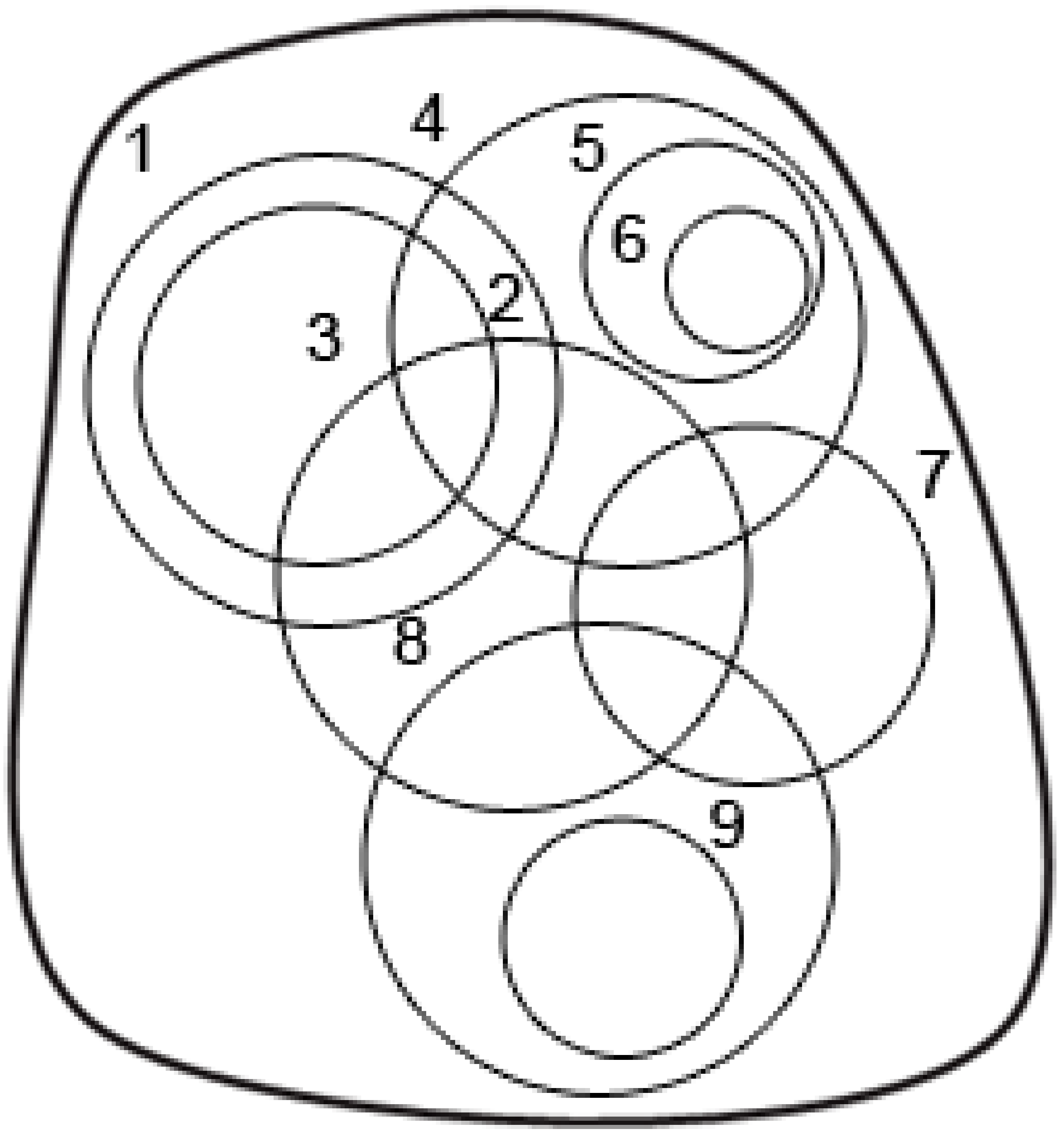
Moscow, 2023

ITERATIVE AND EXPLORATIVE SEARCH

Iterative search



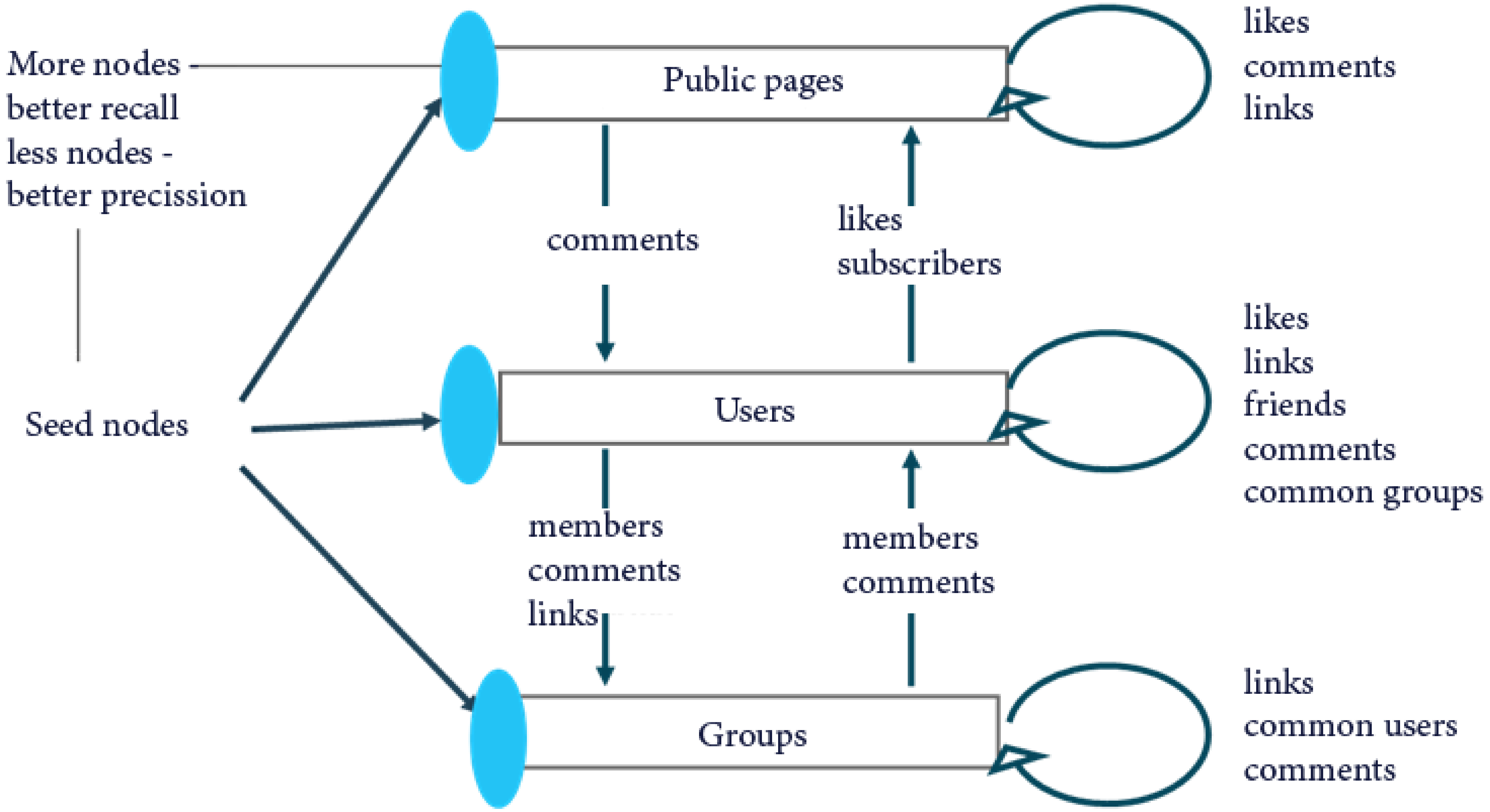
Explorative search



R.W. White and R.A. Roth, “Exploratory Search: Beyond the Query-Response Paradigm”, 2009, pp. 1–98.

ВВЕДЕНИЕ

МЕТОДОЛОГИЯ ЗЕРНОВОГО ПРИРАЩЕНИЯ АКТОРОВ



ТИПОВЫЕ ОТНОШЕНИЯ В СОЦИАЛЬНЫХ МЕДИА

Actor 1 type	Actor 2 type	Tie type	description	direction
Public page	Public page	likes	1 or 0	directed
Public page	Public page	simultaneous likes	amount of users that simultaneously like the page	undirected
Public page	Public page	mutual posting	amount of users that simultaneously wrote post	undirected
Public page	Public page	mutual commenting	amount of users that simultaneously wrote comment to the page post	undirected
Public page	Public page	links	amount of mutual links	directed
Public page	Public page	subscribers intersection	amount of mutual subscribers	undirected
User	Public page	comments	total amount of comments	directed
User	Public page	likes	total amount of post likes	directed
User	Public page	subscribers	1 or 0	directed
Public page	User	likes	total amount of post likes	directed
User	User	friends	1 or 0	directed
User	User	links	amount of mutual links	directed
User	User	post likes	amount of user post likes	directed
User	User	comment likes	amount of user comment likes	directed
User	User	comment	amount of user post comments	directed
User	User	simultaneous comments	amount of users that simultaneously wrote post	undirected
User	User	Publications intersection	amount of that simultaneous comments to external posts	undirected
User	User	membership	amount of co-membership groups	undirected
User	User	joint subscription	amount of co-membership public pages	undirected
User	Group	membership	1 or 0	directed
User	Group	commenting	amount of comments	directed
User	Group	links	amount of links	directed
User	Group	reposts	amount of reposts from the group	directed
Group	Group	direct links	amount of links	directed
Group	Group	members intersection	members intersection	undirected
Group	Group	mutual commenting	amount of active users	undirected
Group	Public page	likes	amount of likes	directed
Public page	Group	likes	amount of likes	directed
Group	Public page	links	amount of links	directed
Public page	Group	links	amount of links	directed

$$p_{ab} = \frac{\sum_{i=1}^{n} x_i^{(ab)}}{\sum_{i=1}^{n} x_i^{(a)}}$$

- связи между акторами в матрице (a)
- пересечение связей между акторами в матрицах (a) и (b)

Group - Group

	GROUP_GROUP _BY_LINKS	GROUP_GROUP _BY_USERS	GROUP_GROUP _COMMENT
GROUP_GROUP_BY_LINKS	1	0,98	0,73
GROUP_GROUP_BY_USERS	0,02	1	0,18
GROUP_GROUP_COMMENT	0,08	0,96	1

Public page - Public page

	PAGE_PAGE _BY_LIKES	PAGE_PAGE _BY_LINKS	PAGE_PAGE _BY_ USERCOMMENTS	PAGE_ PAGE _BY_ USERS
PAGE_PAGE_BY_LIKES	1	0,02	0,35	0,75
PAGE_PAGE_BY_LINKS	0,39	1	0,68	0,98
PAGE_PAGE _BY_USERCOMMENTS	0,47	0,04	1	0,94
PAGE_PAGE_BY_USERS	0,36	0,02	0,34	1

АНАЛИЗ СИЛЫ РАЗЛИЧНЫХ СВЯЗЕЙ В СОЦИАЛЬНЫХ МЕДИА

User - Group

	USER_GROUP	USER_GROUP _BY_LINKS	USER_GROUP _COMMENT	USER_GROUP _REPOST
USER_GROUP	1	0	0,07	0
USER_GROUP _BY_LINKS	0,31	1	0,19	0
USER_GROUP _COMMENT	0,76	0	1	0
USER_GROUP _REPOST	0	0	0	1

User - Public page

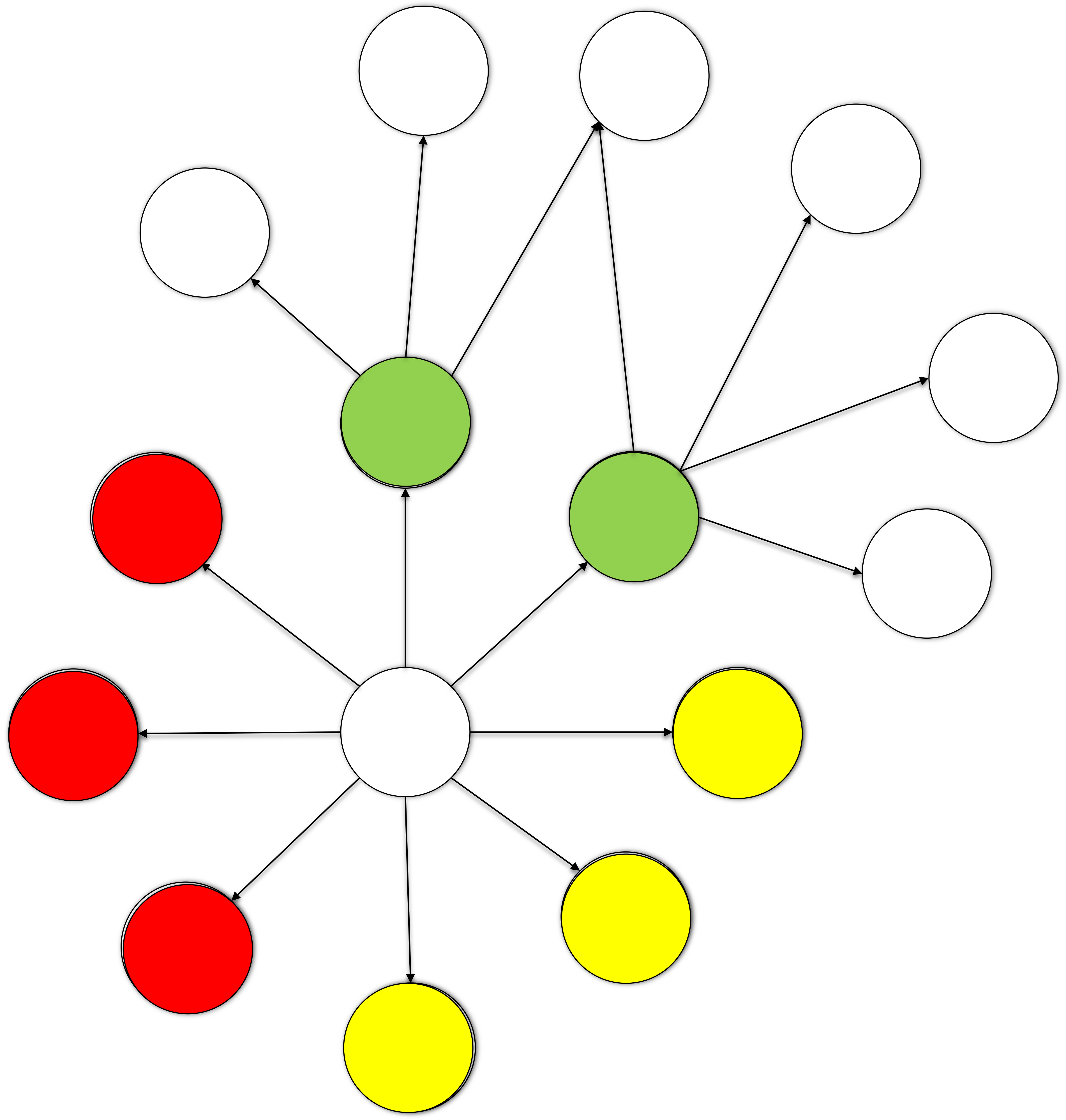
	USER_PAGE	USER_PAGE_BY_ COMMENT	USER_PAGE_BY_ LIKES
USER_PAGE	1	0,05	0,27
USER_PAGE_BY_COMMENT	0,76	1	0,82
USER_PAGE_BY_LIKES	0,2	0,04	1

АНАЛИЗ СИЛЫ РАЗЛИЧНЫХ СВЯЗЕЙ В СОЦИАЛЬНЫХ МЕДИА USER - USER

`	USER_USER_ BY_ANOTHER _COMMENT	USER_USER _BY_ COMMENT	USER_USER _BY_LIKES	USER_ USER _BY _LINKS	USER_USER _BY_ POSTSIN GROUP
USER_USER_ BY_ANOTHER _COMMENT	1	0	0	0	0
USER_USER_ BY_COMMENT	0,27	1	0,43	0	0,478
USER_USER_ BY_LIKES	0,06	0,06	1	0	0,146
USER_USER_ BY_LINKS	0,02	0	0	1	0
USER_USER_ BY_ POSTSIN GROUP	0,03	0	0	0	1

