Oh Pott, Oh Pott!

or how to detect community structure in complex networks

Jörg Reichardt
Interdisciplinary Centre for Bioinformatics,
Leipzig, Germany
(Host of the 2012 Olympics)

Questions to start from…

Why Communities?

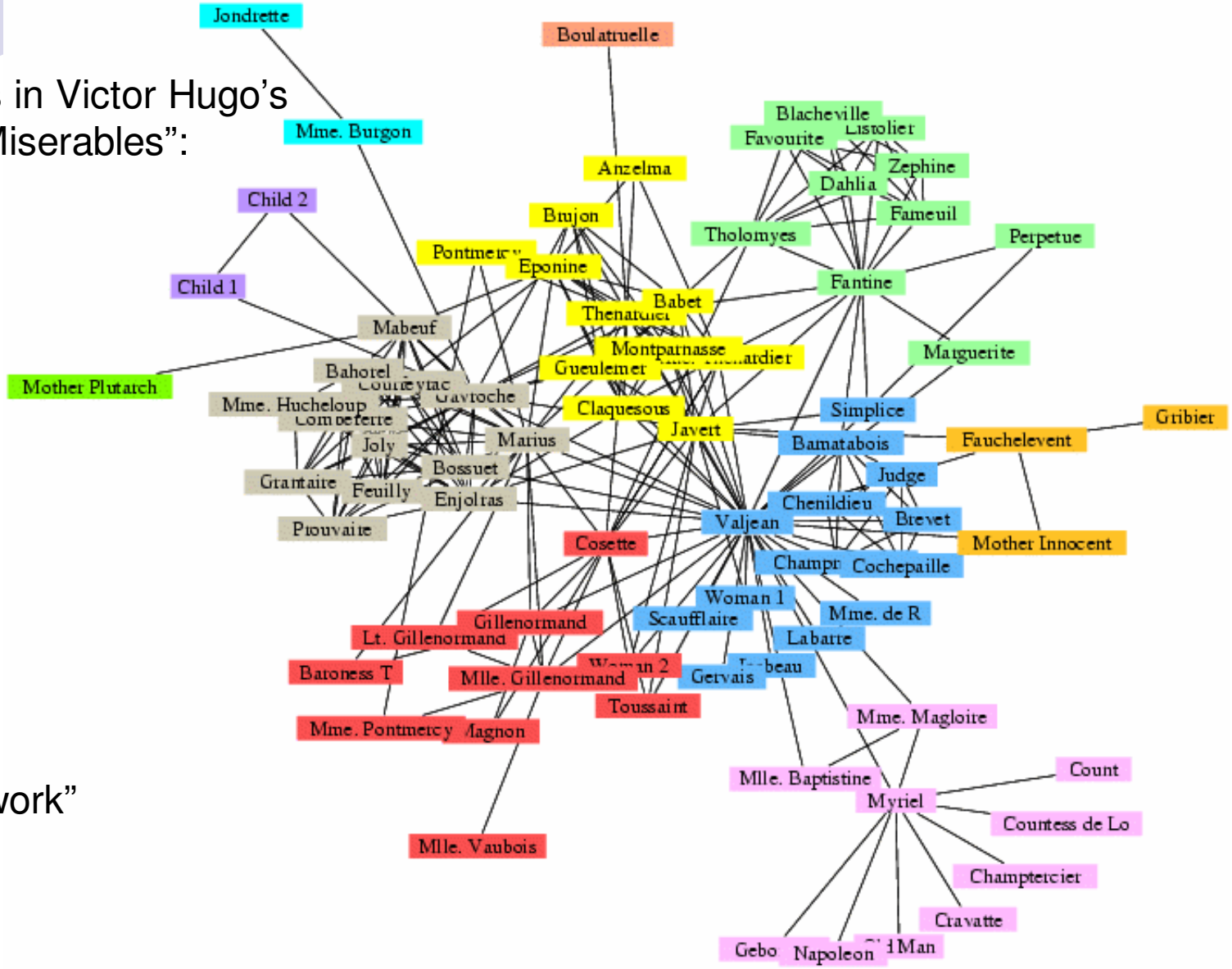
How to find members of communities today?

Practical relevance, Applications?

All finished or are improvements possible? (How to find members of a community tomorrow...)

Different communities in different networks:

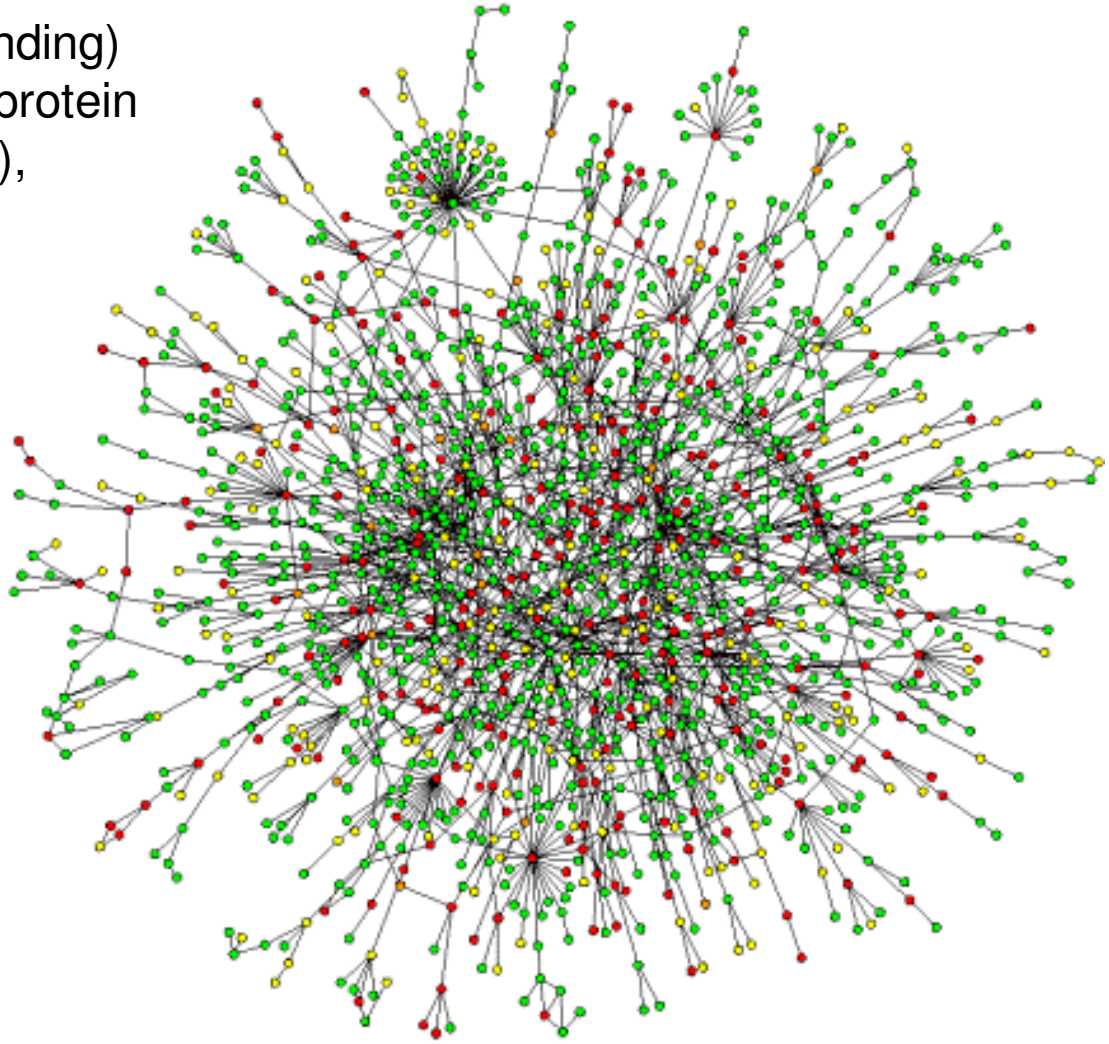
Protagonists in Victor Hugo's novel "Les Miserables":



"Social network"

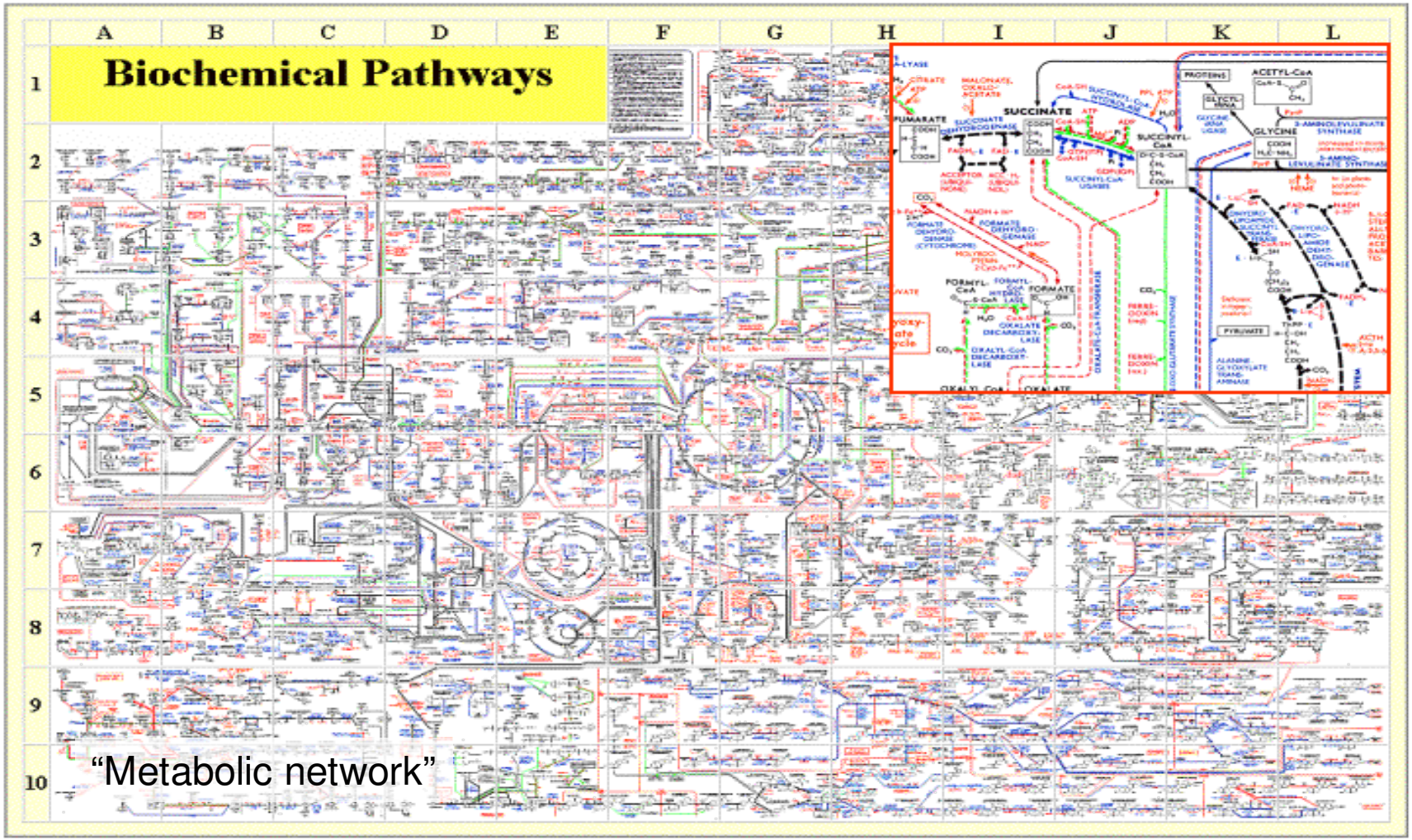
Different communities in different networks:

