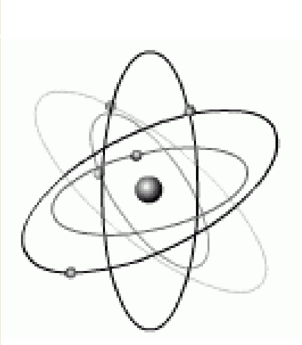
Looking for communities in networks



Krzysztof Kułakowski

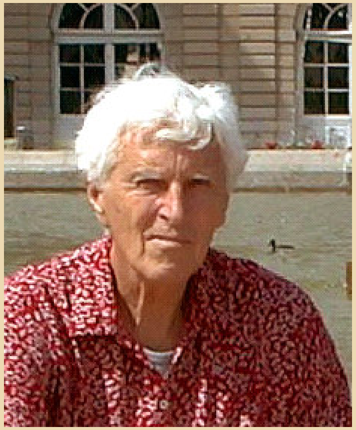
in cooperation with

Małgorzata Krawczyk
Przemysław Gawroński
Piotr Gronek

Faculty of Physics and Applied Computer Science,
AGH University of Sci and Tech, Kraków, Poland

Overview

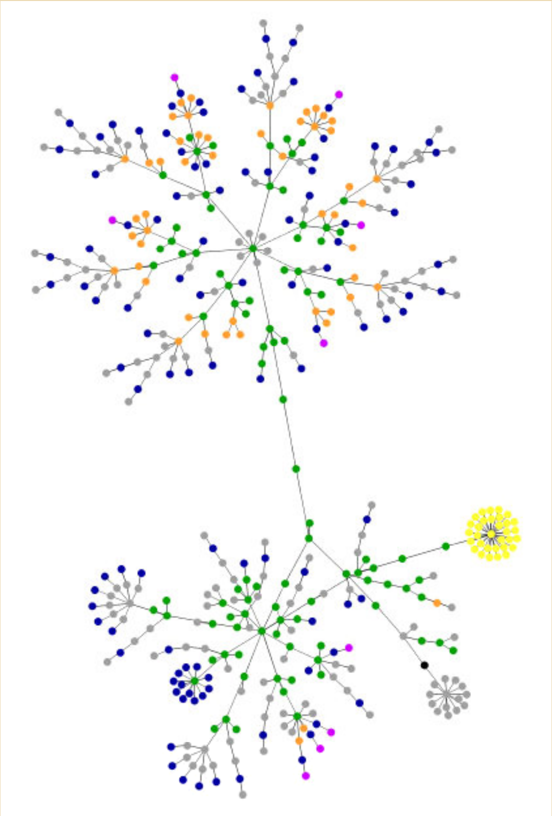
- The task
- A case study – the Southern Women
- The modularity after Newman
- A social problem – the Heider balance
- A continuous approach



Review:
S. Boccaletti et al.,
Complex networks: Structure and dynamics
Physics Reports 424 (2006) 175, section 7.1