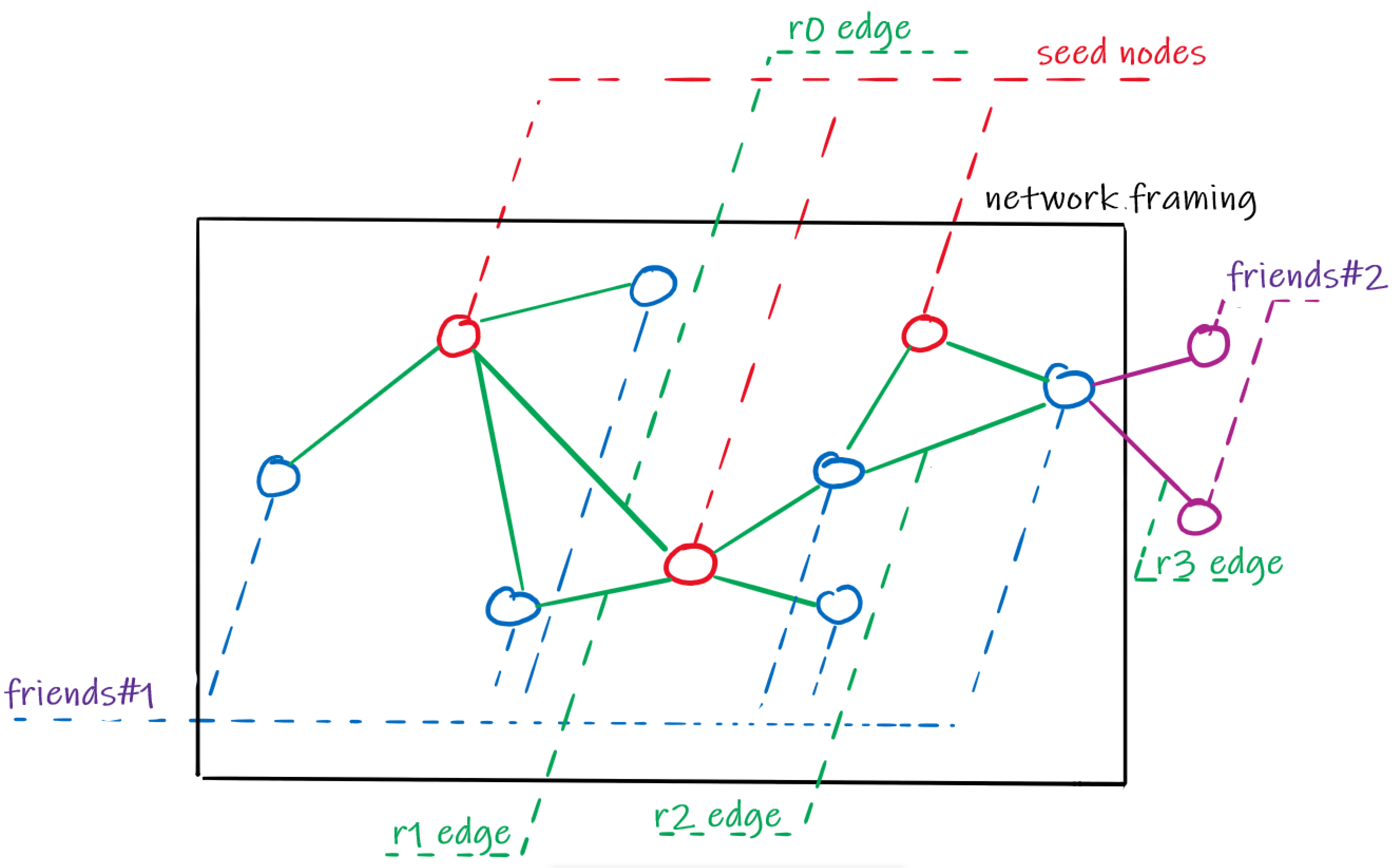
МЕТОД ЗЕРНОВОГО ПРИРАЩЕНИЯ

СТРАТЕГИЯ ЗЕРНОВОГО ПРИРАЩЕНИЯ НА ОСНОВЕ МОДЕЛИ РАНЖИРОВАНИЯ

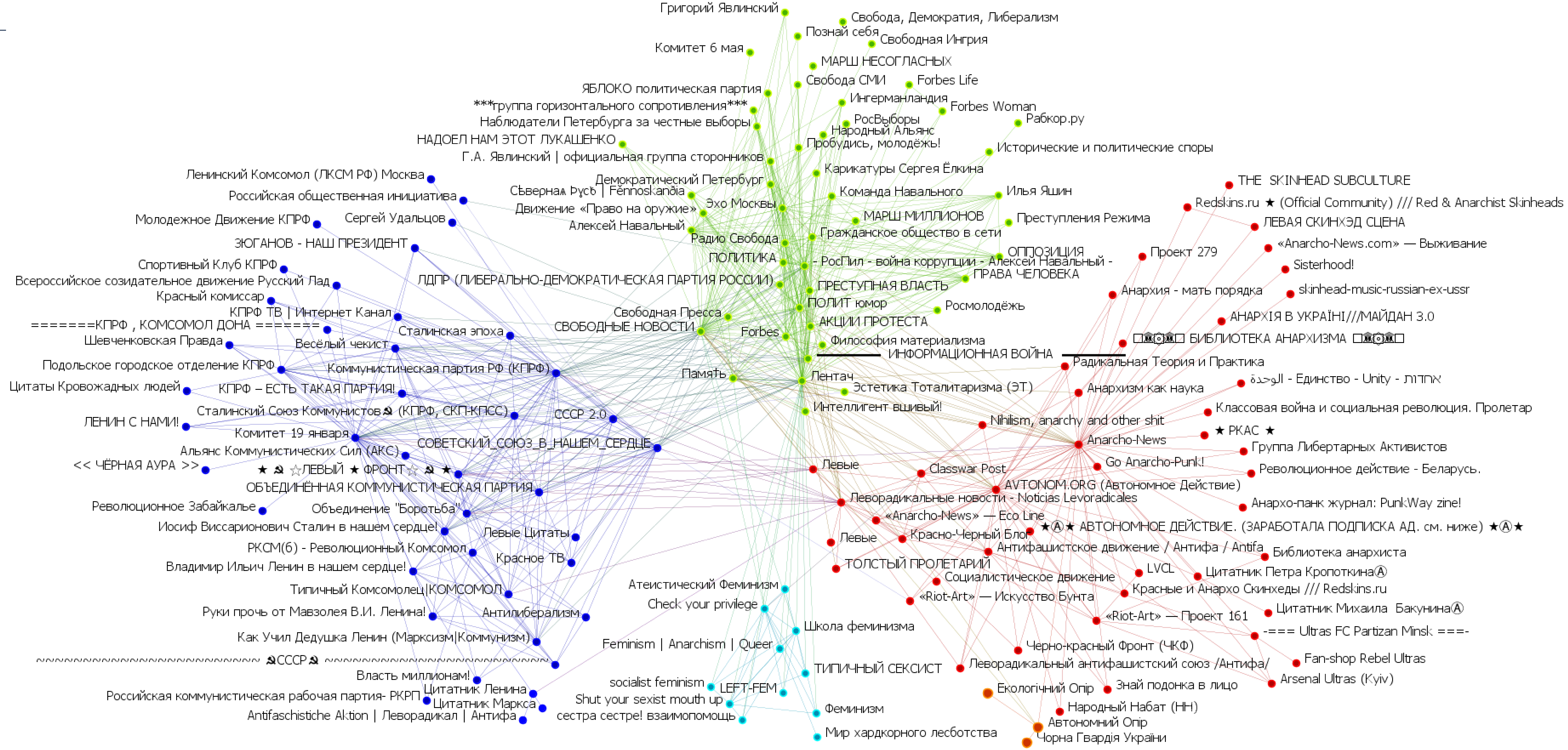


NETWORK PROPERTIES

NETWORK FRAMING

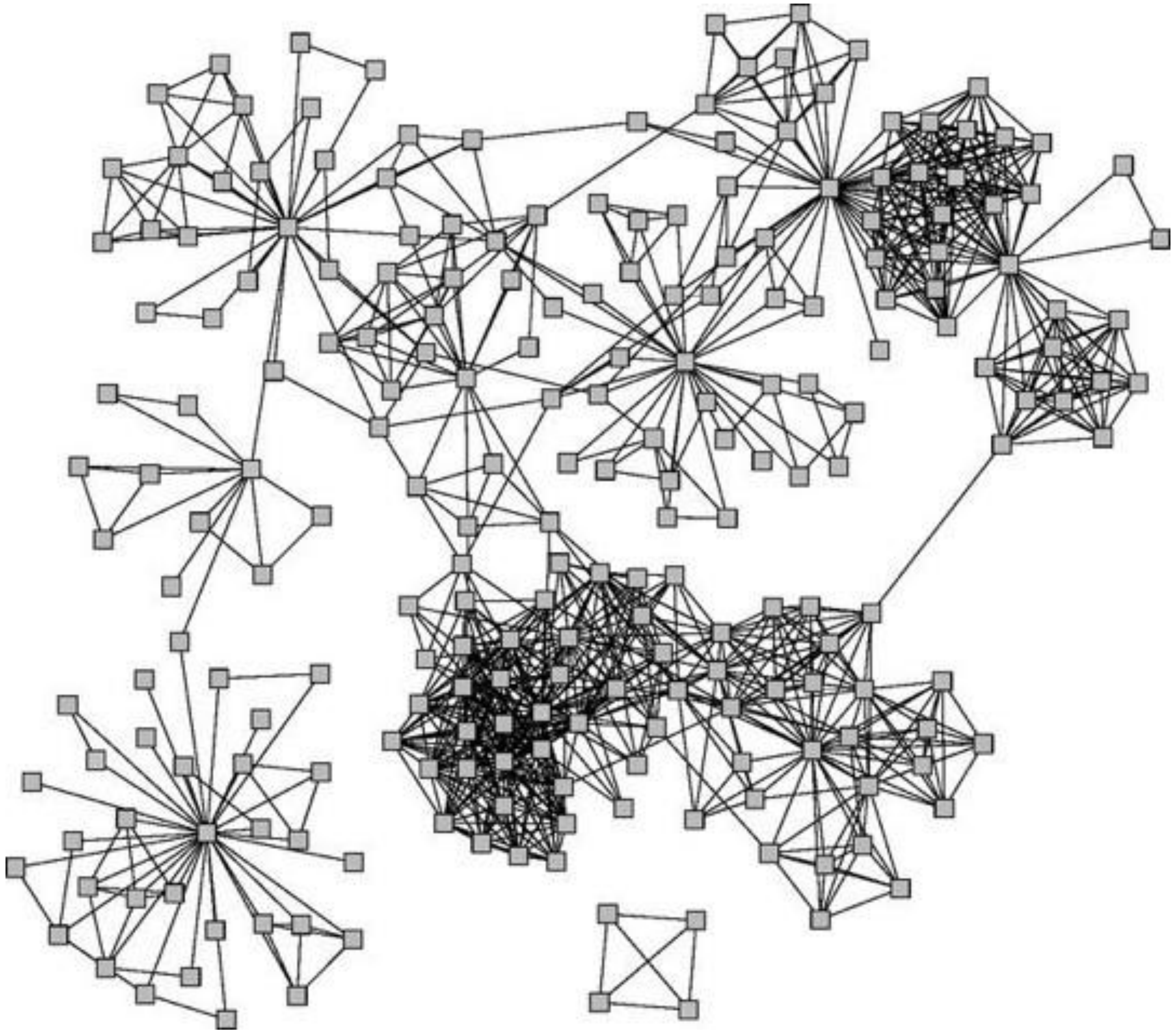


ПРИМЕРЫ ПРИМЕНЕНИЯ МЕТОДА – ПОСТРОЕНИЕ ПОЛИТИЧЕСКОЙ КАРТЫ



NATIONAL RESEARCH UNIVERSITY

COMMUNITY DETECTION



Connected and undirected graphs

NETWORK COMMUNITIES

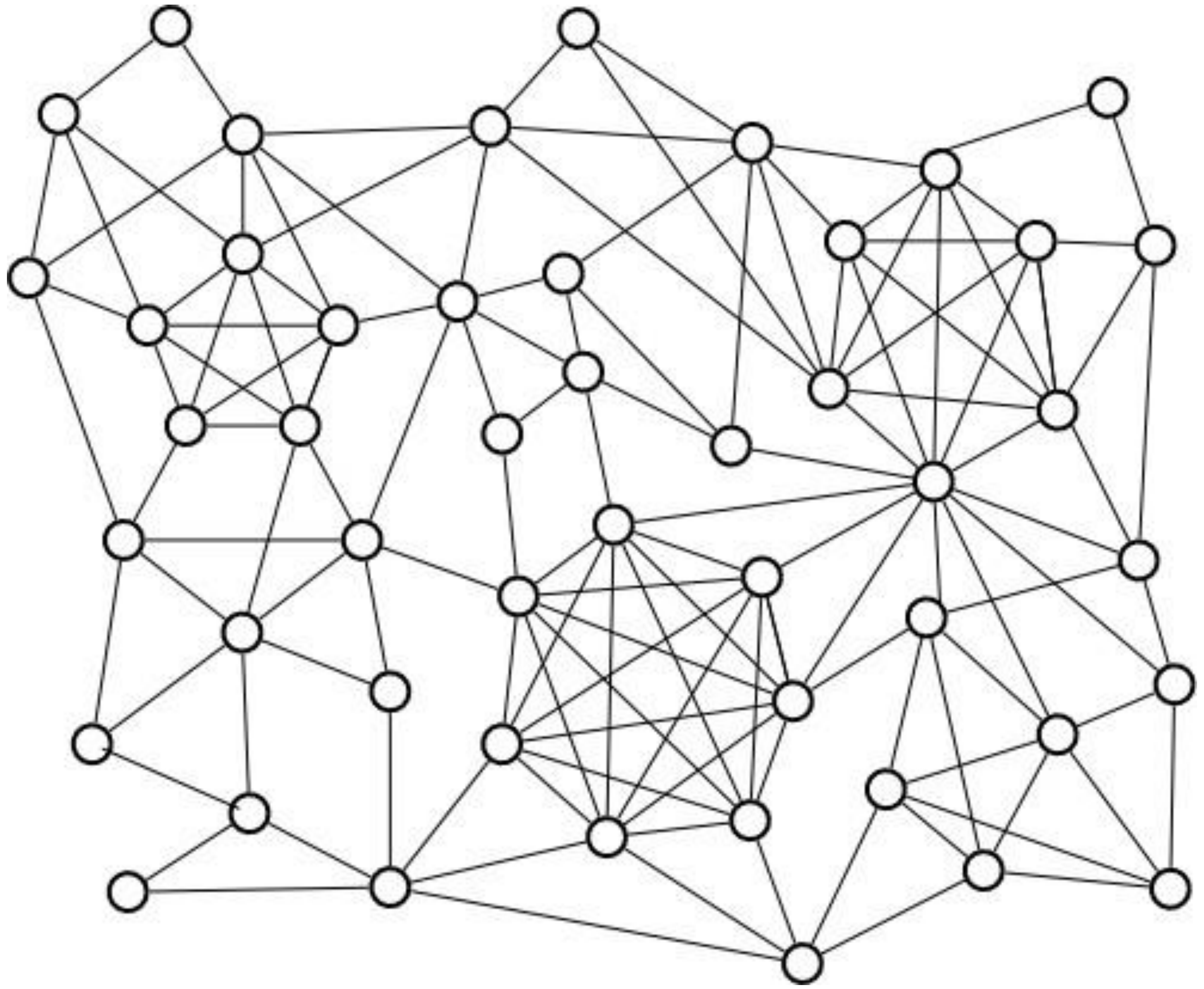
What makes a community (cohesive subgroup):

- Mutuality of ties. Everyone in the group has ties (edges) to one another
- Compactness. Closeness or reachability of group members in small number of steps, not necessarily adjacency
- Density of edges. High frequency of ties within the group
- Separation. Higher frequency of ties among group members compared to non-members