Protein-Protein interaction (binding) in yeast (effect of removal of protein deadly (red), harmless (green), unknown (yellow)



"Biological network"

Different communities in different networks:



Communities in networks:

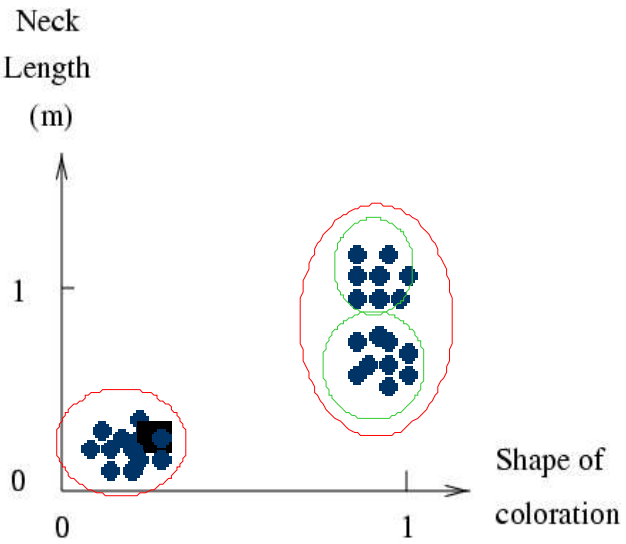
Fact: Many real-world networks display a community structure, e.g. families, groups of close friends in social networks, individual pathways in metabolic networks …

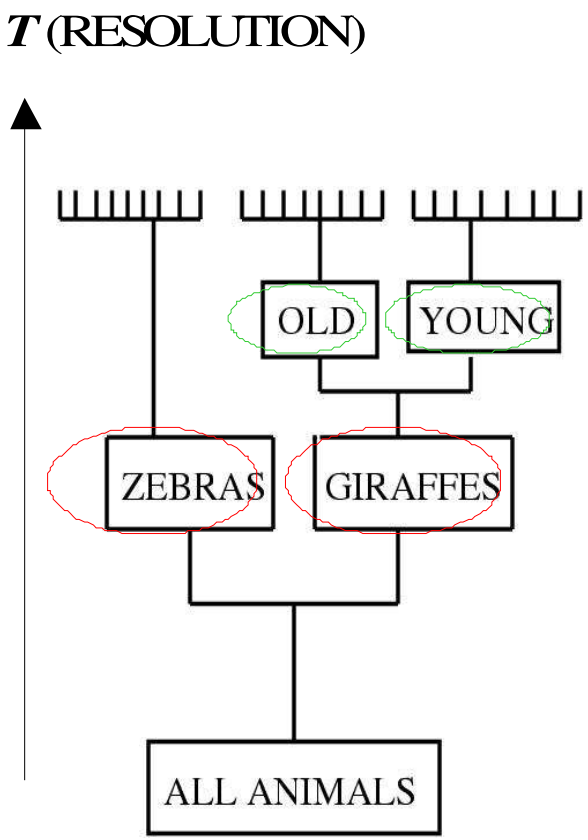
Question: Can we detect the presence of communities in a network and find members of possible communities?

How to find Members of a Community?

The problem of a distance measure for networks…

CLUSTER ANALYSIS YIELDS DENDROGRAM





Multivariate data:
Find groups of points that are close together in attribute space.

Networks:
What does close mean, when L~log(N)

?!

What is a community?

Intuitively: C. is a subgraph V of a graph G with the internal connections denser than the external ones.

Community in a strong sense:

$$k_i^{in}(V) > k_i^{out}(V), \forall i \in V$$

Community in a weak sense:

$$\sum_{i \in V} k_i^{in}(V) > \sum_{i \in V} k_i^{out}(V)$$

Modularity:

$$Q = \sum_c e_{cc} - a_c^2 \qquad a_c = \sum_i e_{ci}$$

Graph-Partioning Heuristics (minimizing the edges to cut, when splitting a graph):

Kernighan-Lin (for balanced partitions):

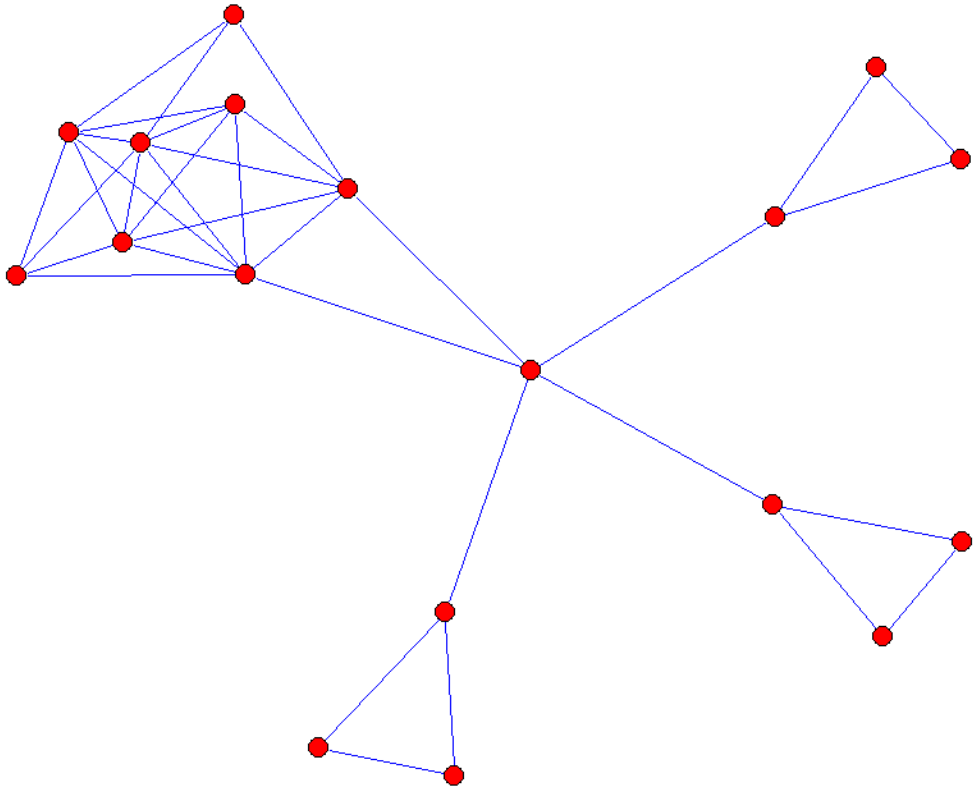
For nodes u,v:

diff(v):= # of links to nodes **out of** community - # of links to nodes **in** community

gain(u,v):=diff(v) + diff(u) – 2 (# of links between u and v)

```
bi-partition graph (randomly)
repeat
        find vertex pair with largest gain
        exchange it
until total number of external edges does not decrease anymore
```

Why it´s not optimal:



Links are not all equal!

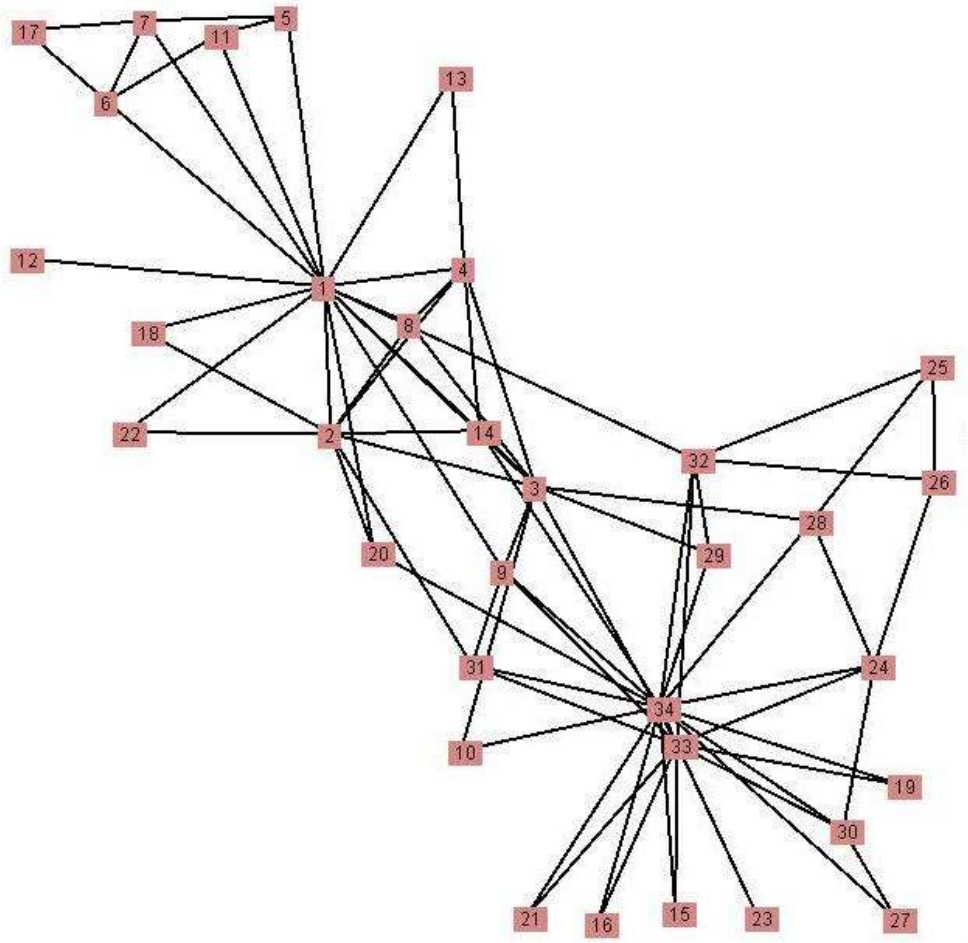
The Girvan-Newman Algorithm:

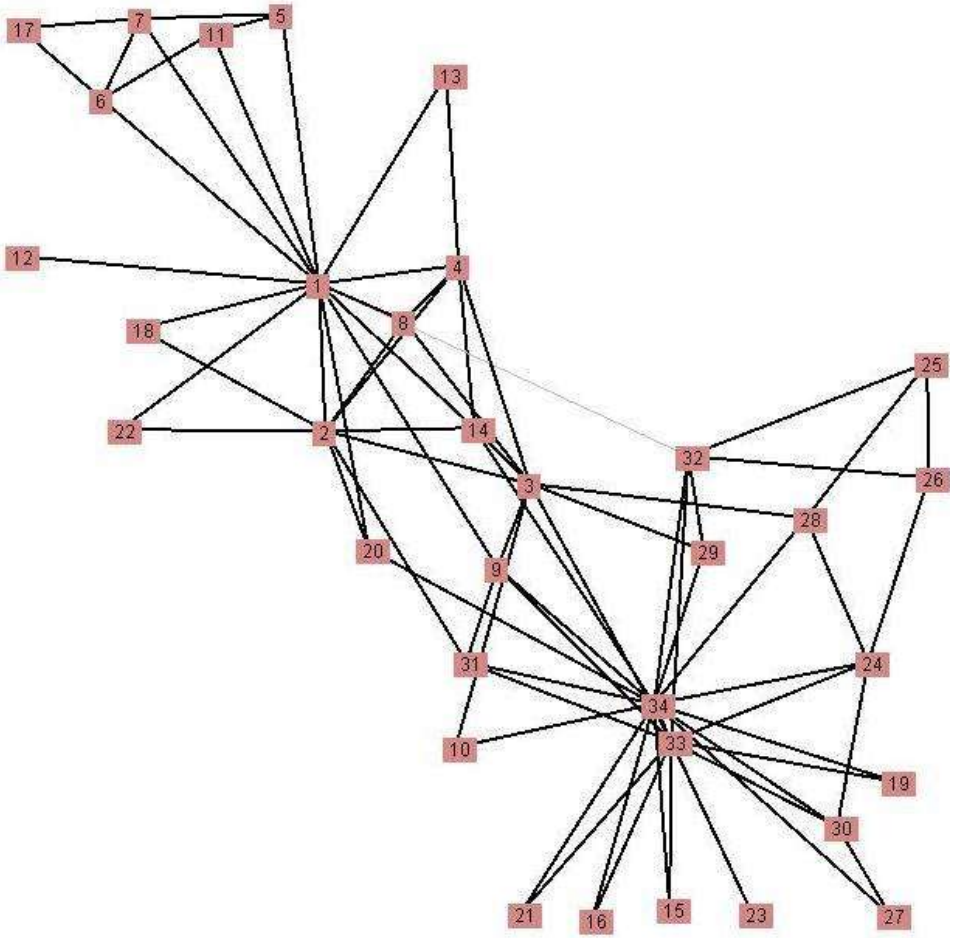
Based on edge betweenness (how many shortest paths between vertices run along a particular edge)

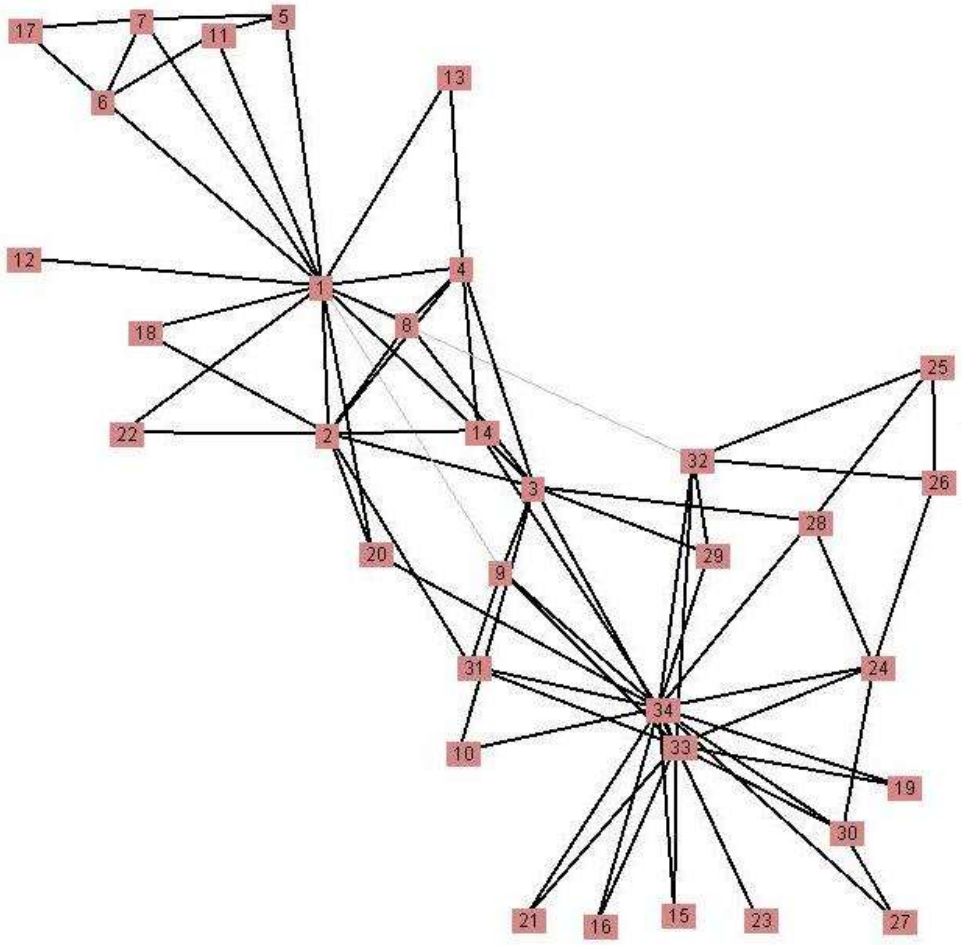
Recursive bi-partioning

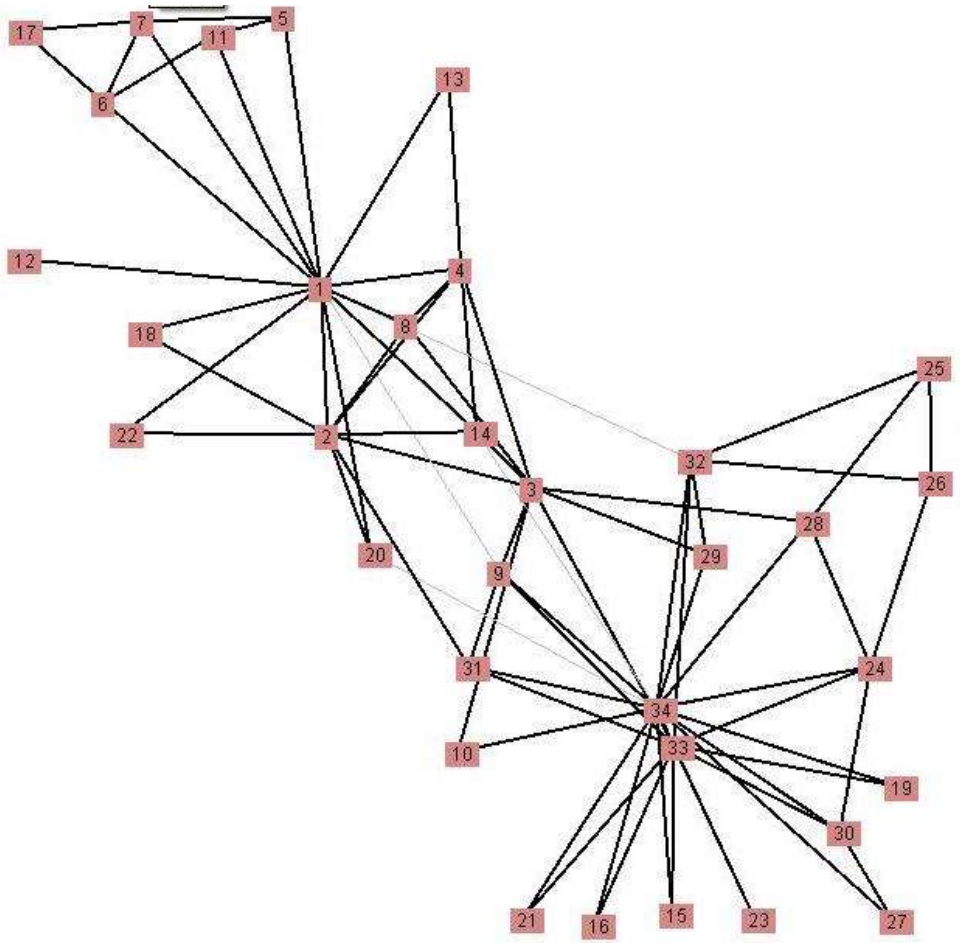
Results in hierarchical clustering

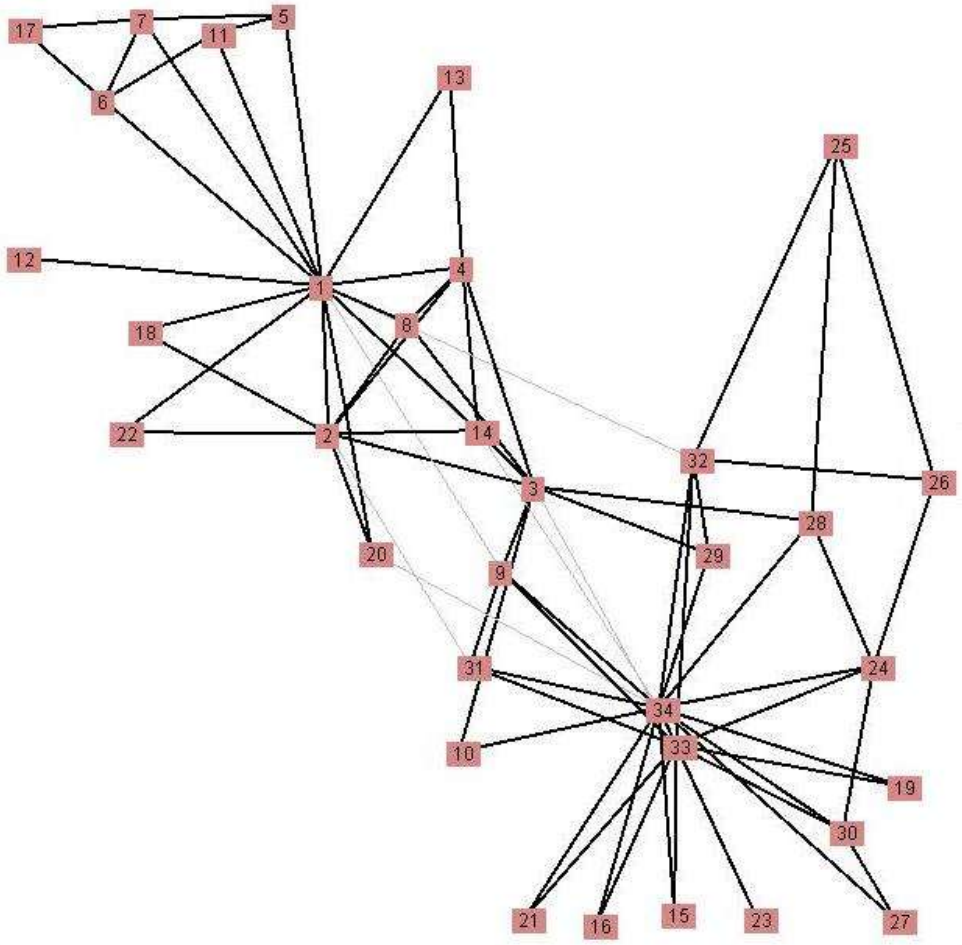
```
repeat
        Calculate edge betweenness of all edges
        Remove edge with highest betweenness
until all edges removed
```