The task

- Graph representation of a network



Jasmine's blog

N nodes

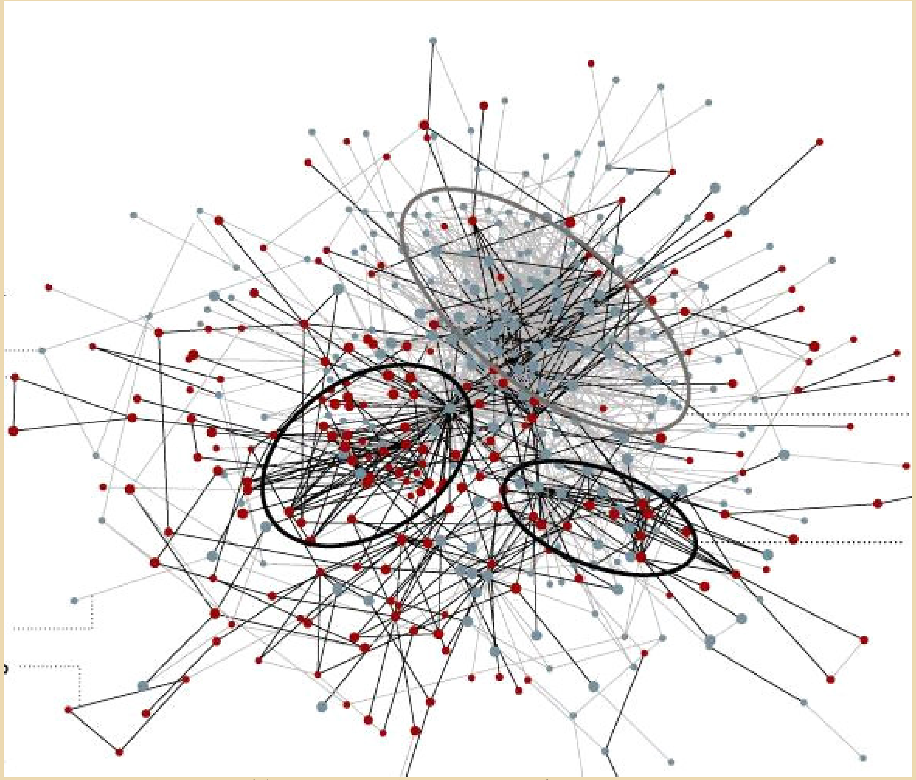
A(i,k)– connectivity matrix

A(i,k)= 0,1 or

A(i,k) ∈ $\mathcal{R}$ (weighted networks)

The task

- The communities:
  - subsets of nodes, internally connected more than with other parts of the network



Cornell > D.Constantine, NYT

Our aim is to find separated communities.

- The task
- A case study – the Southern Women
- The modularity after Newman
- A social problem – the Heider balance
- A continuous approach

The Southern Women, Natchez 1935

L. C. Freeman, *Finding social groups: a meta-analysis of the southern women data*, in Ronald Breiger, Kathleen Carley and Philippa Pattison (eds.) *Dynamic Social Network Modeling and Analysis.* Washington, D.C.:The National Academies Press, 2003.

Participation of the Southern Women in Events

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×		×	×					
2. Miss Laura Mandeville	×	×	×		×	×	×	×						
3. Miss Theresa Anderson		×	×	×	×	×	×	×	×					
4. Miss Brenda Rogers	×		×	×	×	×	×	×						
5. Miss Charlotte McDowd			×	×	×		×							
6. Miss Frances Anderson			×		×	×		×						
7. Miss Eleanor Nye					×	×	×	×						
8. Miss Pearl Oglethorpe						×		×	×					
9. Miss Ruth DeSand					×		×	×	×					
10. Miss Verne Sanderson							×	×	×			×		
11. Miss Myra Liddell								×	×	×		×		
12. Miss Katherine Rogers								×	×	×		×	×	×
13. Mrs. Sylvia Avondale							×	×	×	×		×	×	×
14. Mrs. Nora Fayette						×	×		×	×	×	×	×	×
15. Mrs. Helen Lloyd							×	×		×	×	×		
16. Mrs. Dorothy Murchison								×	×					
17. Mrs. Olivia Carleton									×		×			
18. Mrs. Flora Price									×		×			

Results of 21 methods

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	DGG41	W	W	W	W	W	W	W	W	WW	W	W	W	W	W	W	W	W	W
2	HOM50	W	W	W	W	W	W	W	WW			W	W	W	W	W		W	W
3	P&C72	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
4	BGR74	W	W	W	W	W	W	W		W	W	W	W	W	WW	WW		W	W
5	BBA75	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
6	BCH78	W	W	W	W	W	W				W	W	W	W	W	W			
7	DOR79	W	W	W	W	W	W	W		W	W	W	W	W	W	W			
8	BCH91	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
9	FRE92	W	W	W	W	W	W	W		W	W	W	W	W	W	W	W		
10	E&B93	W	W	W	W	W	W	W		W	W	W	W	W	W	W			
11	FR193	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
12	FR293	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
13	FW193	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	WW	W	W
14	FW293	W	W	W	W	W	W	W		W	W	W	W	W	W	W		W	W
15	BE197	W	W	W	W	W	W	W		W	W	W	W	W	W	W			
16	BE297	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
17	BE397	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
18	S&F99	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W		W	W
19	ROB00	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
20	OSB00	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
21	NEW01	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W