Wasserman and Faust

GRAPH CLIQUES

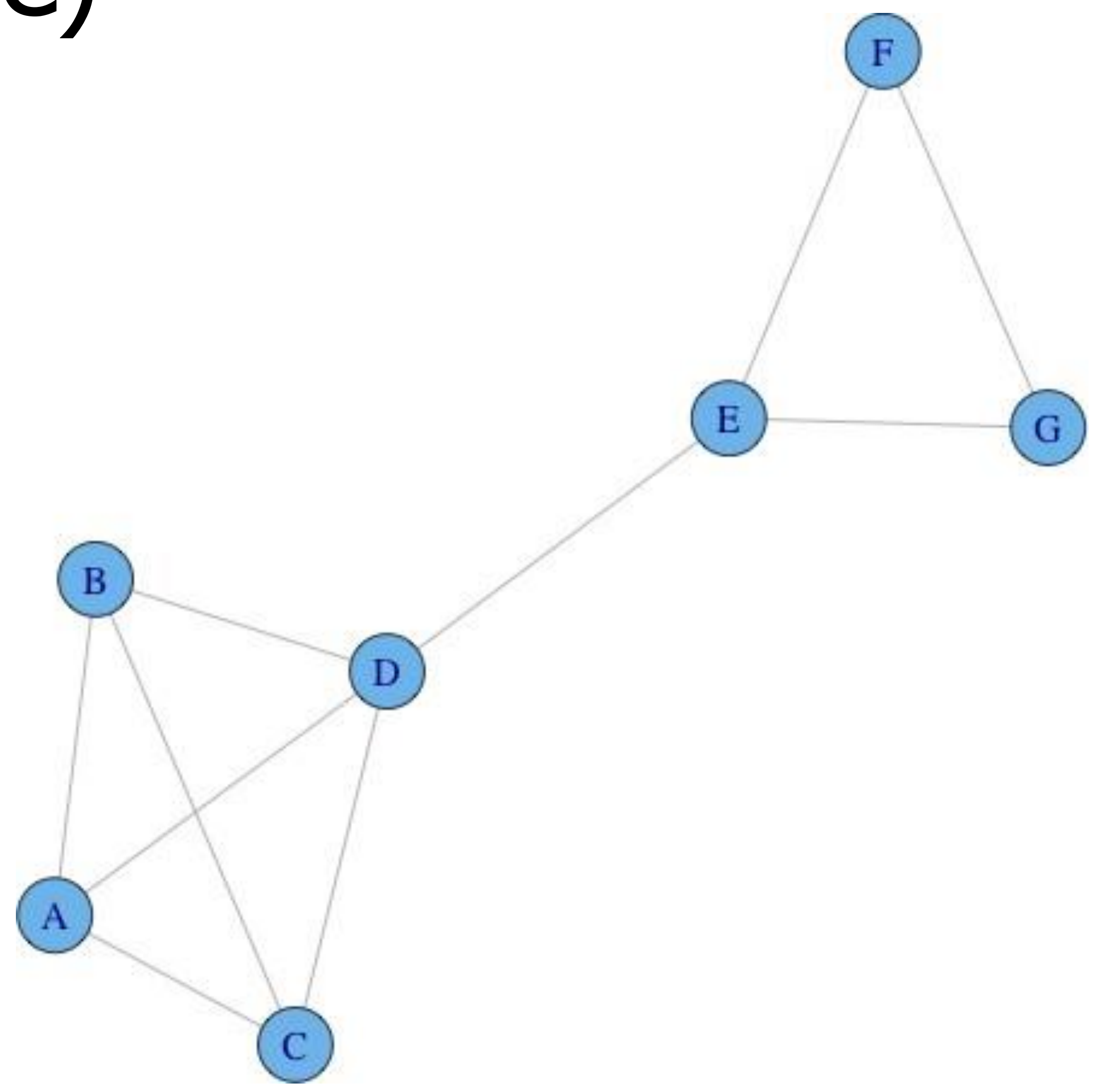
A *clique* is a complete (fully connected) subgraph, i.e. a set of vertices where each pair of vertices is connected.



Cliques can overlap

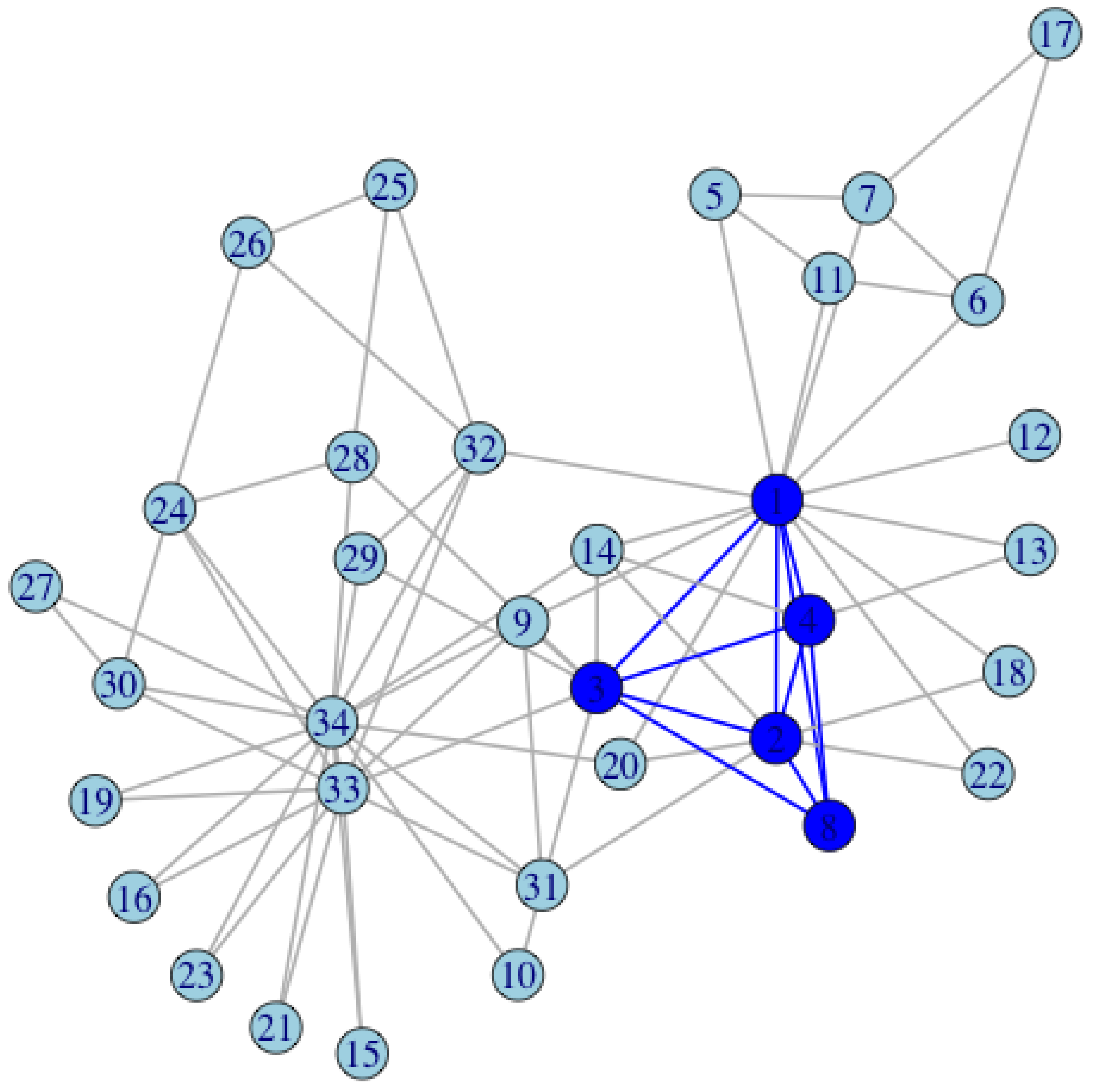
GRAPH CLIQUES

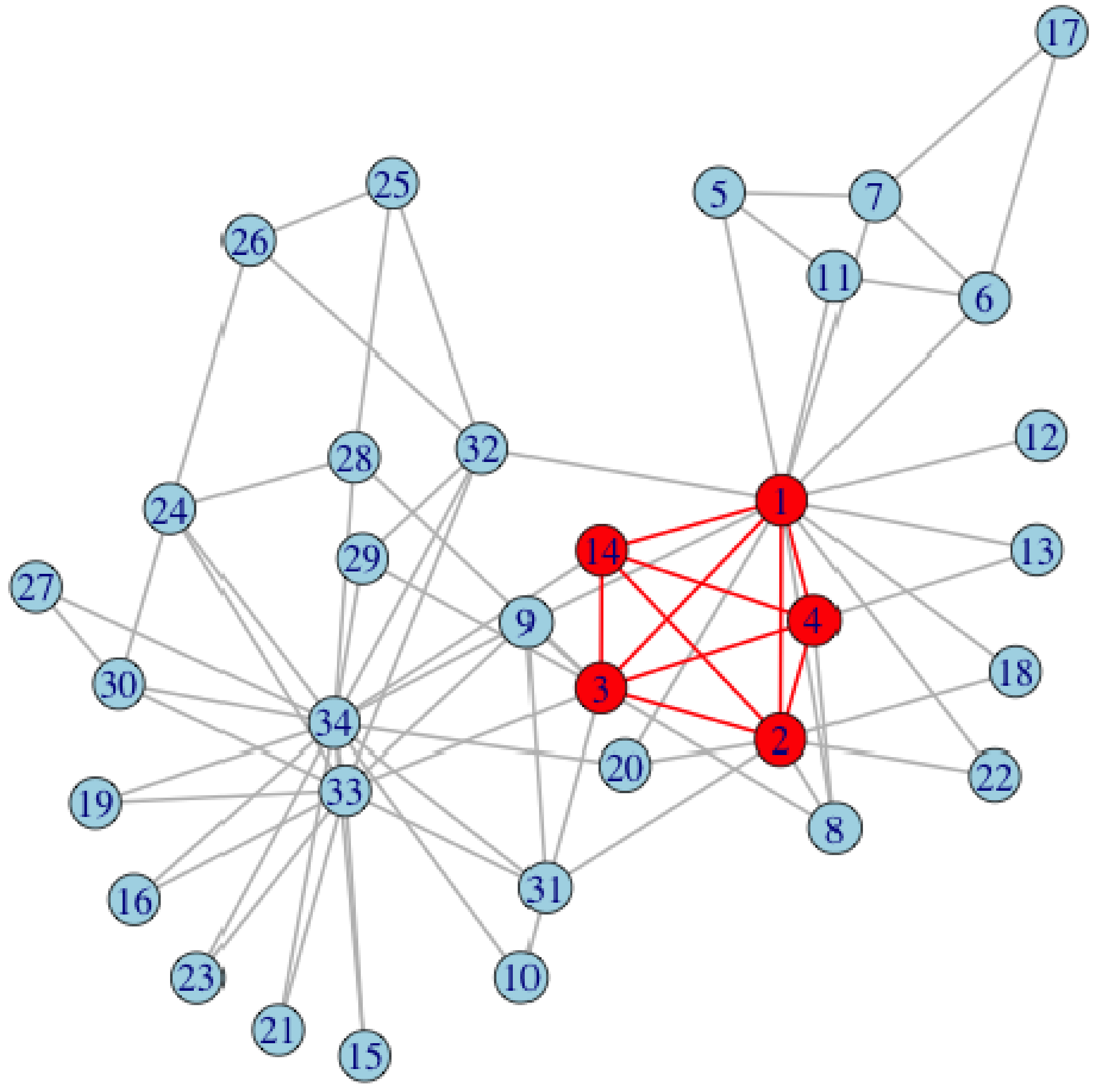
- A **maximal clique** is a clique that cannot be extended by including one more adjacent vertex (not included in larger one)
- A **maximum clique** is a clique of the largest possible size in a given graph
- Graph clique number is the size of the maximum clique



GRAPH CLIQUES

MAXIMUM CLIQUES





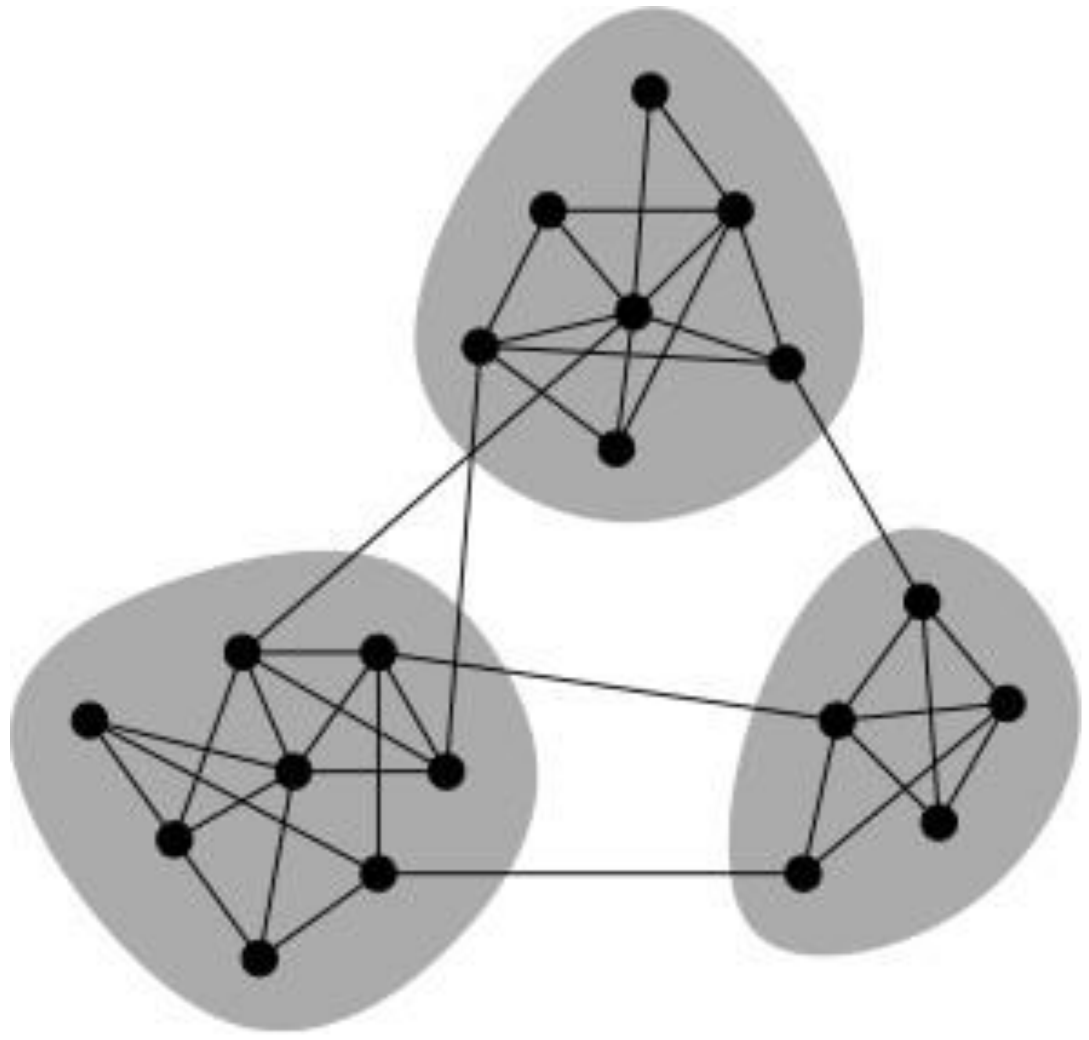
Maximal cliques:

Clique size:	2	3	4	5
Number of cliques:	11	21	2	2

Zachary, 1977

NETWORK COMMUNITIES

Network communities are groups of vertices such that vertices inside the group connected with many more edges than between groups.



Community detection is an assignment of vertices to communities.