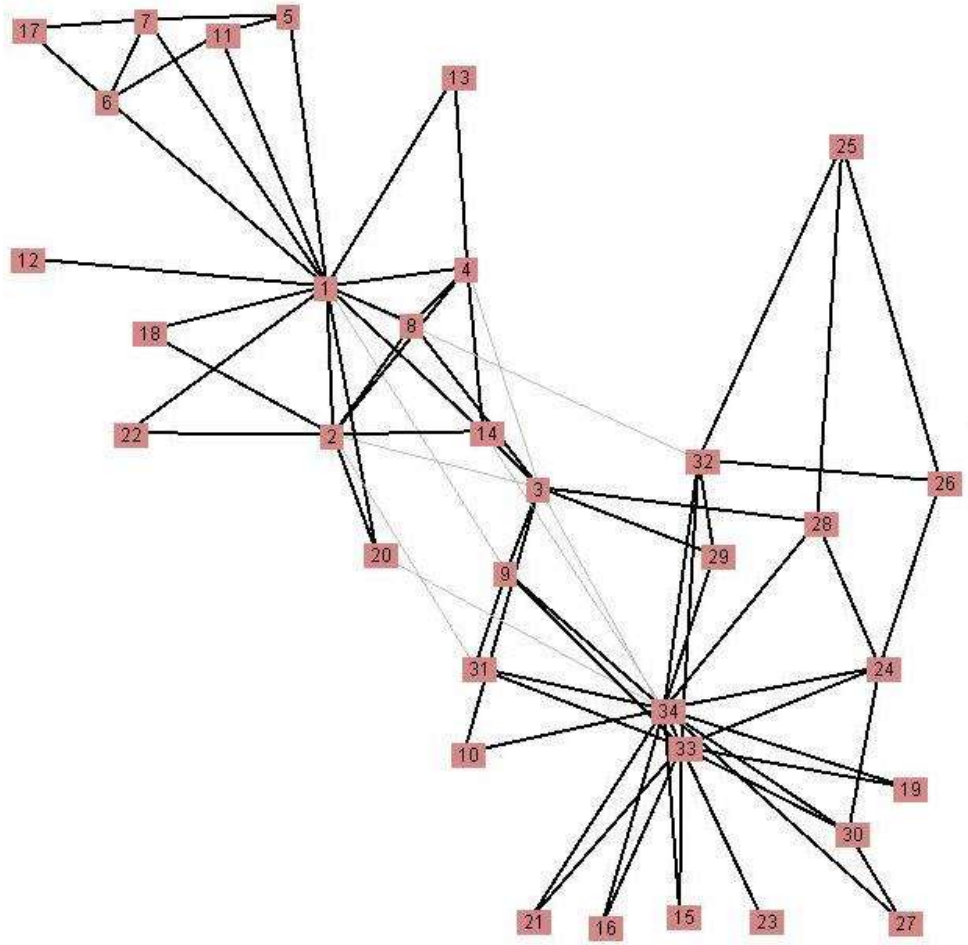

17
7
5
11
6
13
12
4
1
18
8
25
22
2
14
32
26
3
28
20
29
9
31
24
34
10
33
19
30
21
16
15
23
27

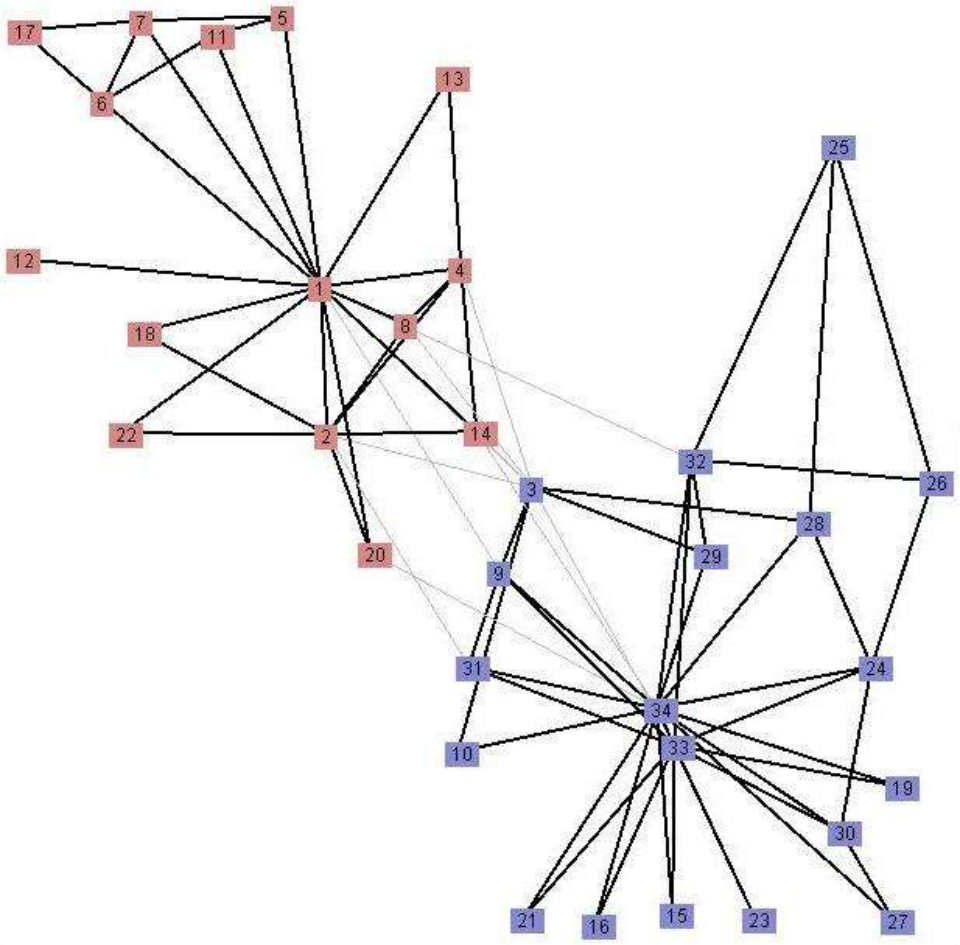

17
7
5
11
6
13
12
1
4
18
8
25
22
2
14
32
26
3
28
20
29
9
31
24
34
10
33
19
30
21
16
15
23
27

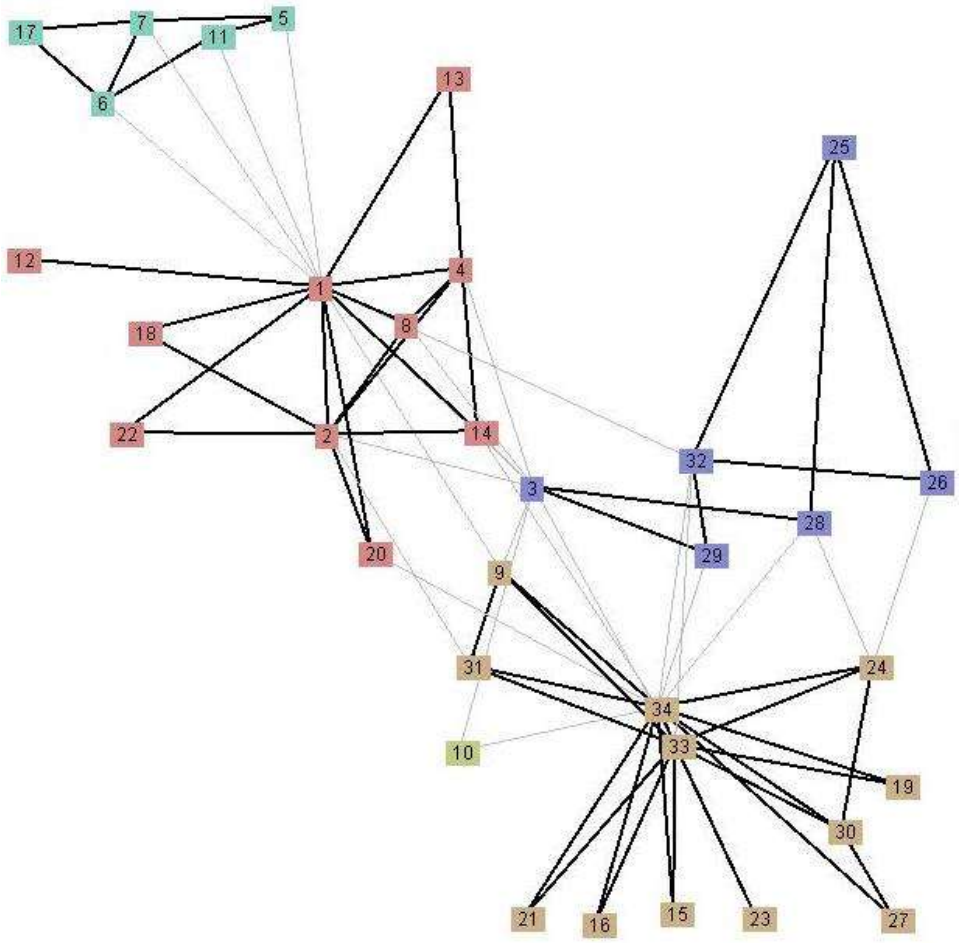

17
7
5
11
6
13
12
1
4
18
8
25
22
2
14
32
26
3
28
20
29
9
31
24
34
10
33
19
30
21
16
15
23
27


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

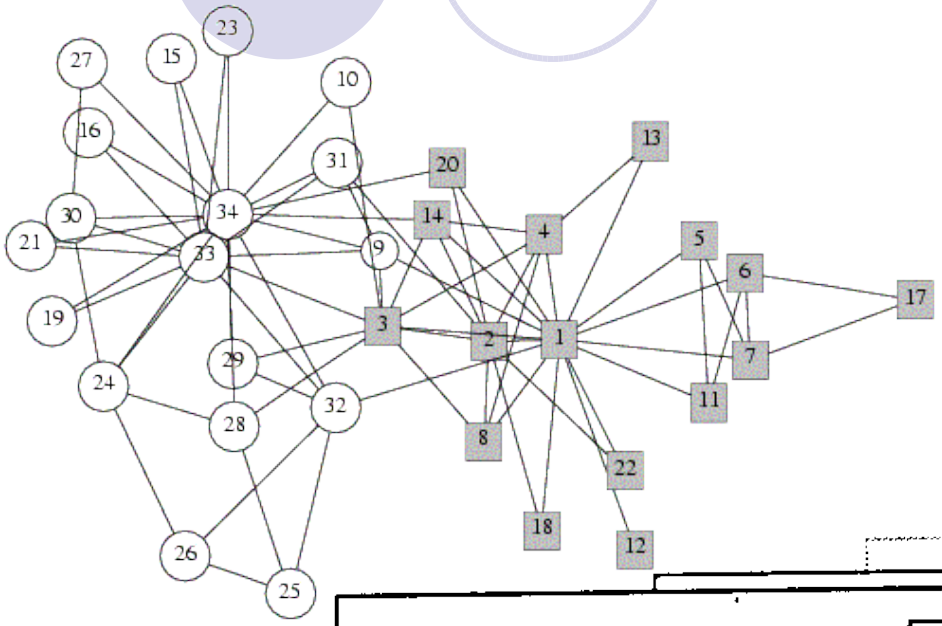

17
7
5
11
6
13
25
12
1
4
18
8
22
2
14
32
26
3
28
20
29
9
31
24
34
10
33
19
30
21
16
15
23
27