Comparison of 21 methods

	Code	Analysis	Closeness to the Matching Criterion
1	DGG41	Davis, Gardner and Gardner, Ethnography	0.920
2	HOM50	Homans, Intuition	0.854
3	P&C72	Phillips and Conviser, Information Theory	0.968
4	BGR74	Breiger, Algebra	0.933
5	BBA75	Breiger, Boorman and Arabie, CONCOR	0.927
6	BCH78	Bonacich, Boolean Algebra	0.841
7	DOR79	Doreian, Algebraic Topology	0.923
8	BCH91	Bonacich, Correspondence Analysis	0.968
9	FRE92	Freeman, G-Transitivity	0.926
10	E&B93	Everett and Borgatti, Regular Coloring	0.916
11	FR193	Freeman, Genetic Algorithm 1	0.968
12	FR293	Freeman, Genetic Algorithm 2	0.842
13	FW193	Freeman and White, Galois Lattice	0.917
14	FW293	Freeman and White, Galois Sub-Lattice	0.954
15	BE197	Borgatti and Everett, Bi-Clique	0.916
16	BE297	Borgatti and Everett, Taboo Search	0.968
17	BE397	Borgatti and Everett, Genetic Algorithm	0.968
18	S&F99	Skvoretz and Faust, p* Model	0.957
19	ROB00	Roberts, SVD with Normalization	0.968
20	OSB00	Osbourn, VERI Algorithm	0.543
21	NEW01	Newman, Weighted Co-Attendance	0.932

- The task
- A case study – the Southern Women
- The modularity after Newman
- A social problem – the Heider balance
- A continuous approach

Basic tool - the modularity

$$Q = \frac{1}{w} \sum_{ij} \left(A(i,j) - \frac{k_i k_j}{w} \right) \delta(c_i, c_j)$$

where

$$w = \sum_{ij} A(i,j)$$

$$k_i = \sum_{j} A(i,j)$$

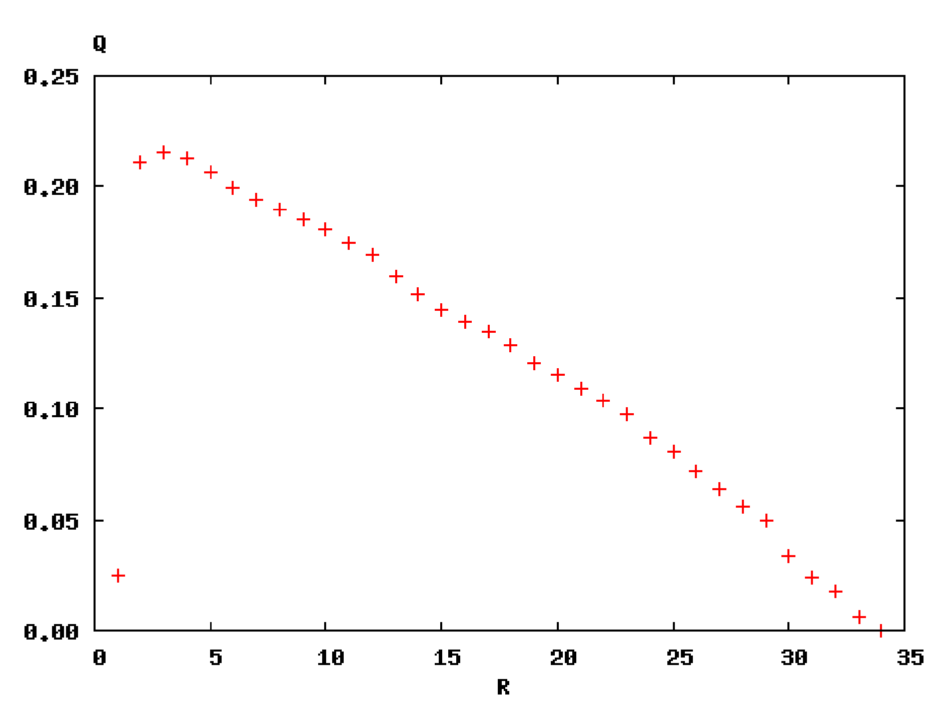
[M. E. J. Newman, *Analysis of weighted networks*, Phys. Rev. E 70, 056131 (2004)]