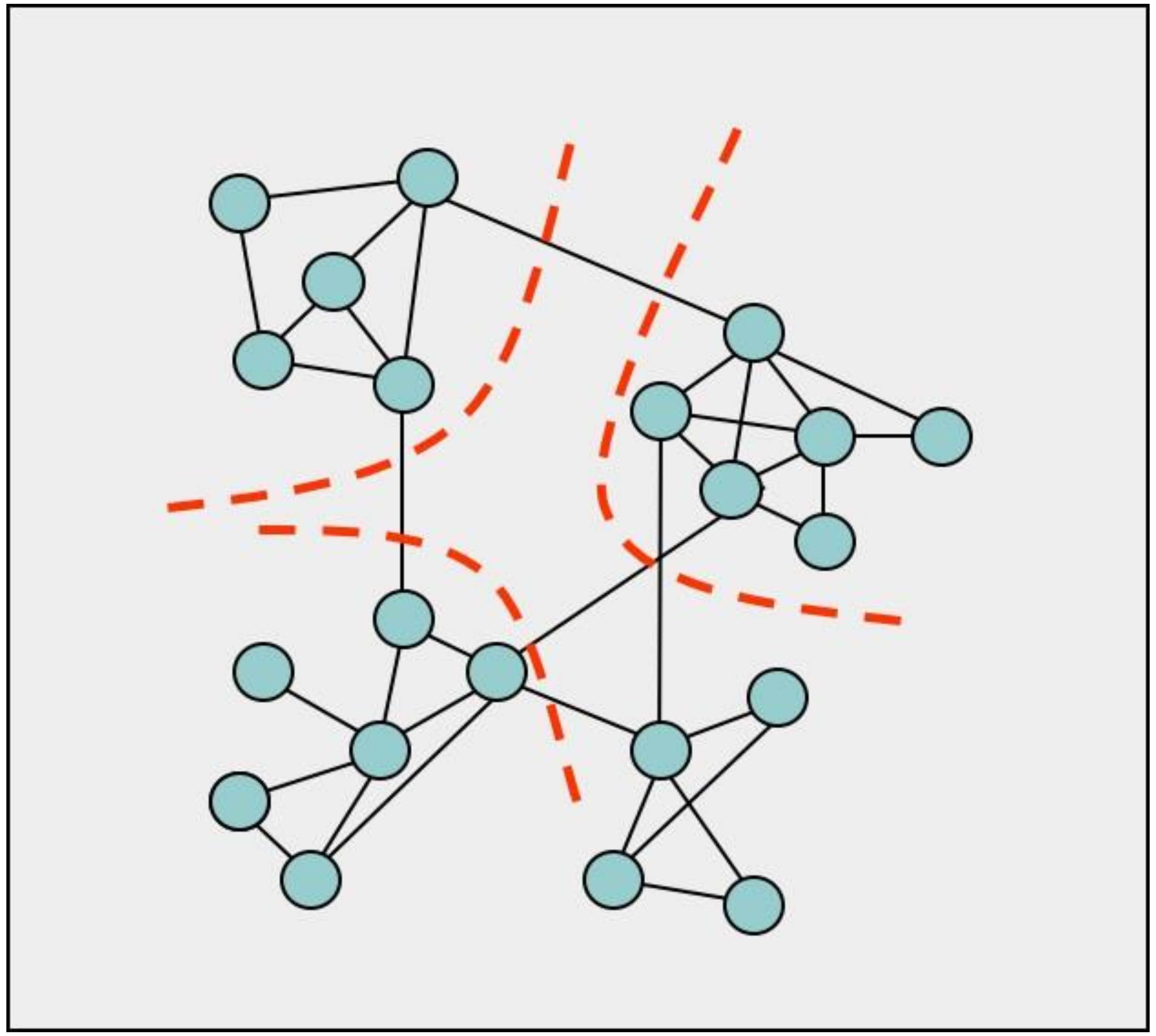
Will consider non-overlapping communities, graph cuts

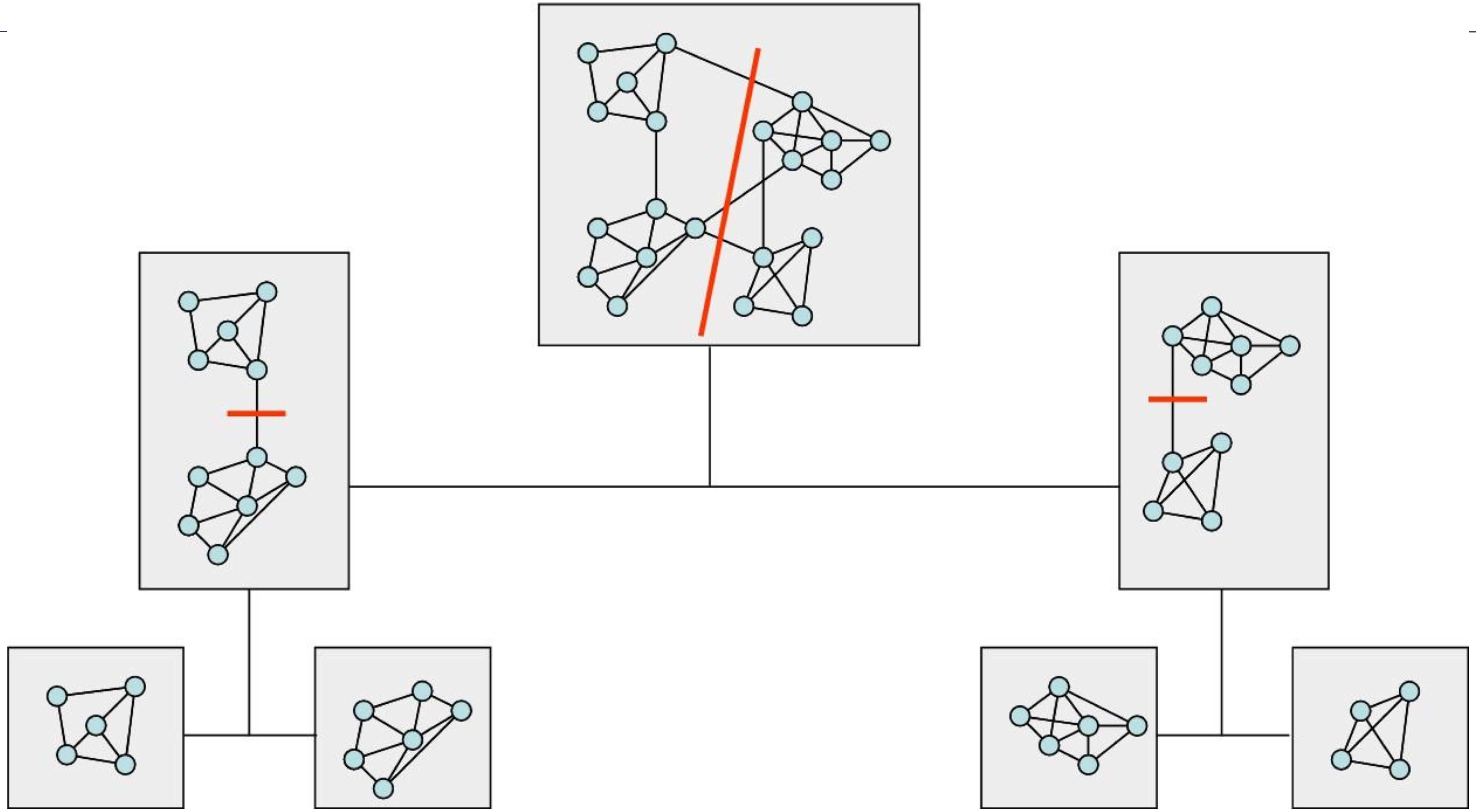
COMMUNITY DETECTION

Consider only sparse graphs $m \ll n^2$ Each community should be connected Combinatorial optimization problem:

- optimization criterion (cut, conductance, modularity)
- optimization method

- Exact solution NP-hard
- (bi-partition: $n = n_1 + n_2$, $n!/(n_1!n_2!)$ combinations)
- Solved by greedy, approximate algorithms or heuristics Recursive top-down 2-way partition, multiway partition Balanced class partition vs communities



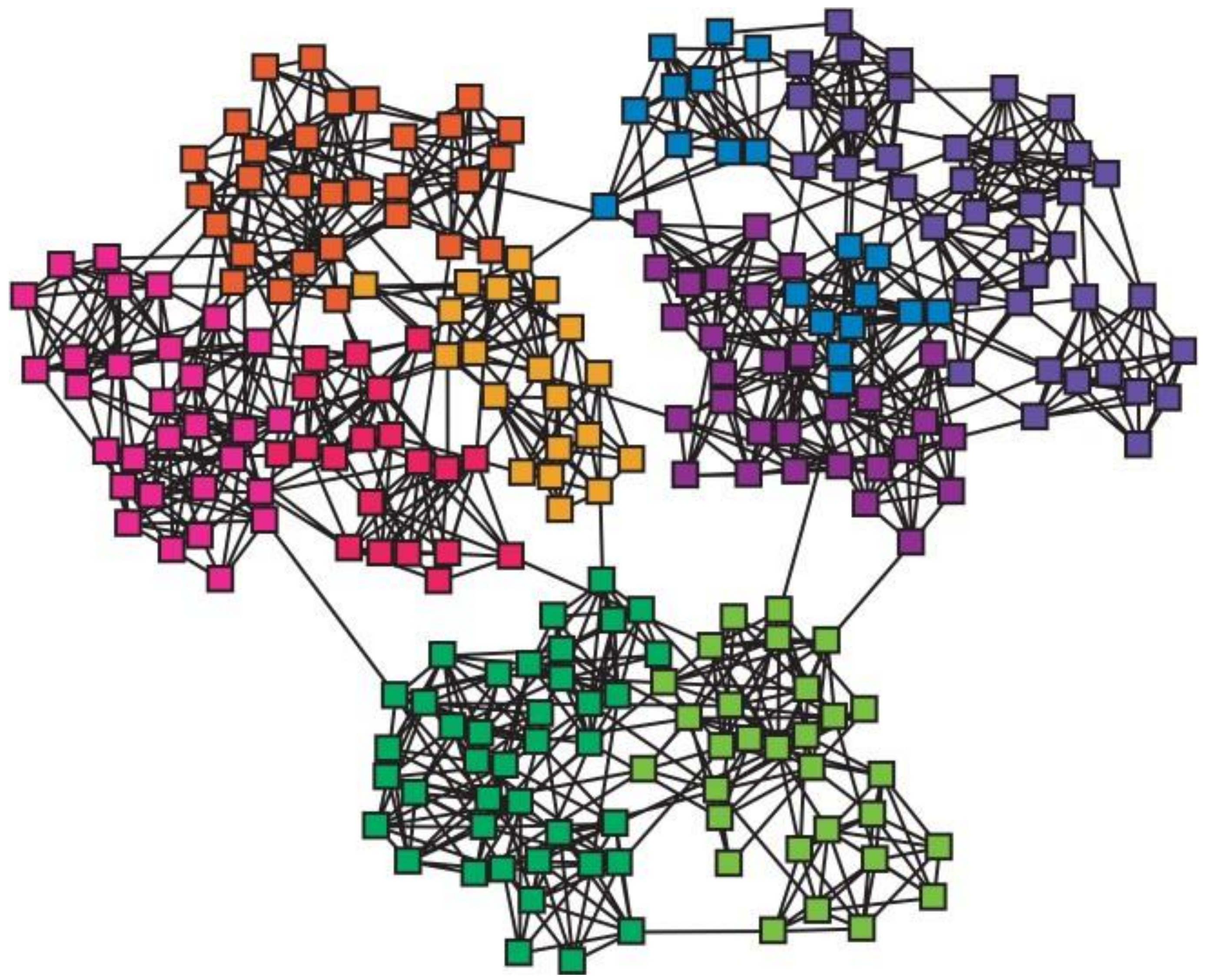


recursive partitioning

EDGE BETWEENNESS

Focus on edges that connect communities.
Edge betweenness -number of shortest paths $\sigma_{st}(e)$ going through edge e

$$C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$$



Construct communities by progressively removing edges

EDGE BETWEENNESS ALGORITHM

Newman-Girvan, 2004

Algorithm: Edge Betweenness

Input: graph G(V,E)

Output: Dendrogram/communities

Repeat

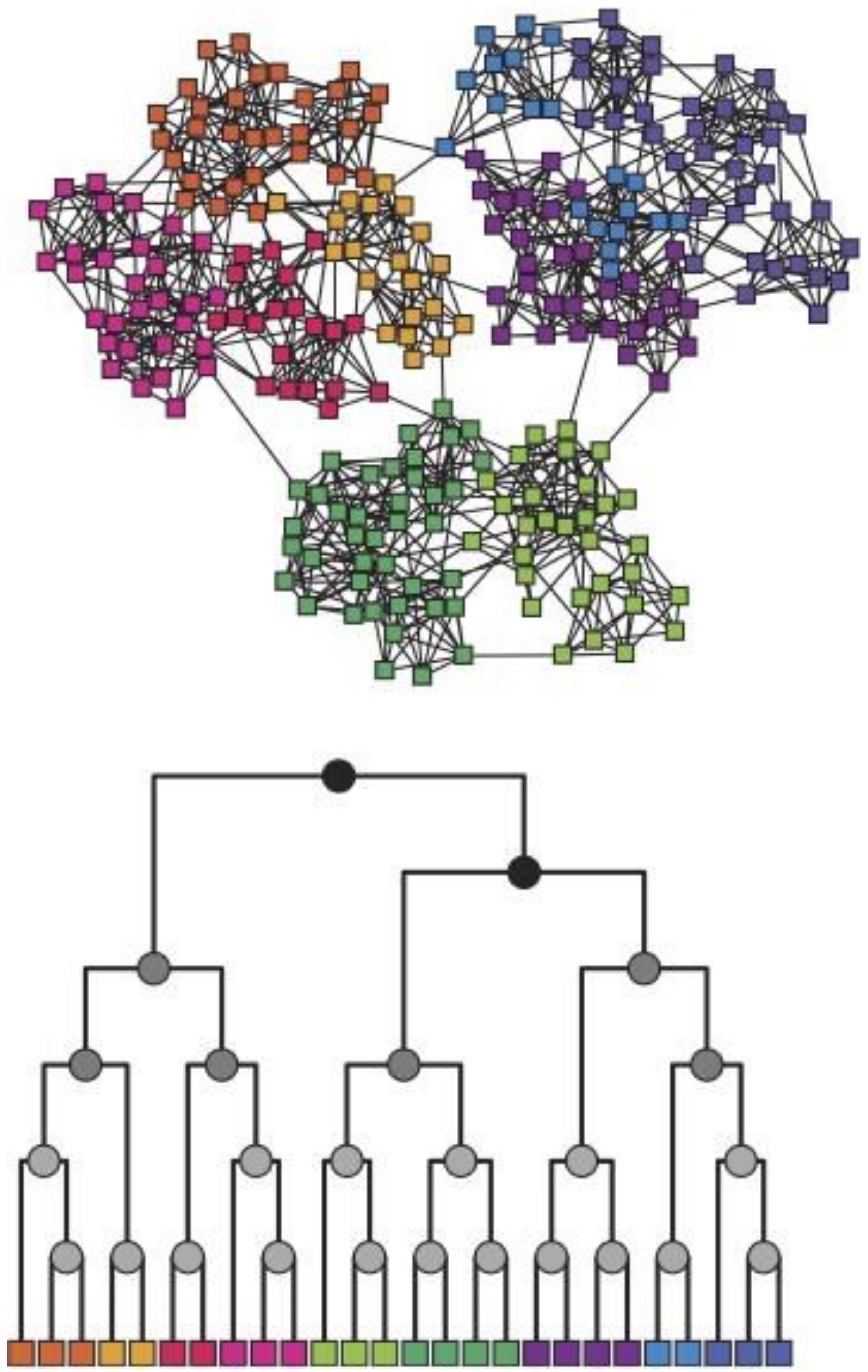
For all $e \in E$ compute edge betweenness $C_B(e)$;

remove edge e_i with largest $C_B(e_i)$;