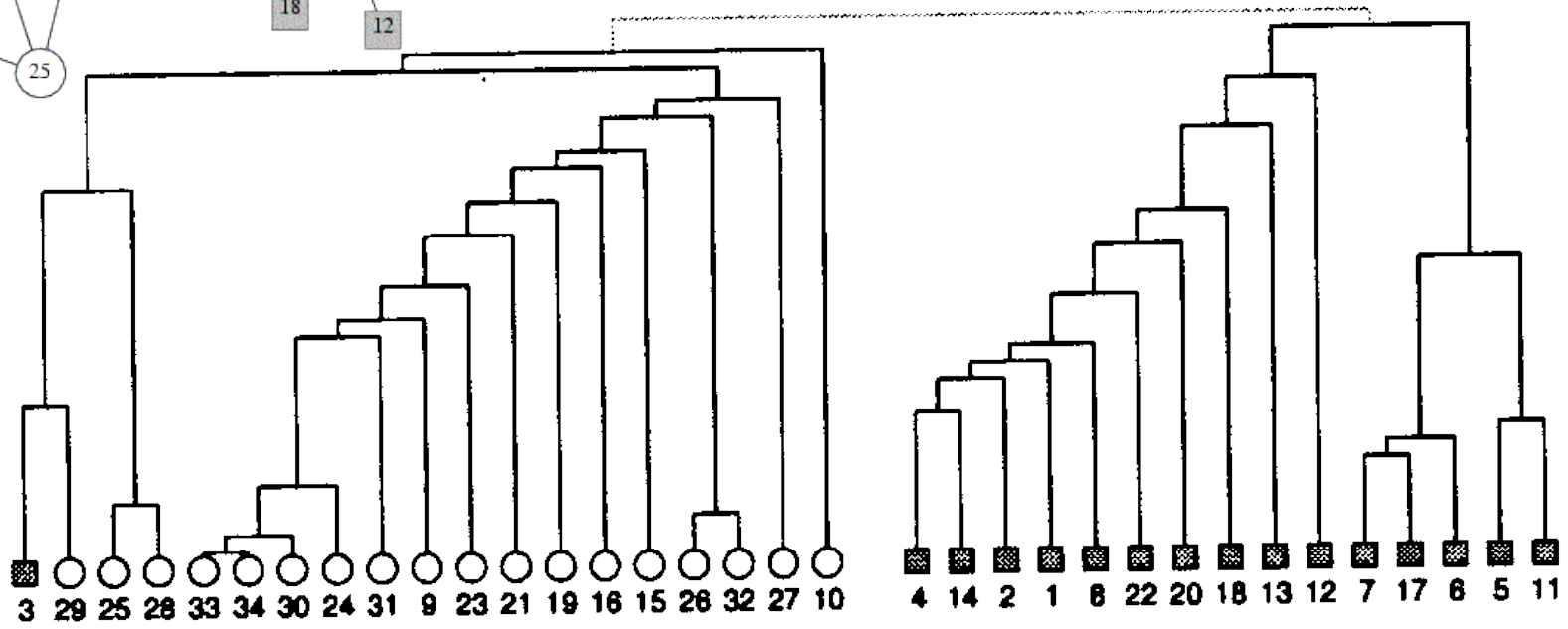

17
7
5
11
6
13
25
12
4
1
18
8
22
2
14
32
26
3
28
20
29
9
31
24
34
10
33
19
30
21
16
15
23
27


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


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

The Zachary Karate Club:



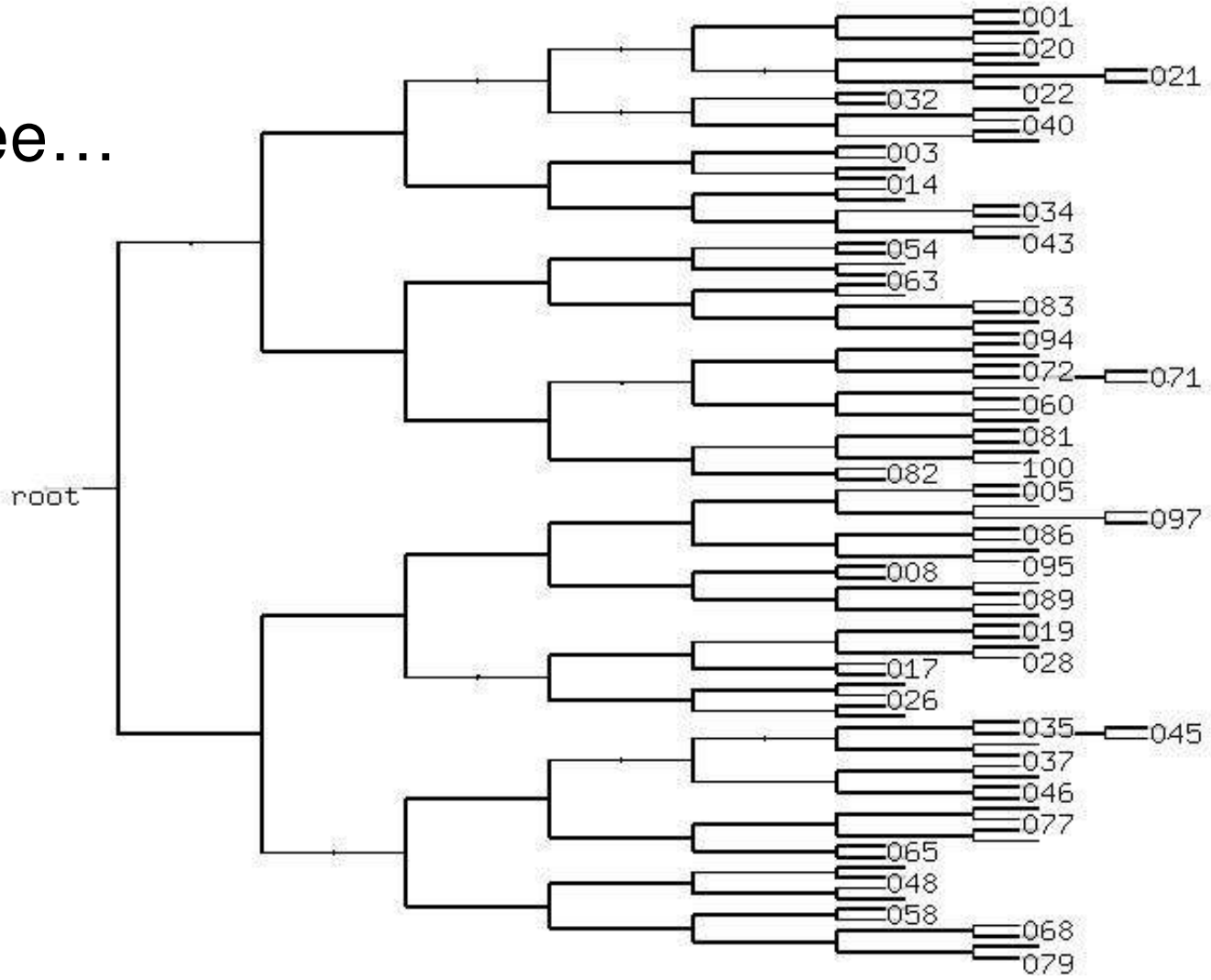


Published Applications (which I know of):

Subnetworks in biochemical pathways
Co-citation networks of genes
Collaboration of Jazz musicians
Subgroups in communication networks
Putative function of proteins

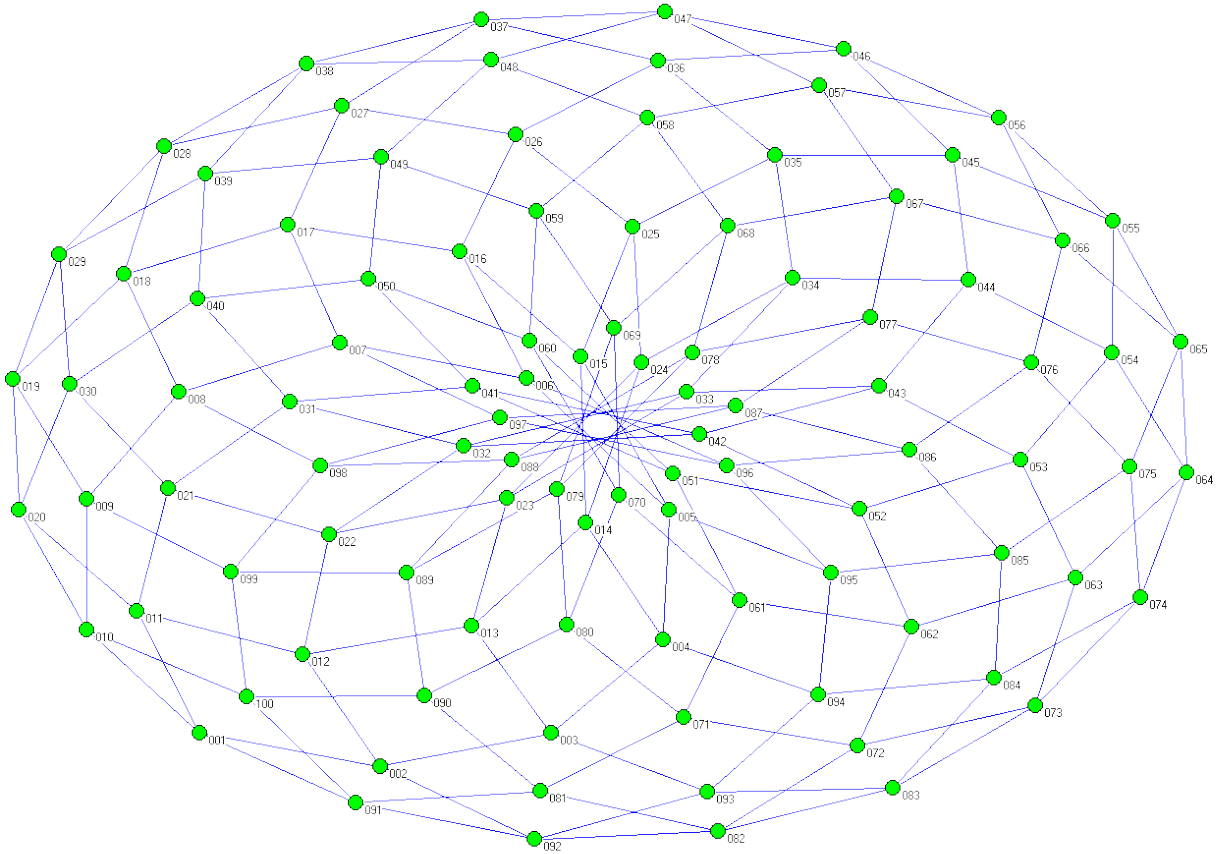
Disadvantages of the GN-algorithm:

A tree…



… and its network…

… 10x10 grid with periodic boundary conditions …



Community detection is not graph partitioning!

Furthermore:

Algorithm is deterministic – but how robust is the result?

Bi-partioning is conceptually questionable.

Does not allow for overlapping or “fuzzy” communities – a node may not belong to more than one community at the same time.

Detecting Community Structure with a q-state Potts model:

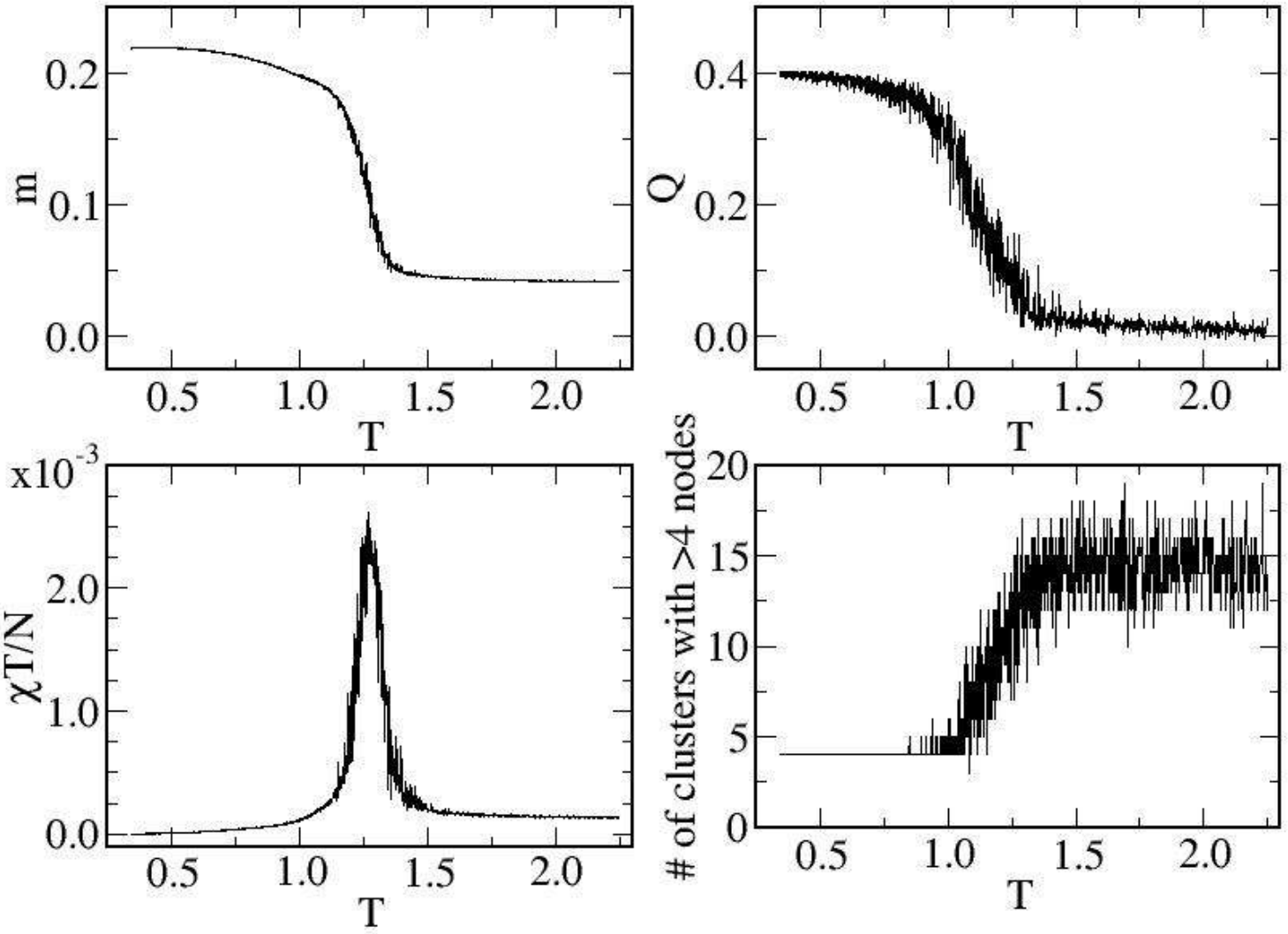
Put Potts spins 1 to q onto the nodes. Then use:

$$H(\{\sigma_i\}) = -J \sum_{(i,j)\in E} \delta(\sigma_i, \sigma_j) + \gamma \sum_{s=1}^{q} \frac{n_s(n_s-1)}{2}$$

Homogeneity Diversity

Find spin configuration for which the energy is minimal. Read off communities as sets of nodes with equal spin.

Cooling down the new Hamiltonian – finding the ground state with simulated annealing:



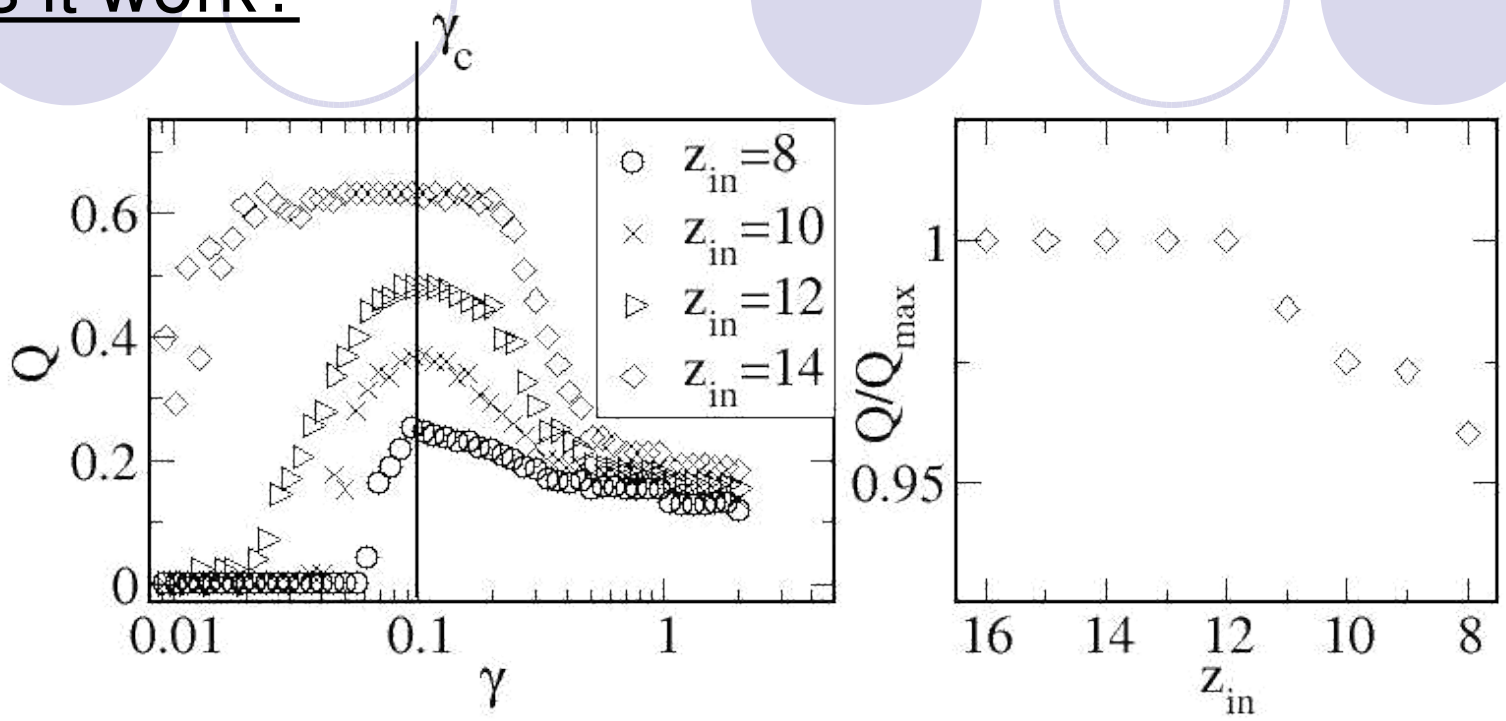
How to set ?

Assuming no knowledge of the network topology, at which does the energy of a homogeneous configuration equal that of completely inhomogeneous system?

$$-Jp\frac{N(N-1)}{2}+\gamma_c\frac{N(N-1)}{2}=-Jpq\frac{N/q(N/q-1)}{2}+\gamma_c q\frac{N/q(N/q-1)}{2}$$

$$Jp=\gamma_c$$

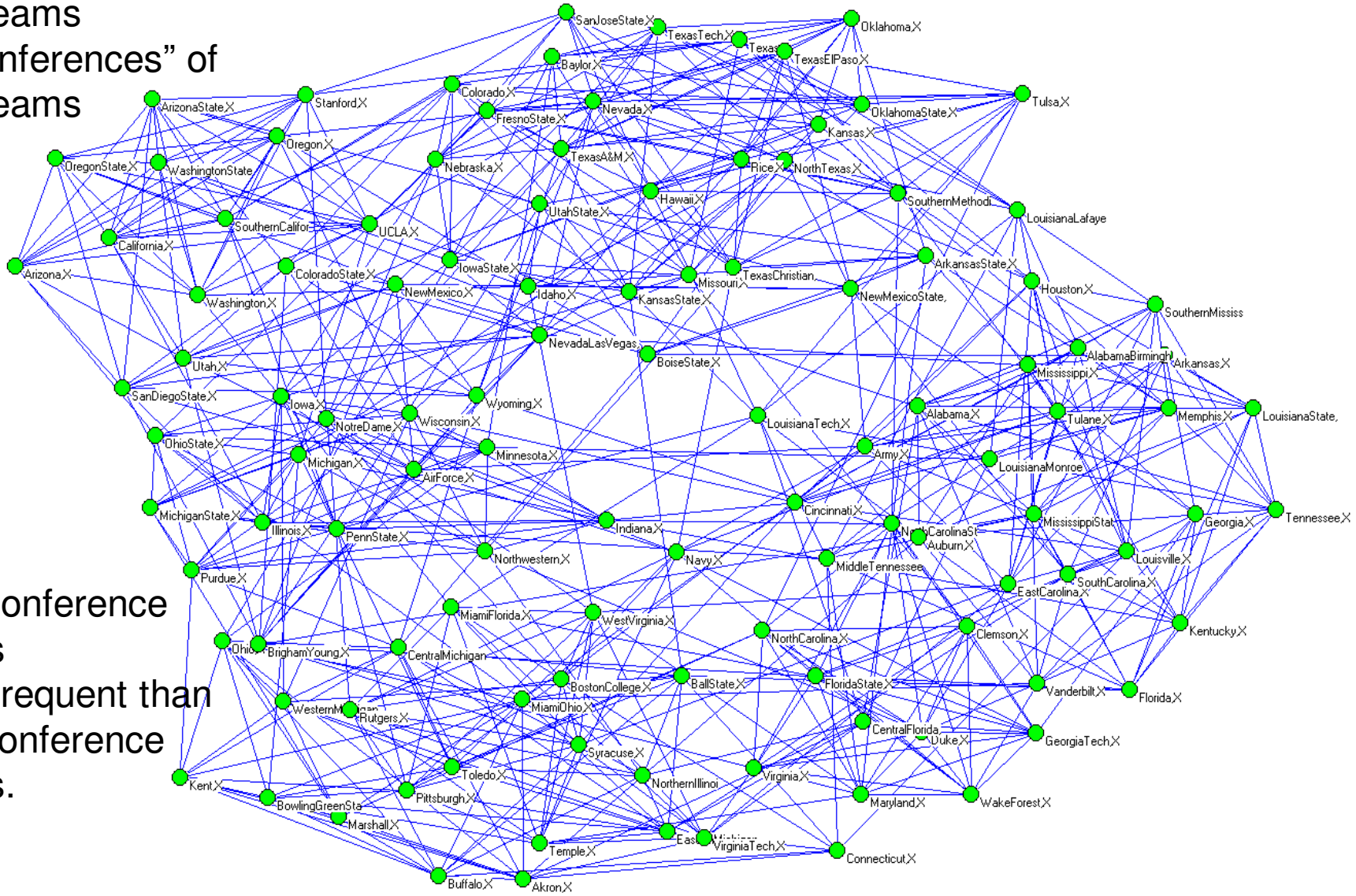
Does it work?