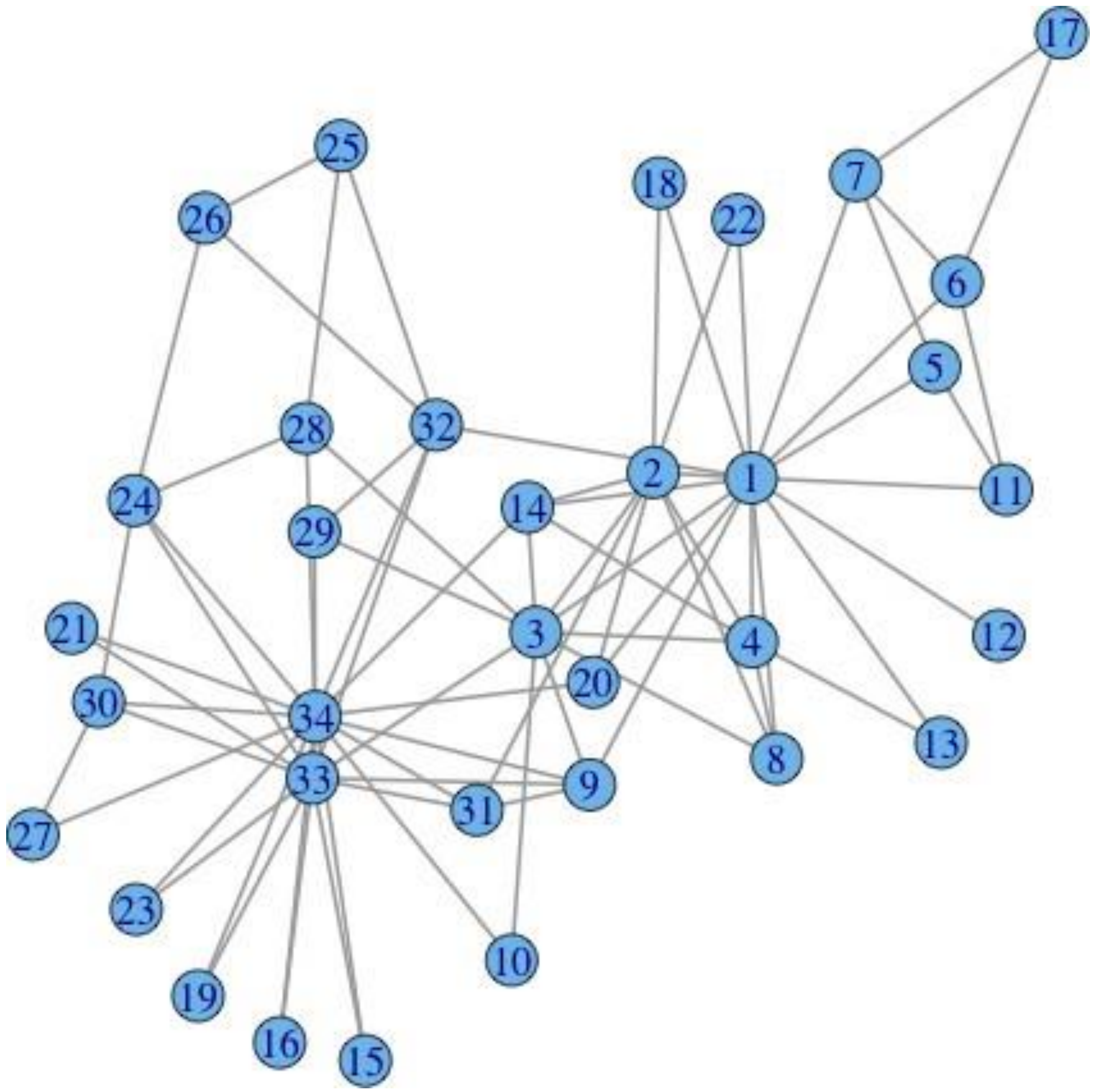
until *edges left*;

If bi-partition, then stop when graph splits in two components (check for connectedness)

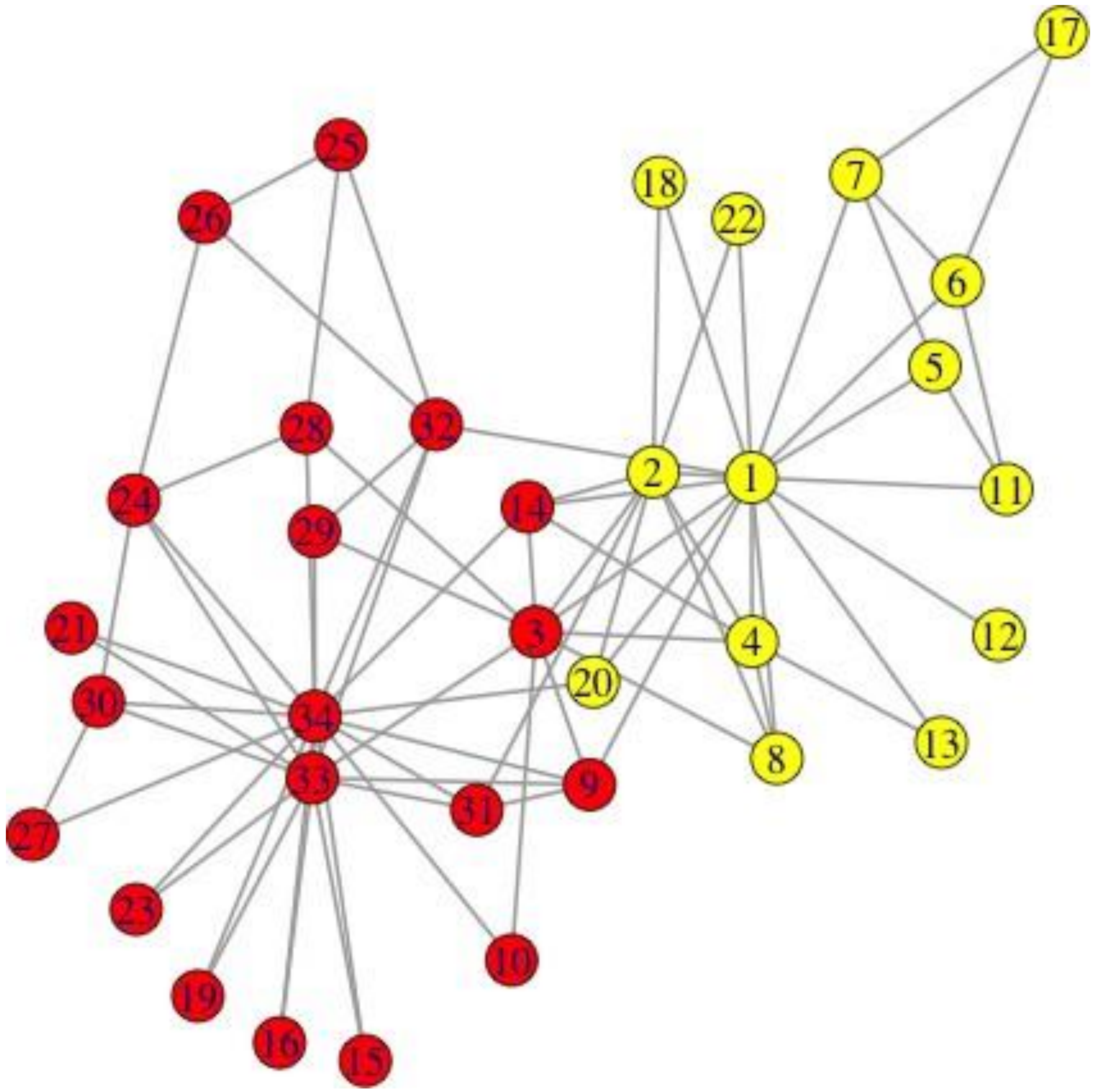
HIERARCHICAL ALGORITHM, DENDROGRAM



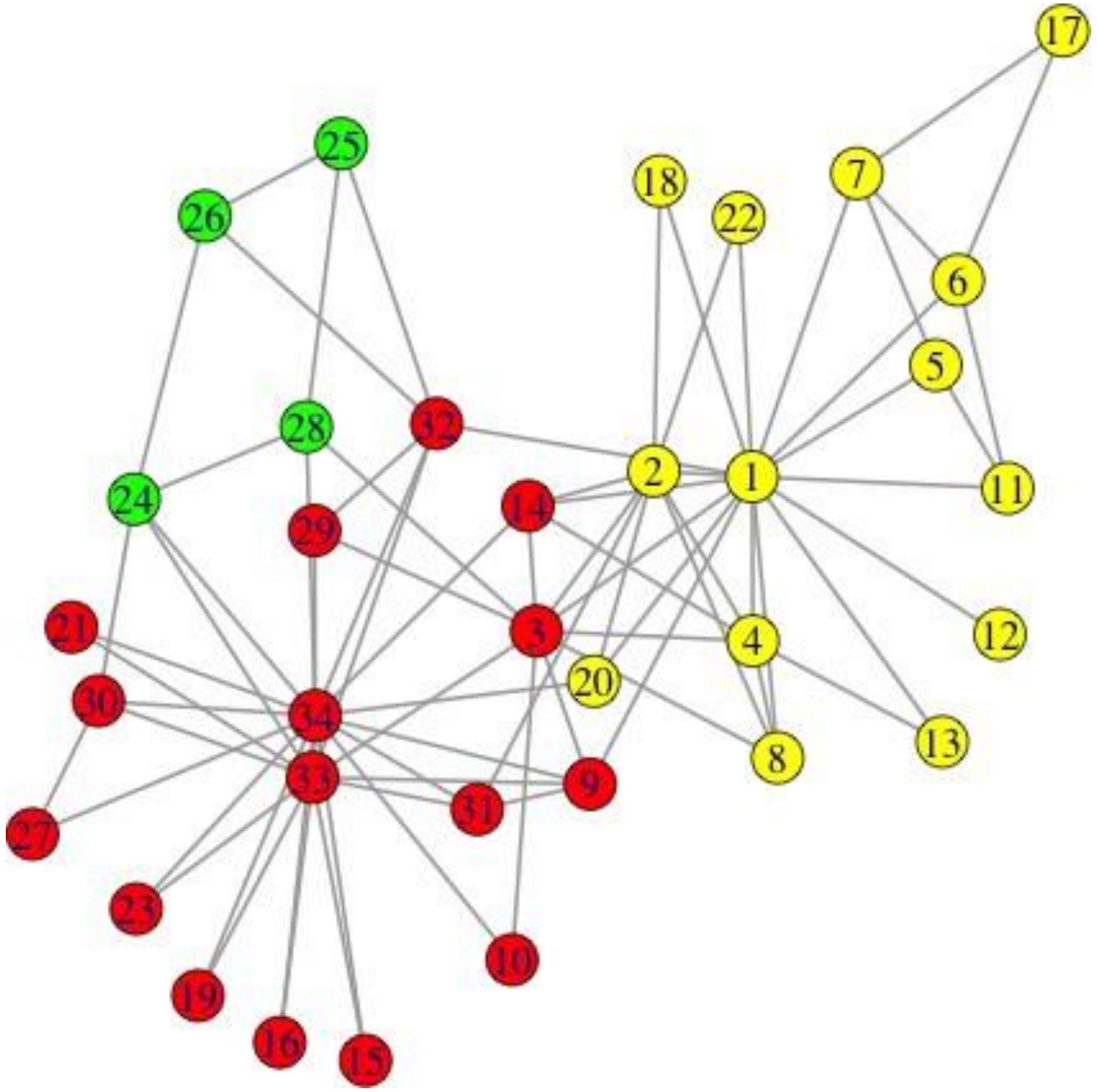
ZACHARY KARATE CLUB



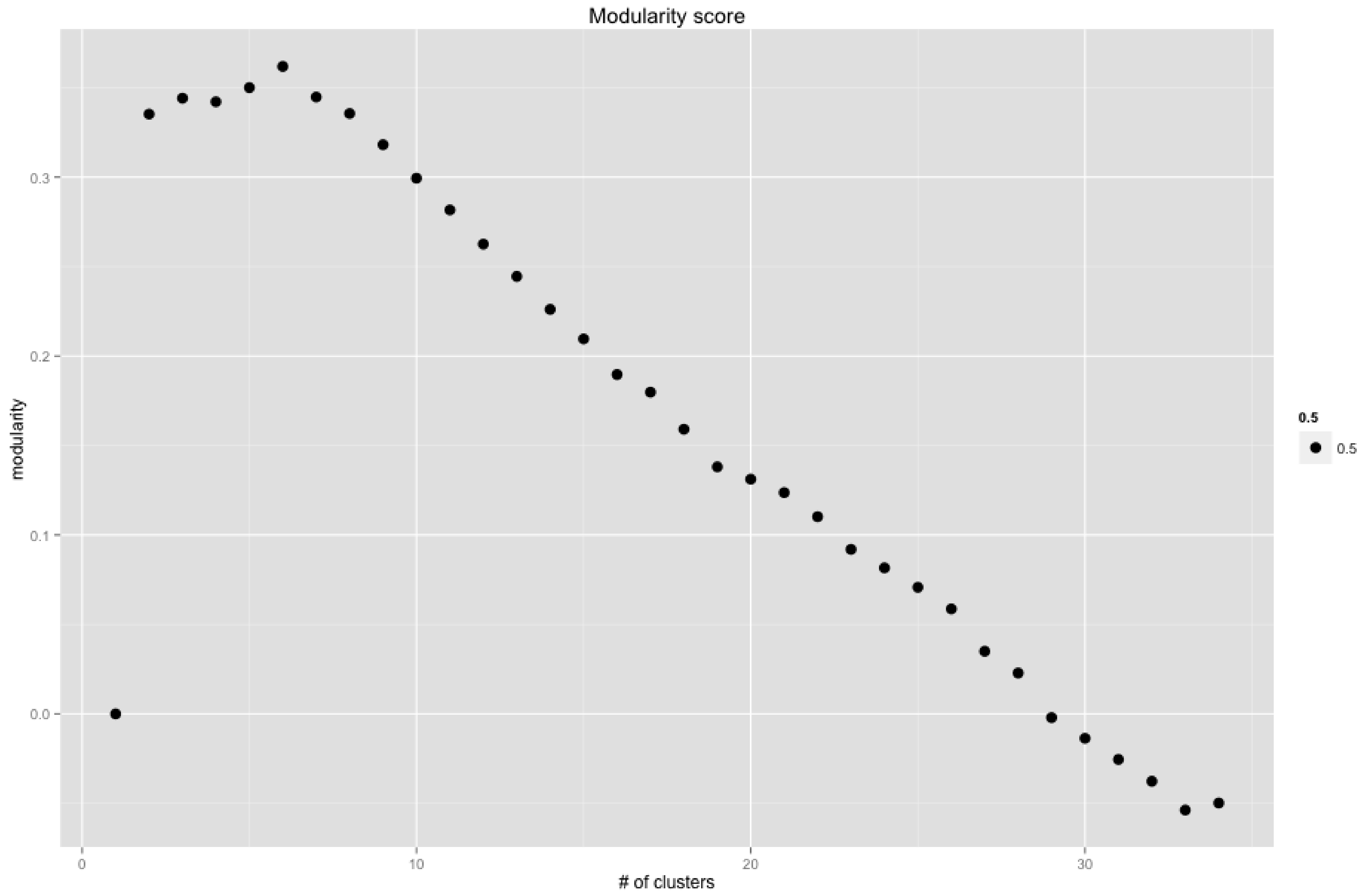
ZACHARY KARATE CLUB



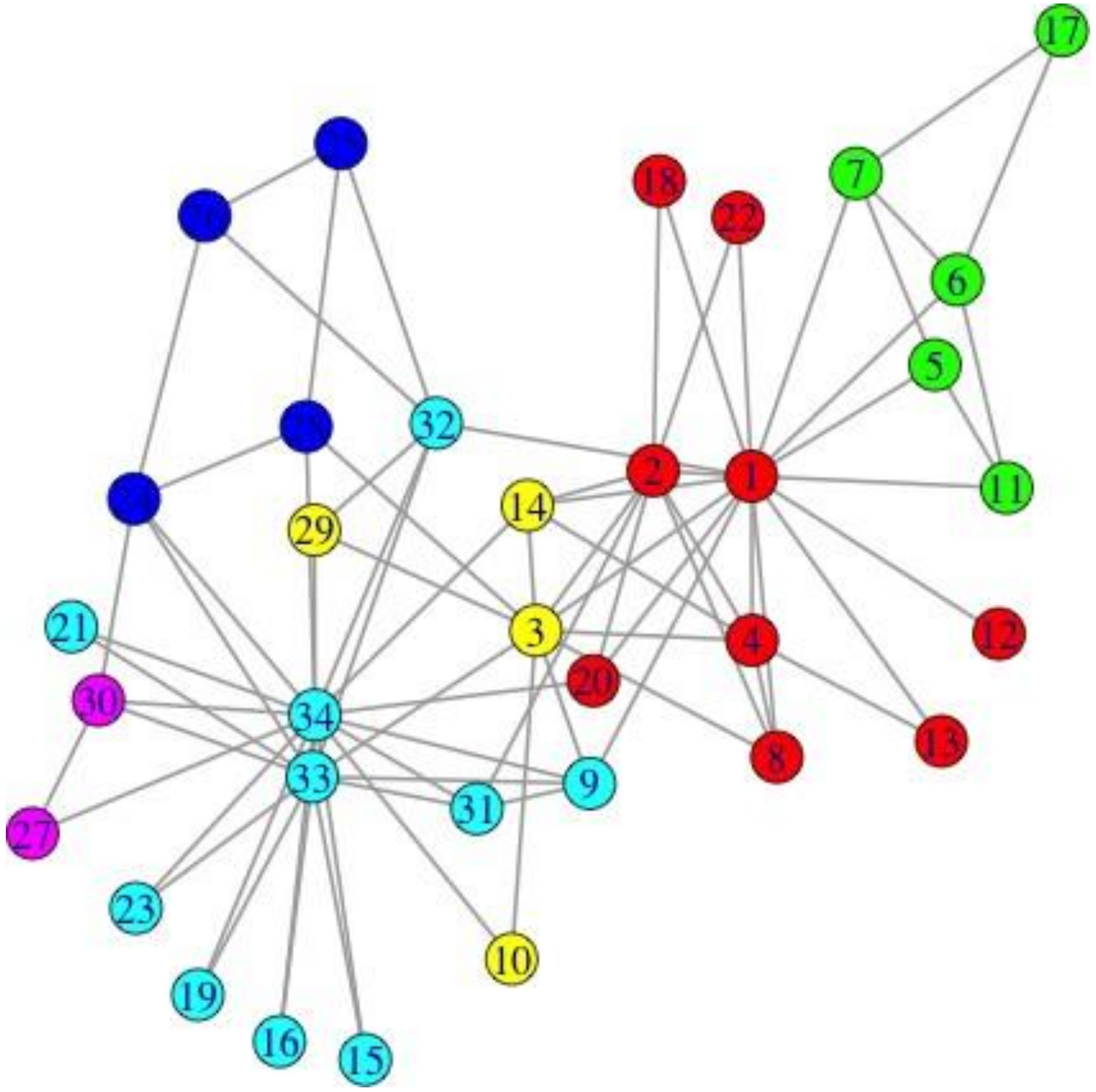
ZACHARY KARATE CLUB

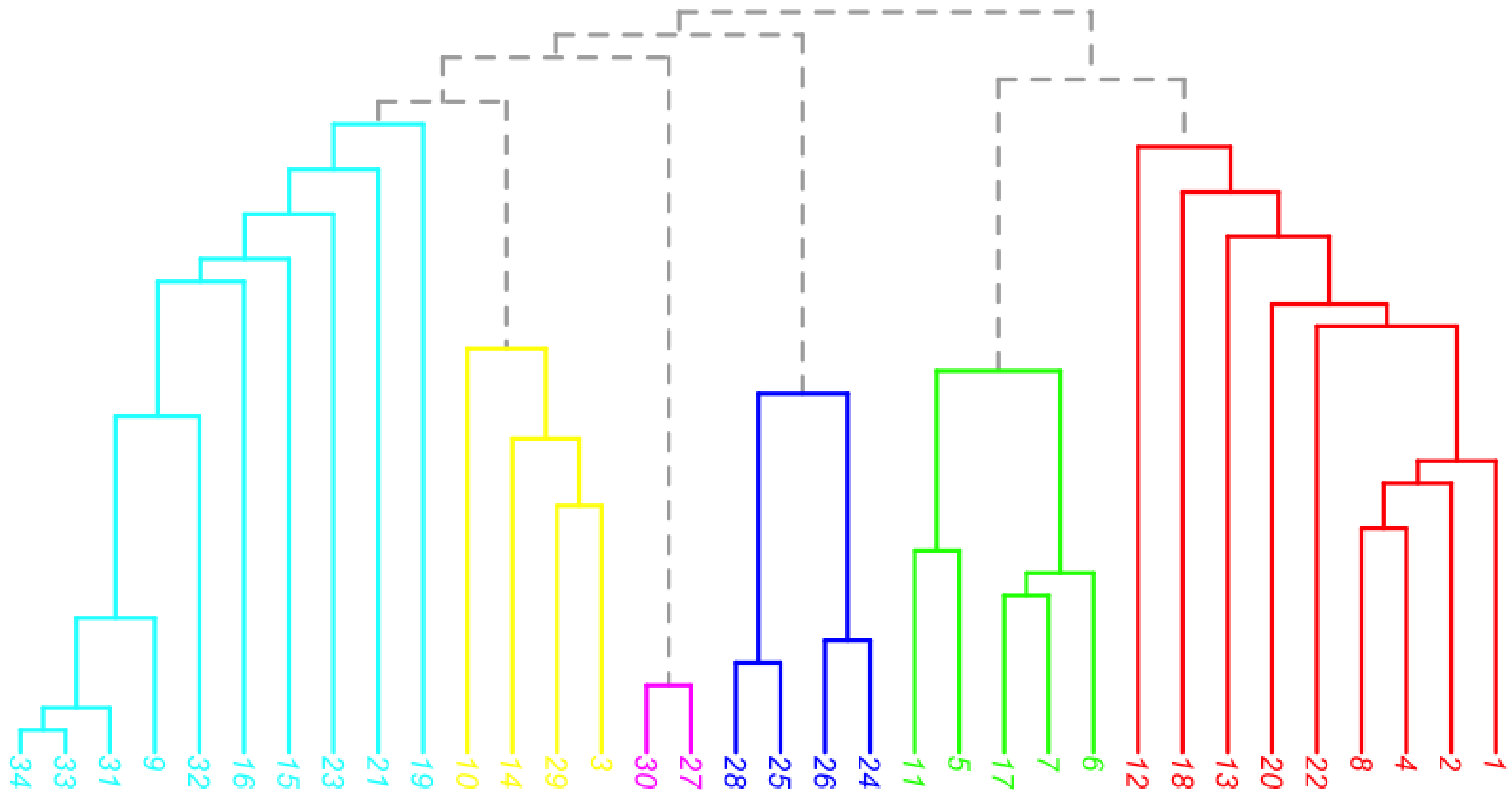


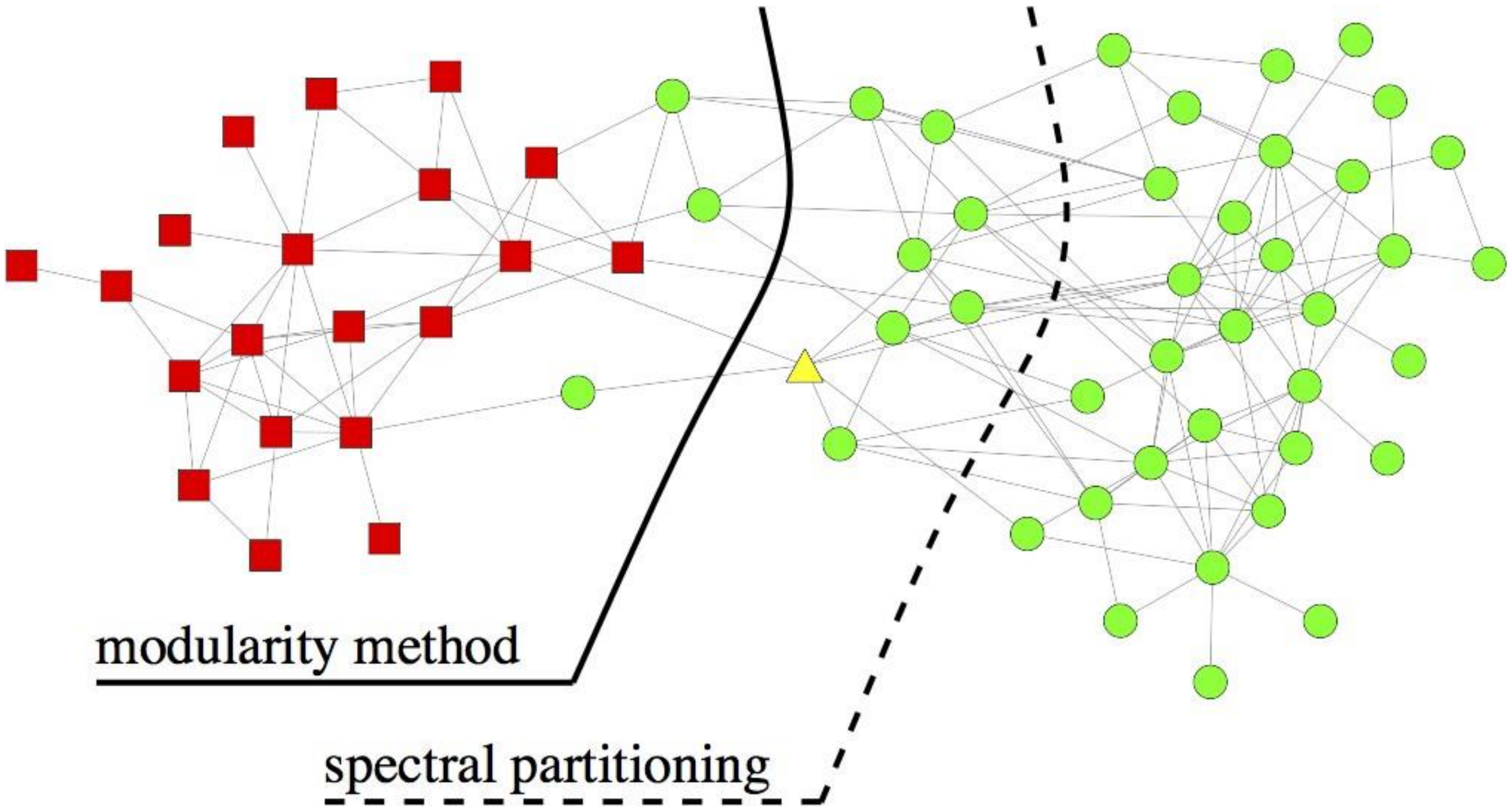
MODULARITY SCORE



best: clusters = 6, modularity = 0.345




34
33
31
9
32
16
15
23
21
19
10
14
29
3
30
27
28
25
26
24
11
5
17
7
6
12
18
13
20
22
8
4
2
1


modularity method
spectral partitioning

SPECTRAL MODULARITY MAXIMIZATION

M. Newman, 2006

Algorithm: Spectral modularity maximization: two-way partition

Input: adjacency matrix $\mathbf{A}$

Output: class indicator vector $\mathbf{s}$

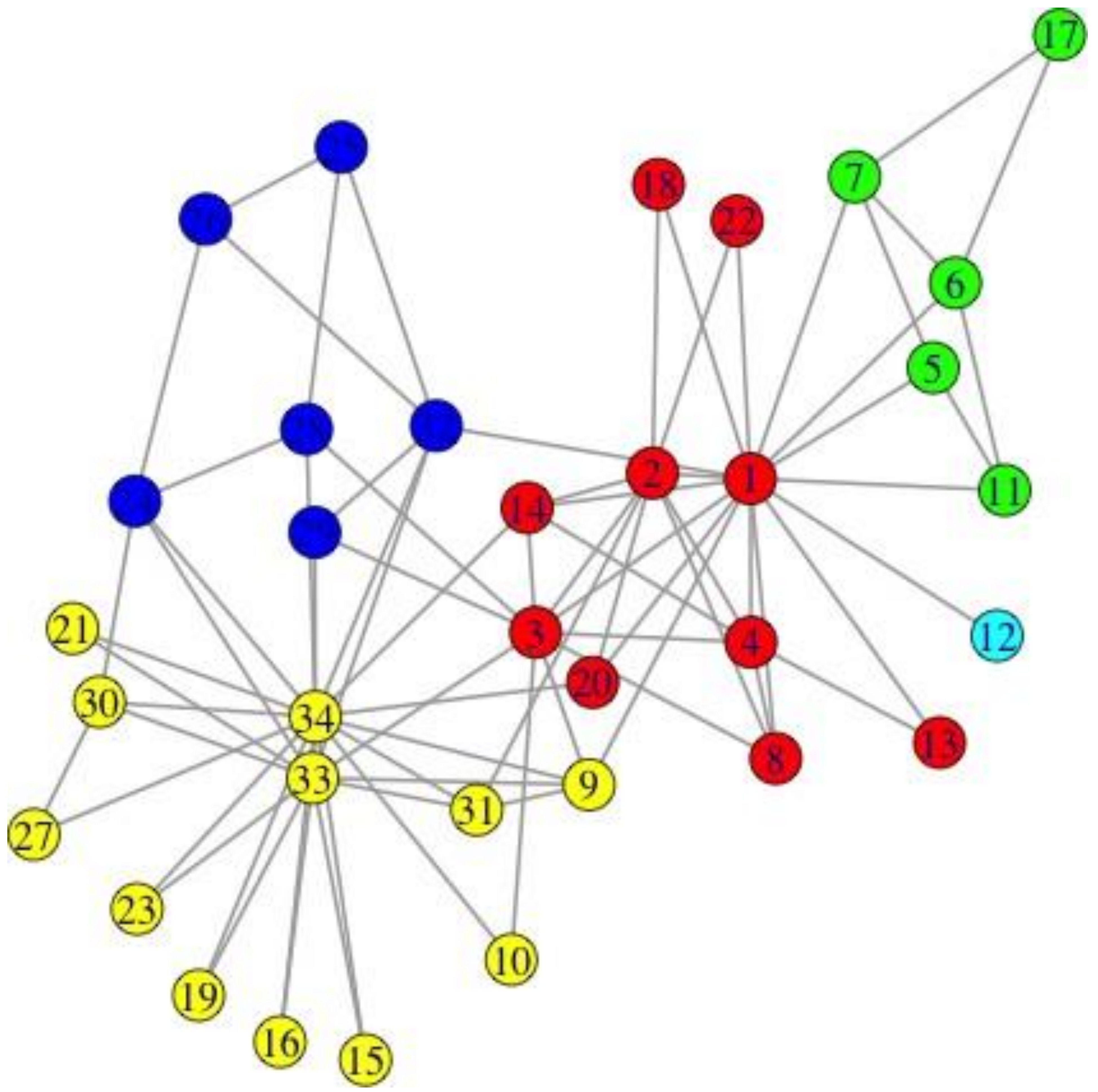
compute $\mathbf{k} = deg(\mathbf{A})$;

compute $\mathbf{B} = \mathbf{A} - \frac{1}{2m}\mathbf{k}\mathbf{k}^T$;

solve for maximal eigenvector $\mathbf{B}\mathbf{x} = \lambda\mathbf{x}$;

set $\mathbf{s} = sign(\mathbf{x}_{max})$

clusters = 5, modularity = 0.437



LABEL PROPAGATION ALGORITHM

U.N. Raghavan, R. Albert, S. Kumara, 2007

Algorithm: Label propagation

Input: Graph G(V,E)

Output: Communities

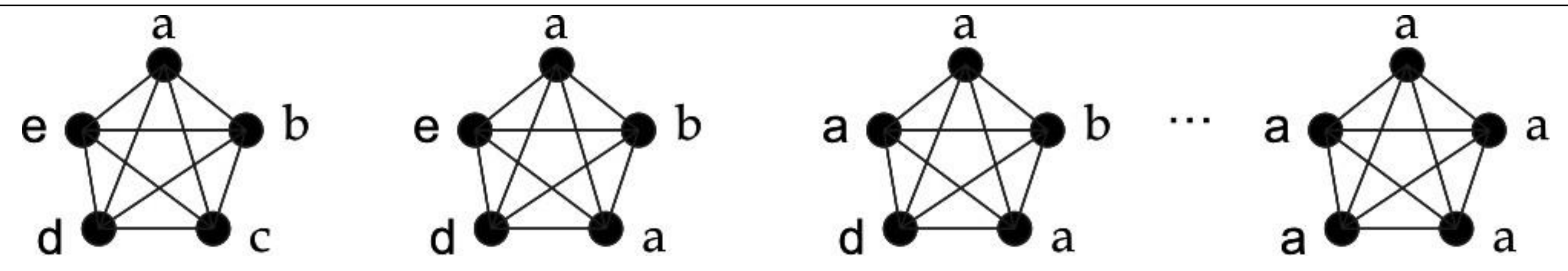
Initialize labels on all nodes;