- Computer-generated networks with known community structure (4x32 nodes) and known $Q_{max}=Z_{in}/16-1/4$.
- Z_{in}:intra-community links, z_{out}:inter-community links.
- $Z_{in}+z_{out}$=const.=16, e.g. p=const=0.126, but link density within the communities is different!

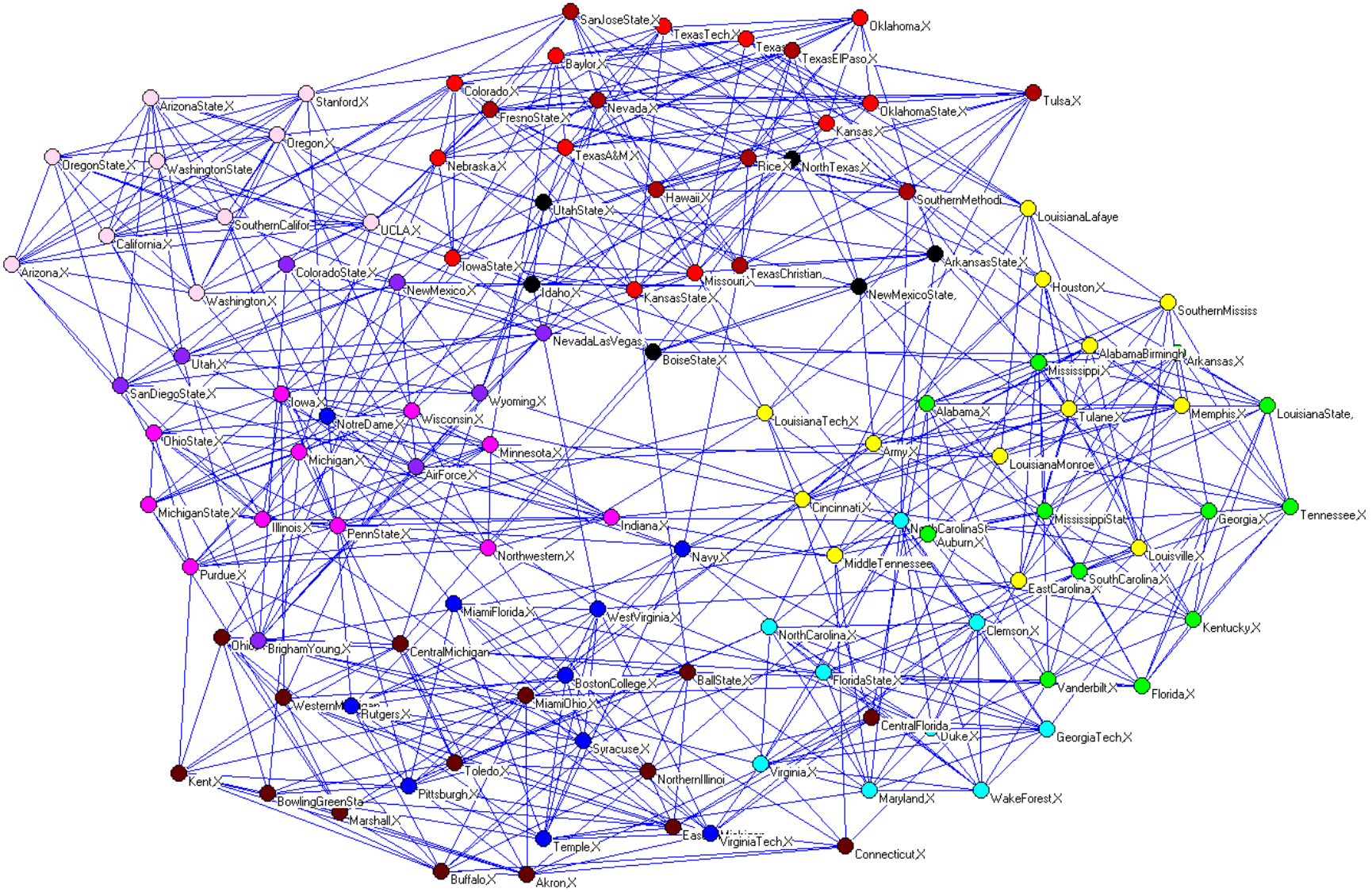
Another example: College Football

115 Teams
11 “conferences” of
7-13 teams

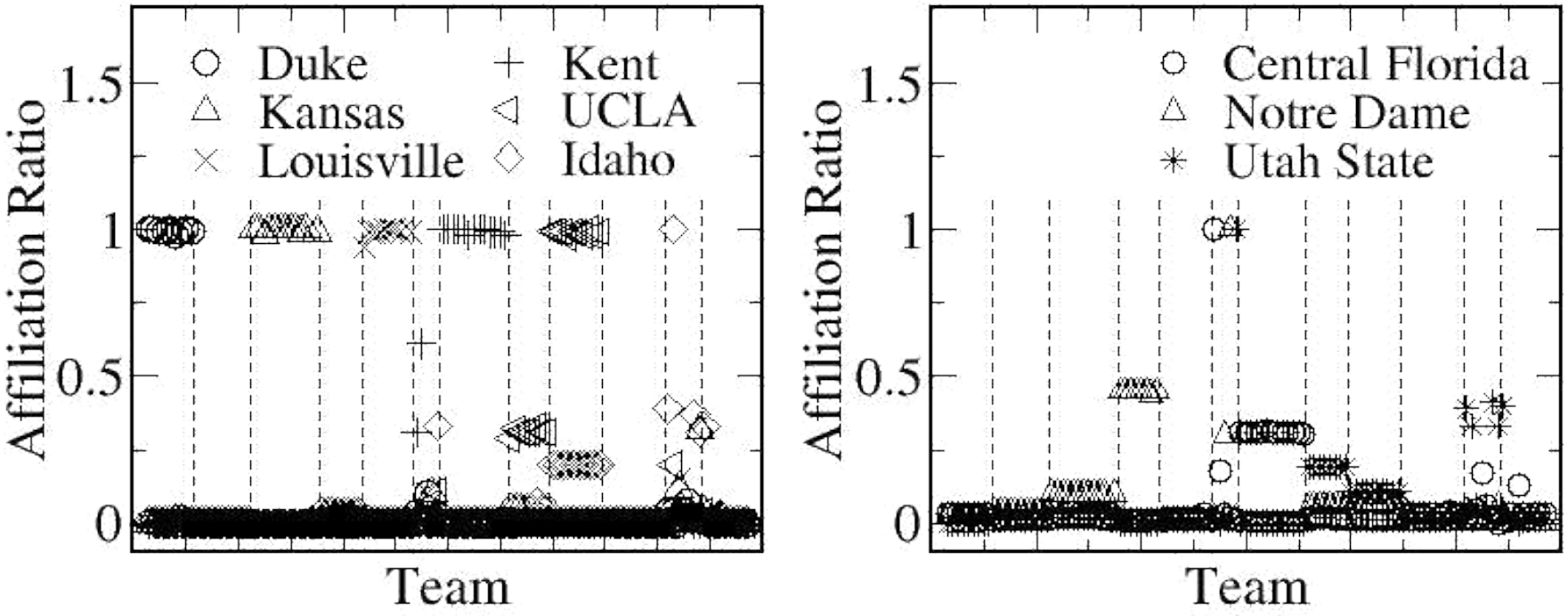


Intra-conference
games
more frequent than
inter-conference
games.

Communities obtained with the Potts-Model



The conference structure is regained, but how stable is it?



Measure how often any two teams end up in the same community when starting from different initial conditions or measure at temperature slightly above “freezing”, e.g. acceptance ratio of 10% or so.

Summary:

- Short introduction to the problem of community detection in complex networks
- Community detection is not graph partioning!
- Presentation of a new algorithm for the detection of “fuzzy” communities that allows for assessment of the stability of the communities.

Acknowledgements:

Stefan Bornholdt

Konstantin Klemm and Thimo Rohlf

Mark Newman for providing network data

Diffusion Approach:

Consider a large number of random walkers on the network:

$$\rho_i(t+\Delta t) = \sum_j T_{ij}\rho_j(t)$$

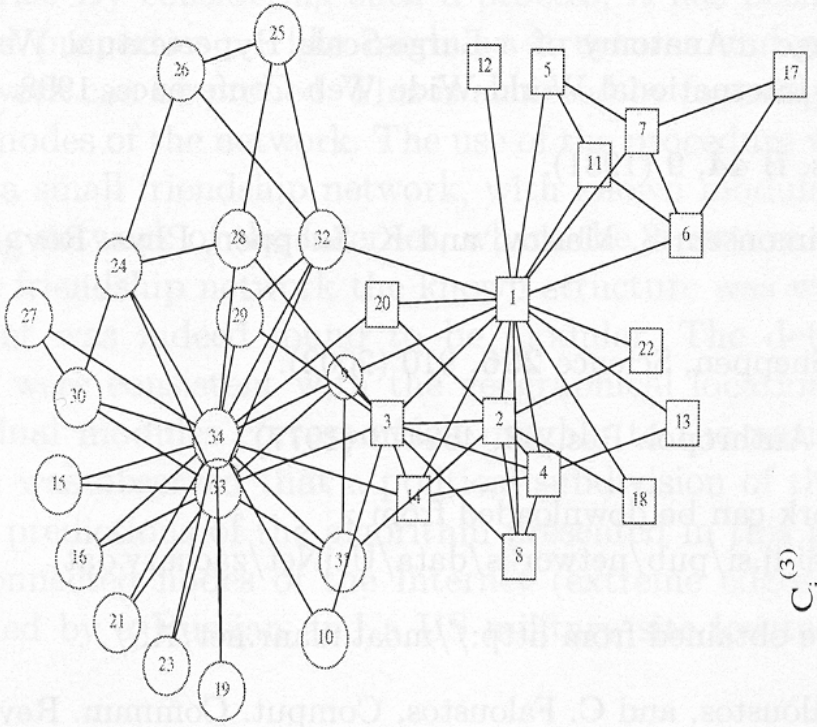
$_i$ is the expectation value of the number of random walkers on node i.

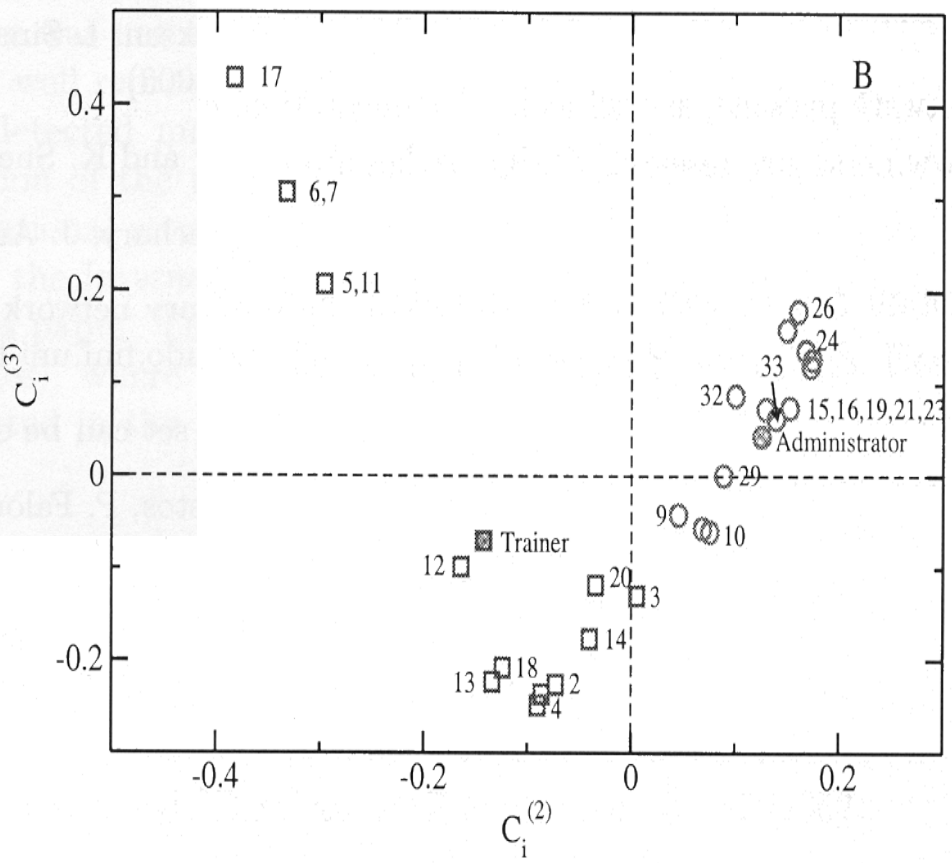
T_{ij} is the transfer matrix.

The elements of Tij are $1/\mathrm{K_j}$ for connected nodes and zero otherwise.

The relaxation of any initial distribution $_{\mathrm{i}}(0)$ towards the steady state $_{\mathrm{i}}(\infty)$ is governed by the spectral properties of Tij.

How does it look like?





1. Nodes of different type: **Assortative Mixing**

Let's assume the network consists of nodes of different type – are nodes preferably connected to nodes of their own type, e.g. are the communities representations of the type of the nodes?

Defining the assortativity coefficient r:

		women				
		black	hispanic	white	other	a_i
men	black	0.258	0.016	0.035	0.013	0.323
	hispanic	0.012	0.157	0.058	0.019	0.247
	white	0.013	0.023	0.306	0.035	0.377
	other	0.005	0.007	0.024	0.016	0.053
	b_i	0.289	0.204	0.423	0.084	

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

dissortive r < 0 < r assortative

But what to do, if we know nothing about the types of nodes?!

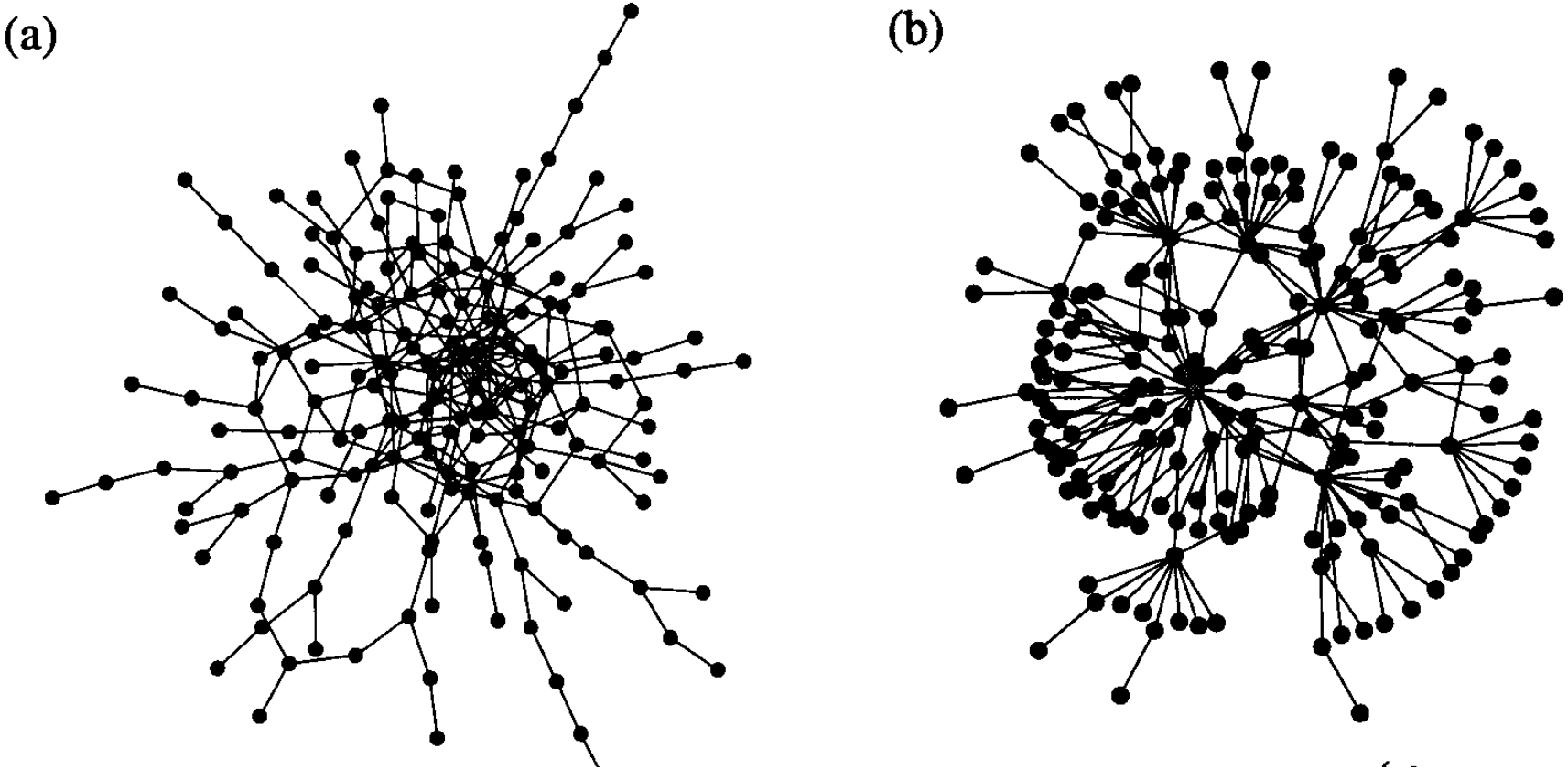
Well – ask for assortative mixing by degree! Do highly connected nodes primarily connect to other highly connected nodes?

	network	type	size n	assortativity r
social	physics coauthorship	undirected	52 909	0.363
	biology coauthorship	undirected	1 520 251	0.127
	mathematics coauthorship	undirected	253 339	0.120
	film actor collaborations	undirected	449 913	0.208
	company directors	undirected	7 673	0.276
	email address books	directed	16 881	0.092
technol.	Internet	undirected	10 697	−0.189
	World-Wide Web	directed	269 504	−0.067
	software dependencies	directed	3 162	−0.016
biological	protein interactions	undirected	2 115	−0.156
	metabolic network	undirected	765	−0.240
	neural network	directed	307	−0.226
	marine food web	directed	134	−0.263
	freshwater food web	directed	92	−0.326

Assortative

Dissortive

How does assortive mixing by degree look like?



assortative

dissortive

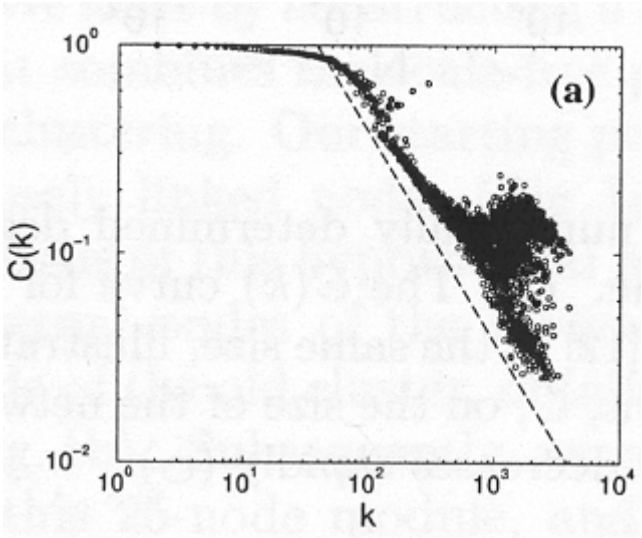
2. What else can we expect?

So if there exist communities of highly interconnected nodes (high clustering coeff. c) which are interconnected by high degree nodes (dissortive mixing by degree), then we can expect a dependence of c on k.

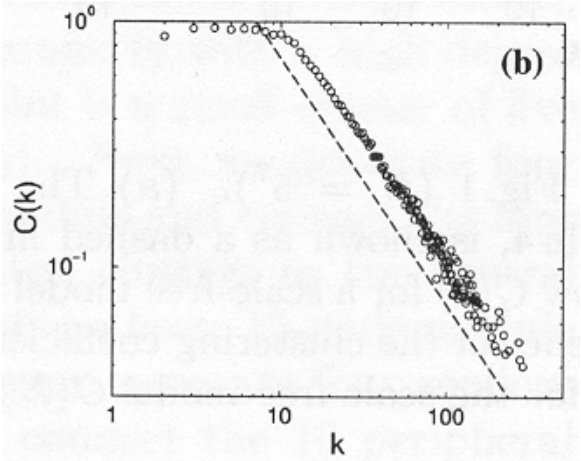
Indeed, some networks show:

$$c(k) \propto \frac{1}{k}$$

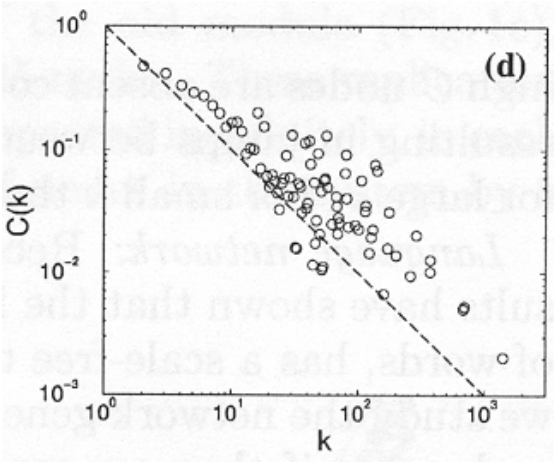
Some examples:



Actor Network



Synonyms in Merriam Webster



Internet AS level