Randomized node order;

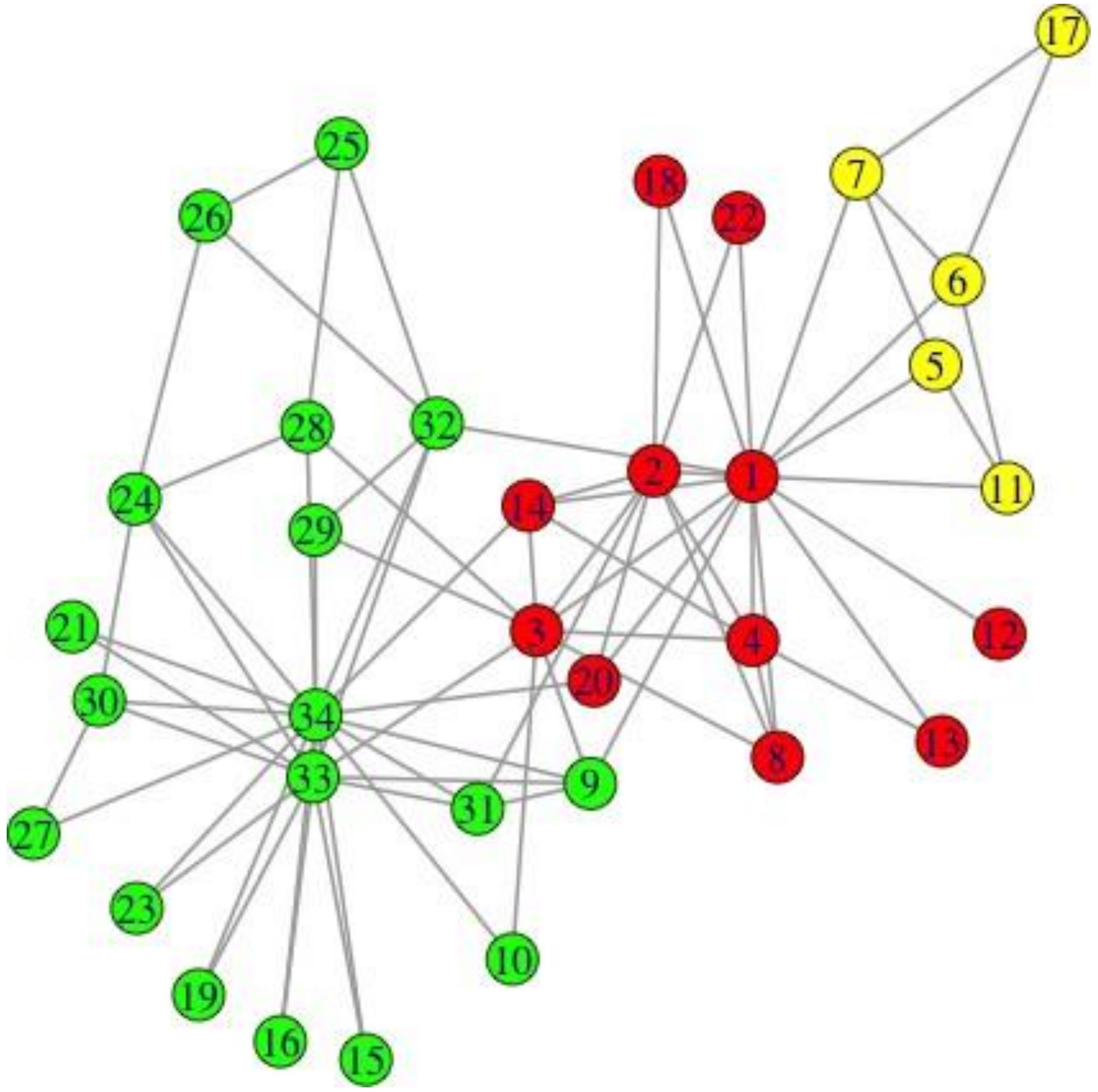
repeat

For every node replace its label with occurring with the highest frequency among neighbors (ties are broken uniformly randomly);

until *every node has a label that the maximum number of the neighbors have*;



clusters = 3, modularity = 0.435



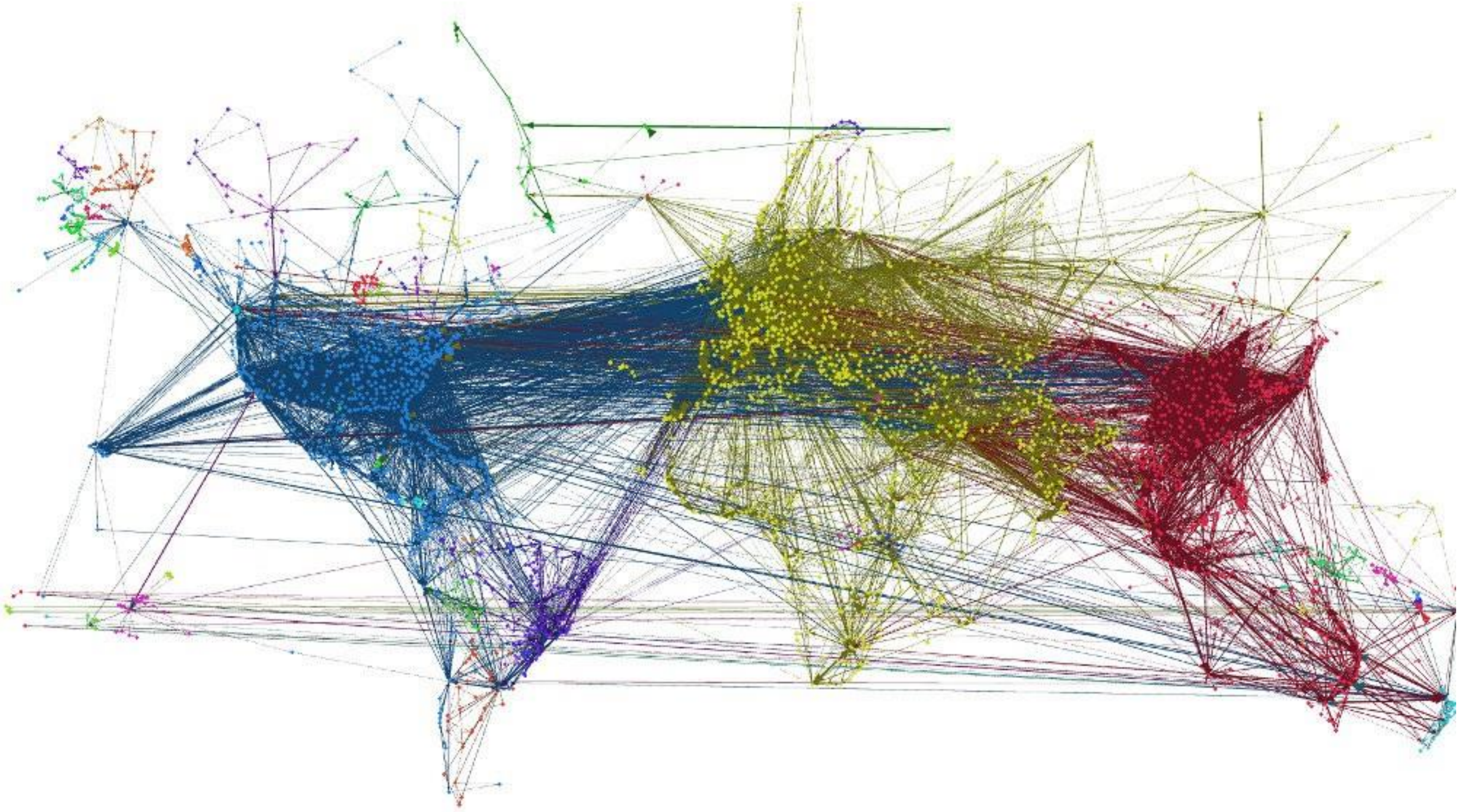


image from Lab41 blog

FAST COMMUNITY UNFOLDING

V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, 2008 ”The Louvain method”

- Heuristic method for greedy modularity optimization
- Find partitions with high modularity
- Multi-level (multi-resolution) hierarchical scheme
- Scalable

V. Blondel et.al., 2008

FAST COMMUNITY UNFOLDING ALGORITHM

Algorithm: Fast unfolding

Input: Graph G(V,E)

Output: Communities

Assign every node to its own community;

repeat

  repeat

    For every node evaluate modularity gain from removing node from its community and placing it in the community of its neighbor;

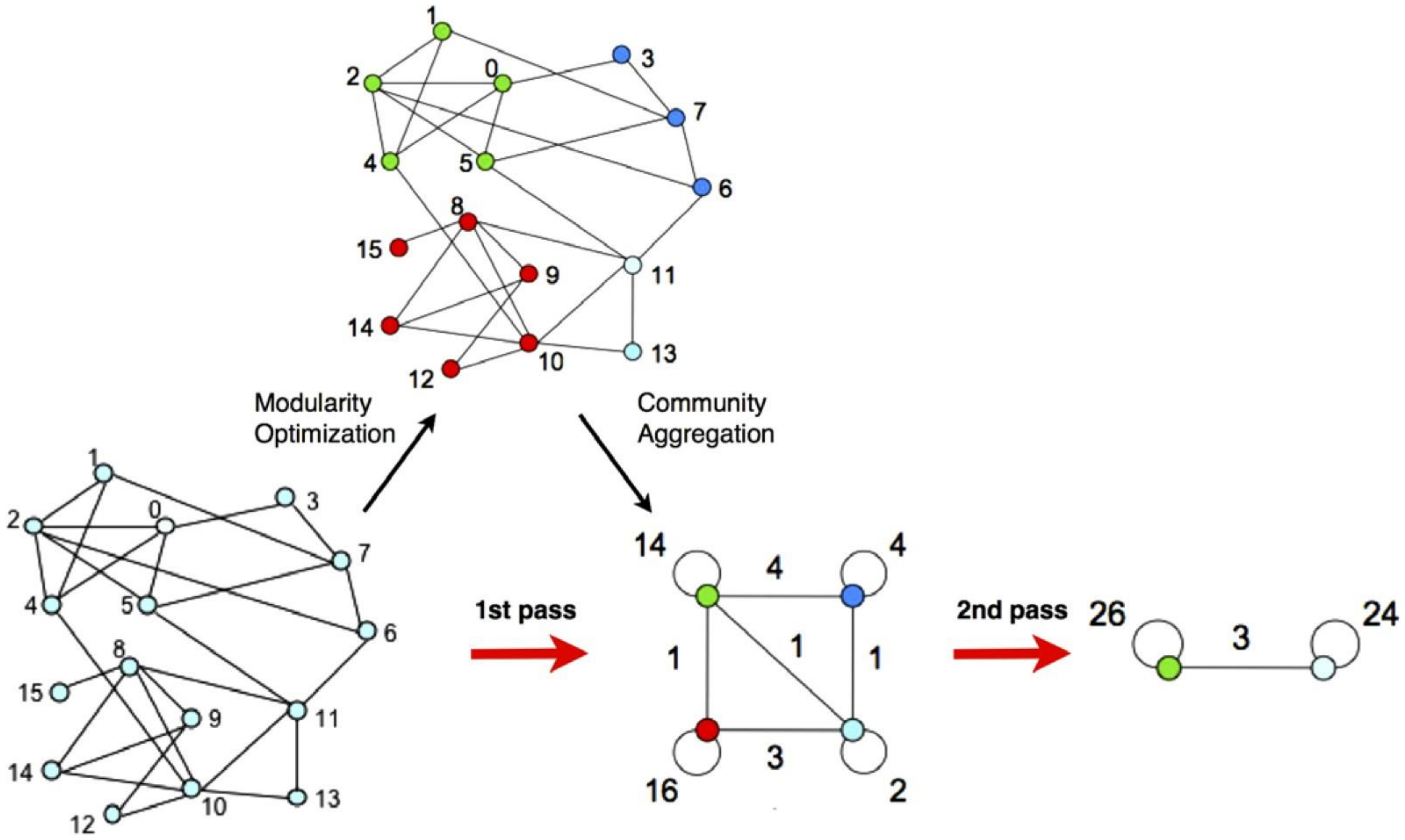
    Place node in the community maximizing modularity gain;

  until *no more improvement (local max of modularity)*;

  Nodes from communities merged into "super nodes" ;

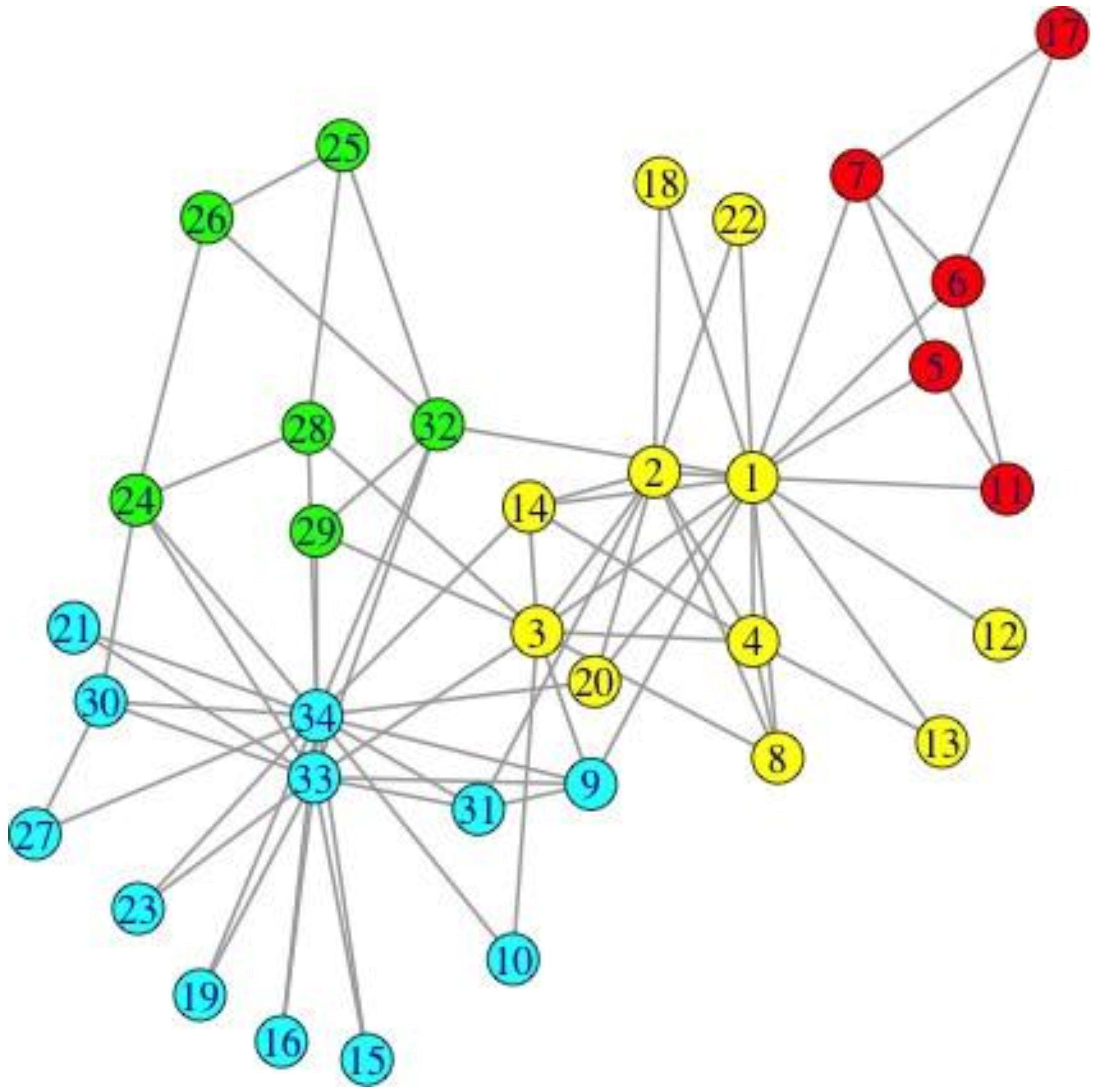
  Weight on the links added up

until *no more changes (max modularity)*;



V. Blondel et.al., 2008

clusters = 4, modularity = 0.445



WALKTRAP

Algorithm: Walktrap community detection

Input: Graph G(V,E)
Output: Dendrogram/communities

Assign each vertex to its own community;

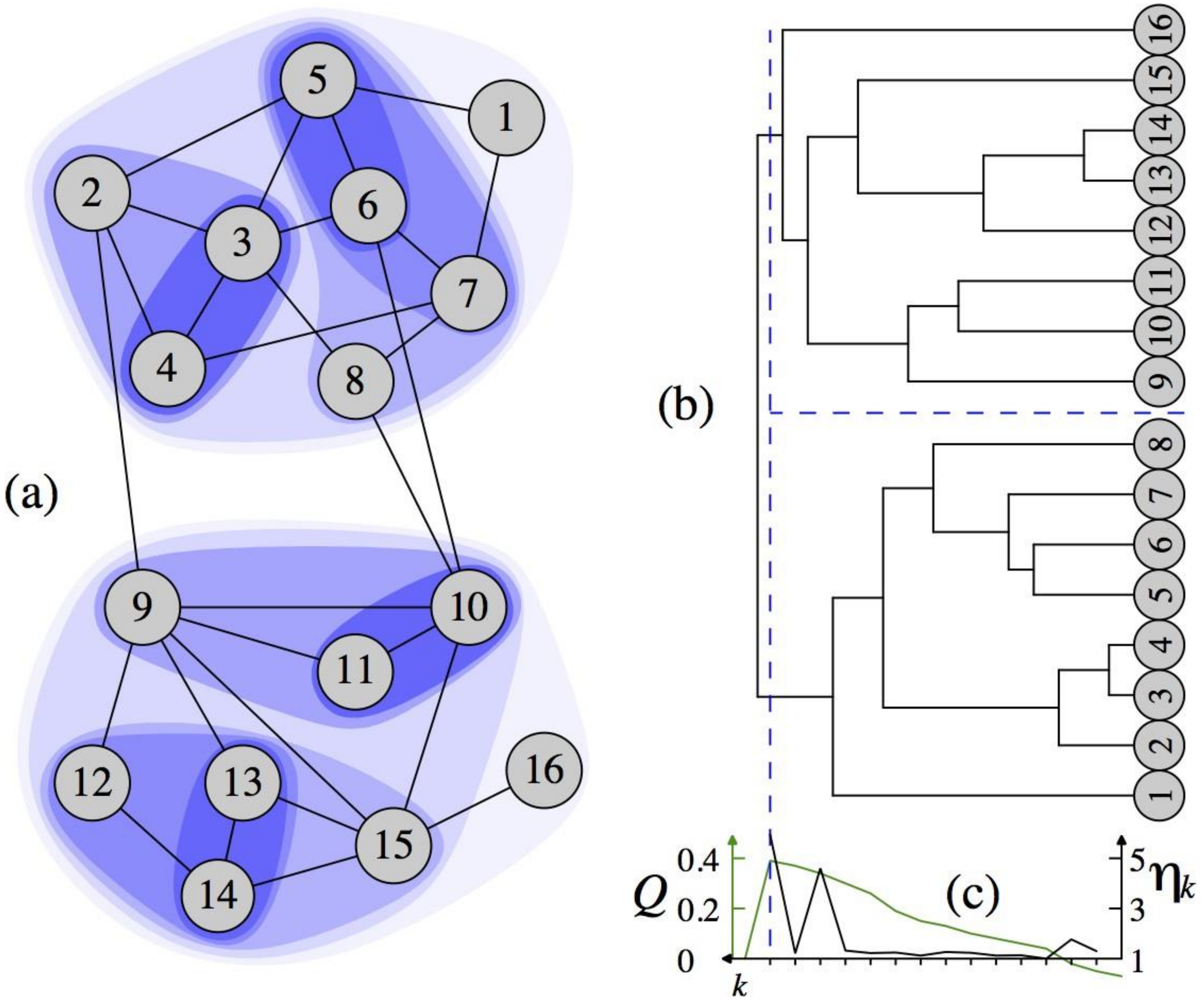
Compute random walk distance between adjacent vertices;

for *n-1 steps* **do**

  choose two ”closest” communities and merge them ;

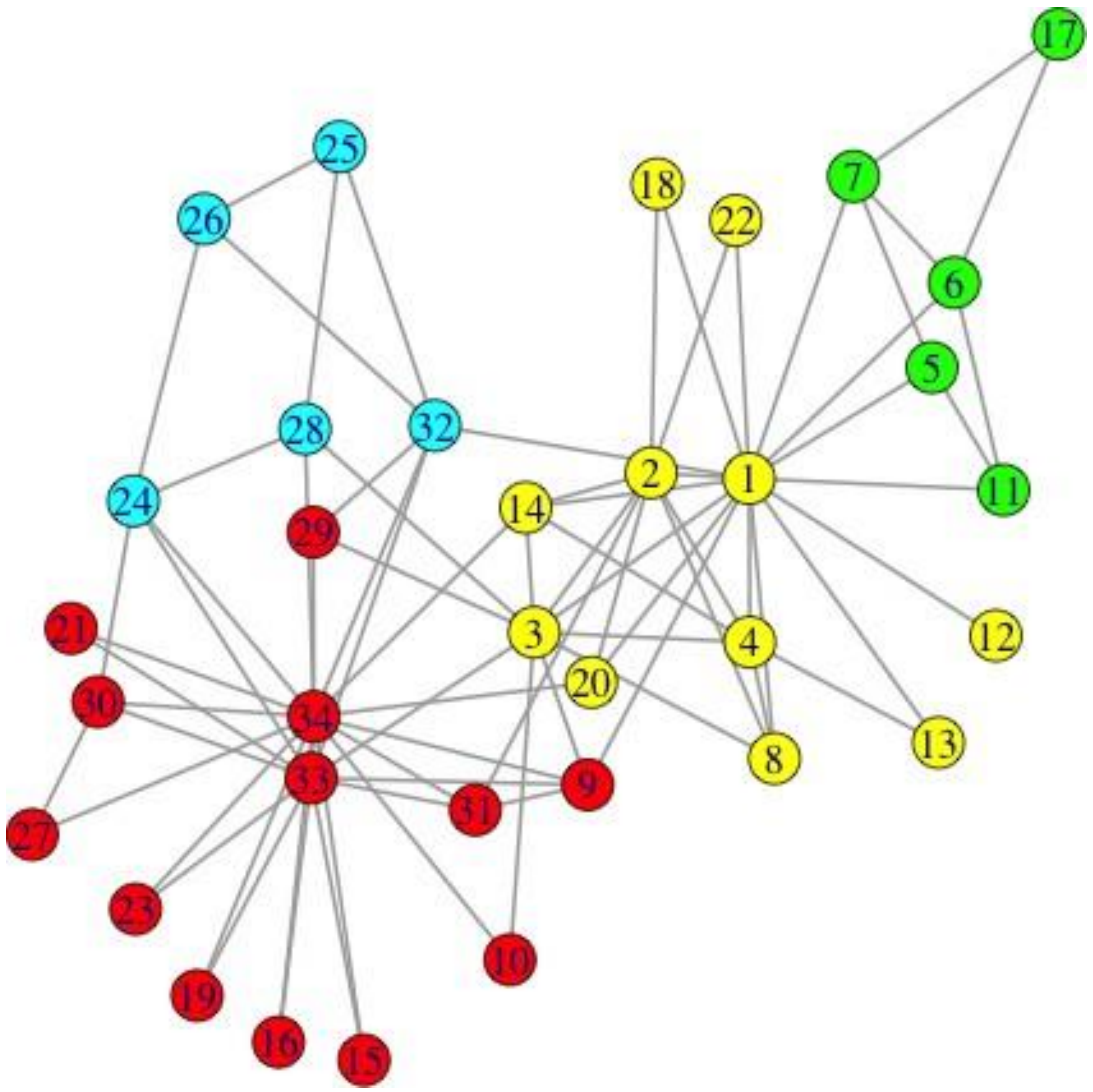
  update distance between communities

end

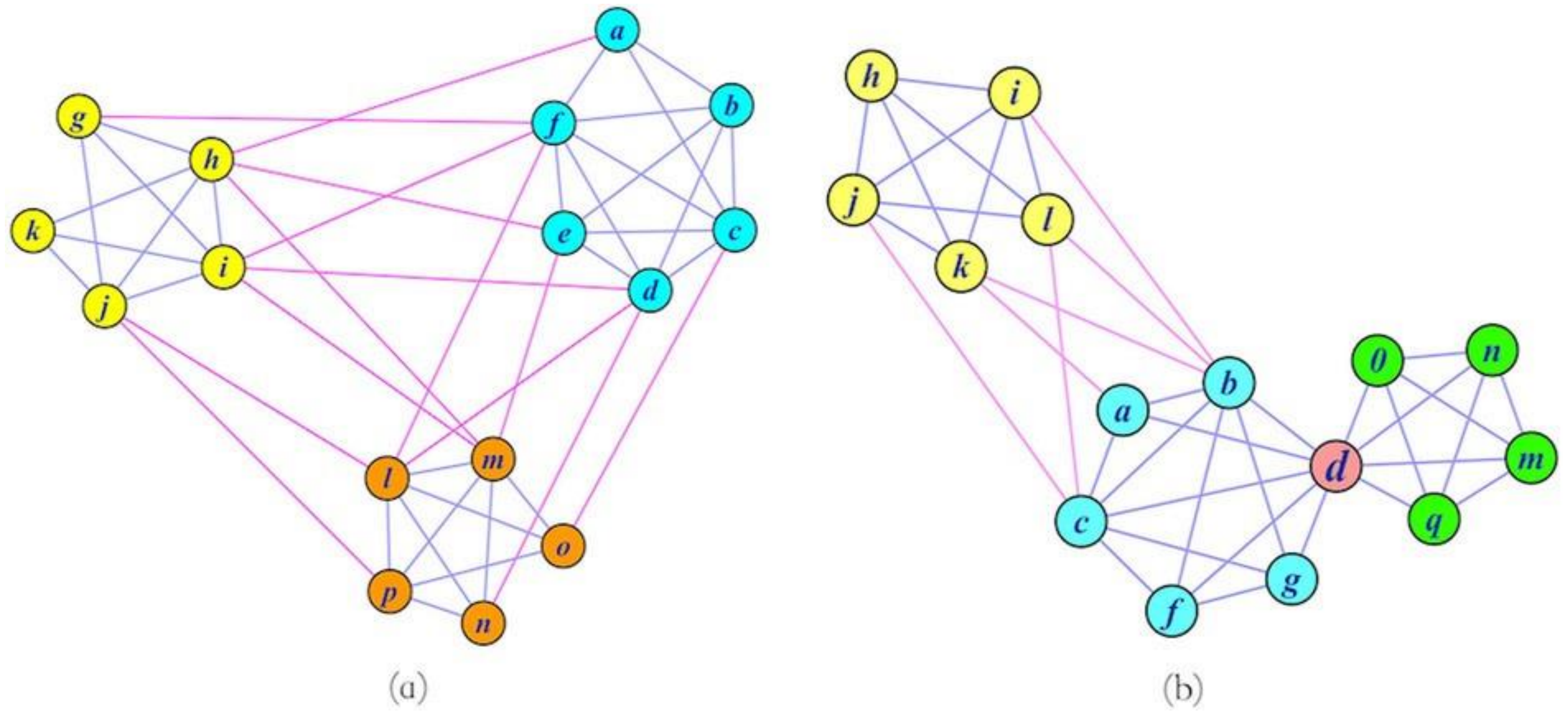


P. Pons and M. Latapy, 2006

clusters = 4, modularity = 0.440



OVERLAPPING COMMUNITIES



Community detection:

Graph partitioning (sparse cuts)

Vertex clustering (vertex similarity)

image from W. Liu , 2014

Author	Ref.	Label	Order
Eckmann & Moses	(Eckmann and Moses, 2002)	EM	$O(m\langle k^2\rangle)$
Zhou & Lipowsky	(Zhou and Lipowsky, 2004)	ZL	$O(n^3)$
Latapy & Pons	(Latapy and Pons, 2005)	LP	$O(n^3)$
Clauset et al.	(Clauset *et al.*, 2004)	NF	$O(n \log^2 n)$
Newman & Girvan	(Newman and Girvan, 2004)	NG	$O(nm^2)$
Girvan & Newman	(Girvan and Newman, 2002)	GN	$O(n^2 m)$
Guimerà et al.	(Guimerà and Amaral, 2005; Guimerà *et al.*, 2004)	SA	parameter dependent
Duch & Arenas	(Duch and Arenas, 2005)	DA	$O(n^2 \log n)$
Fortunato et al.	(Fortunato *et al.*, 2004)	FLM	$O(m^3 n)$
Radicchi et al.	(Radicchi *et al.*, 2004)	RCCLP	$O(m^4/n^2)$
Donetti & Muñoz	(Donetti and Muñoz, 2004, 2005)	DM/DMN	$O(n^3)$
Bagrow & Bollt	(Bagrow and Bollt, 2005)	BB	$O(n^3)$
Capocci et al.	(Capocci *et al.*, 2005)	CSCC	$O(n^2)$
Wu & Huberman	(Wu and Huberman, 2004)	WH	$O(n+m)$
Palla et al.	(Palla *et al.*, 2005)	PK	$O(\exp(n))$
Reichardt & Bornholdt	(Reichardt and Bornholdt, 2004)	RB	parameter dependent

Author	Ref.	Label	Order
Girvan & Newman	(Girvan and Newman, 2002; Newman and Girvan, 2004)	GN	$O(nm^2)$
Clauset et al.	(Clauset *et al.*, 2004)	Clauset et al.	$O(n \log^2 n)$
Blondel et al.	(Blondel *et al.*, 2008)	Blondel et al.	$O(m)$
Guimerà et al.	(Guimerà and Amaral, 2005; Guimerà *et al.*, 2004)	Sim. Ann.	parameter dependent
Radicchi et al.	(Radicchi *et al.*, 2004)	Radicchi et al.	$O(m^4/n^2)$
Palla et al.	(Palla *et al.*, 2005)	Cfinder	$O(\exp(n))$
Van Dongen	(Dongen, 2000a)	MCL	$O(nk^2)$, $k < n$ parameter
Rosvall & Bergstrom	(Rosvall and Bergstrom, 2007)	Infomod	parameter dependent
Rosvall & Bergstrom	(Rosvall and Bergstrom, 2008)	Infomap	$O(m)$
Donetti & Muñoz	(Donetti and Muñoz, 2004, 2005)	DM	$O(n^3)$
Newman & Leicht	(Newman and Leicht, 2007)	EM	parameter dependent
Ronhovde & Nussinov	(Ronhovde and Nussinov, 2009)	RN	$O(m^\beta \log n)$, $\beta \sim 1.3$

Fortunato, 2010

REFERENCES

- S. Fortunato. Community detection in graphs, Physics Reports, Vol. 486, Iss. 35, pp 75-174, 2010
- S. E. Schaeffer. Graph clustering. Computer Science Review, 1(1):2764, 2007.
- Modularity and community structure in networks, M.E.J. Newman, PNAS, vol 103, no 26, pp 8577-8582, 2006
- Finding and evaluating community structure in networks, M.E.J. Newman, M. Girvan, Phys. Rev E, 69, 2004
- U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (3) (2007) 036106.
- G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814?818.
- P. Pons and M. Latapy, Computing communities in large networks using random walks, Journal of Graph Algorithms and Applications, 10 (2006), 191-218.
- V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. P10008 (2008).

CLUSTERING METHODS