Special cases

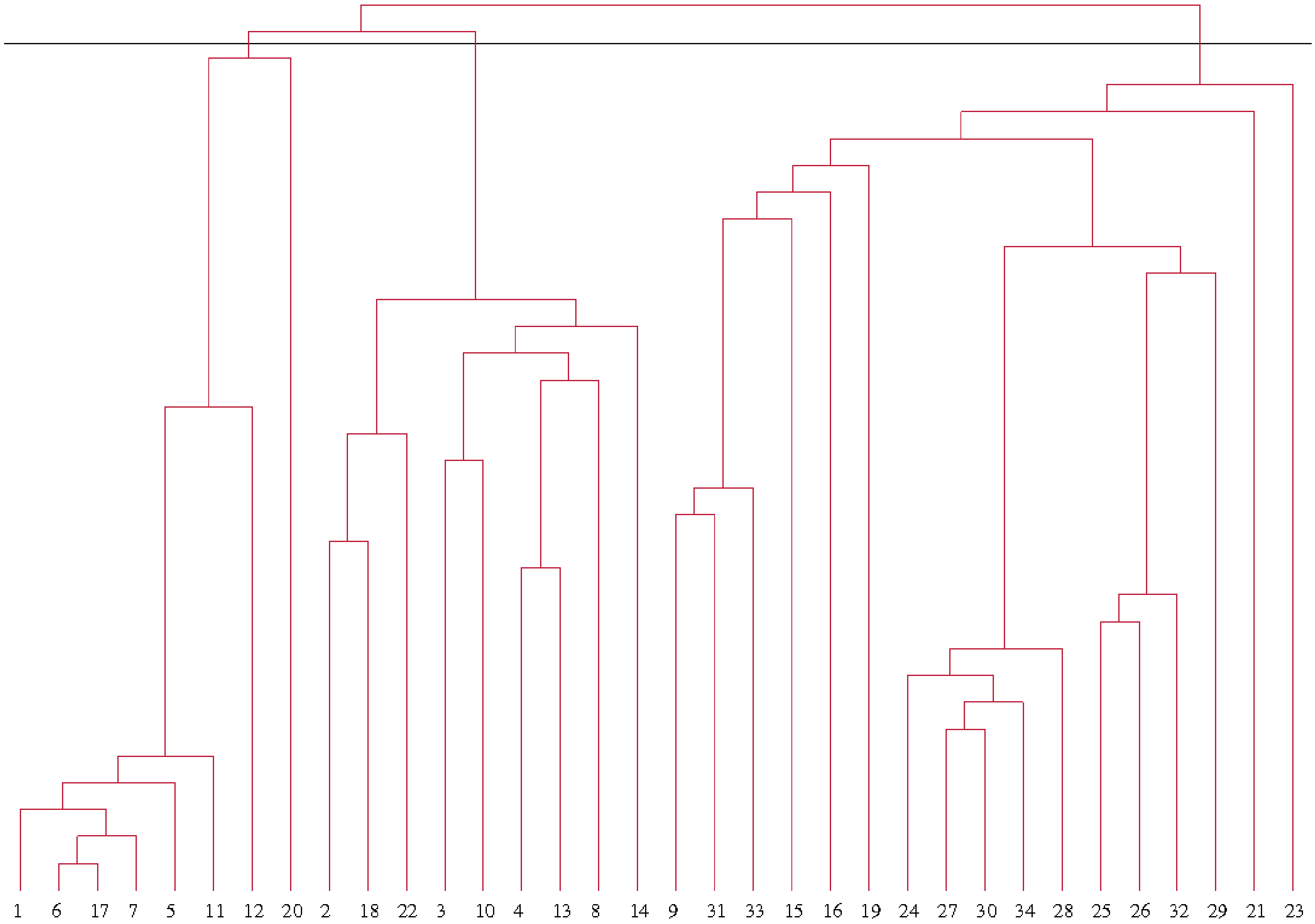
- One fully connected cluster of N nodes plus one separated node: $Q = 1/N$. When the node is added to the cluster, $Q=0$.
- R separated fully connected clusters, each of N nodes: $Q=1-(N-1)/(RN) \approx 1$ if $R>>1$. When one fully connected cluster is formed, $Q=1/(RN) \approx 0$.

Algorithm

- For a given set of data (N, $\{A(i,k)\}$) we have to construct a new connectivity matrix $\{c_{ik}=0,1\}$, which reveals the communities.
- start from $c_{ik}=0$ for all i,k.
- find the pair (n,m) such that when we put $c_{nm}=1$, Q increases most.
- repeat the search of such pairs, until Q starts to decrease. We accept the division where Q has a maximum.



Results – a dendrogram



- The task
- A case study – the Southern Women
- The modularity after Newman
- A social problem – the Heider balance
- A continuous approach

The Heider balance

- Let us transform the connectivity matrix as:

  $1 \rightarrow \ 1$ (friendly relation)

  $0 \rightarrow -1$ (hostile relation)

- Nodes = group members.
- In the balanced state, some rules are obeyed:

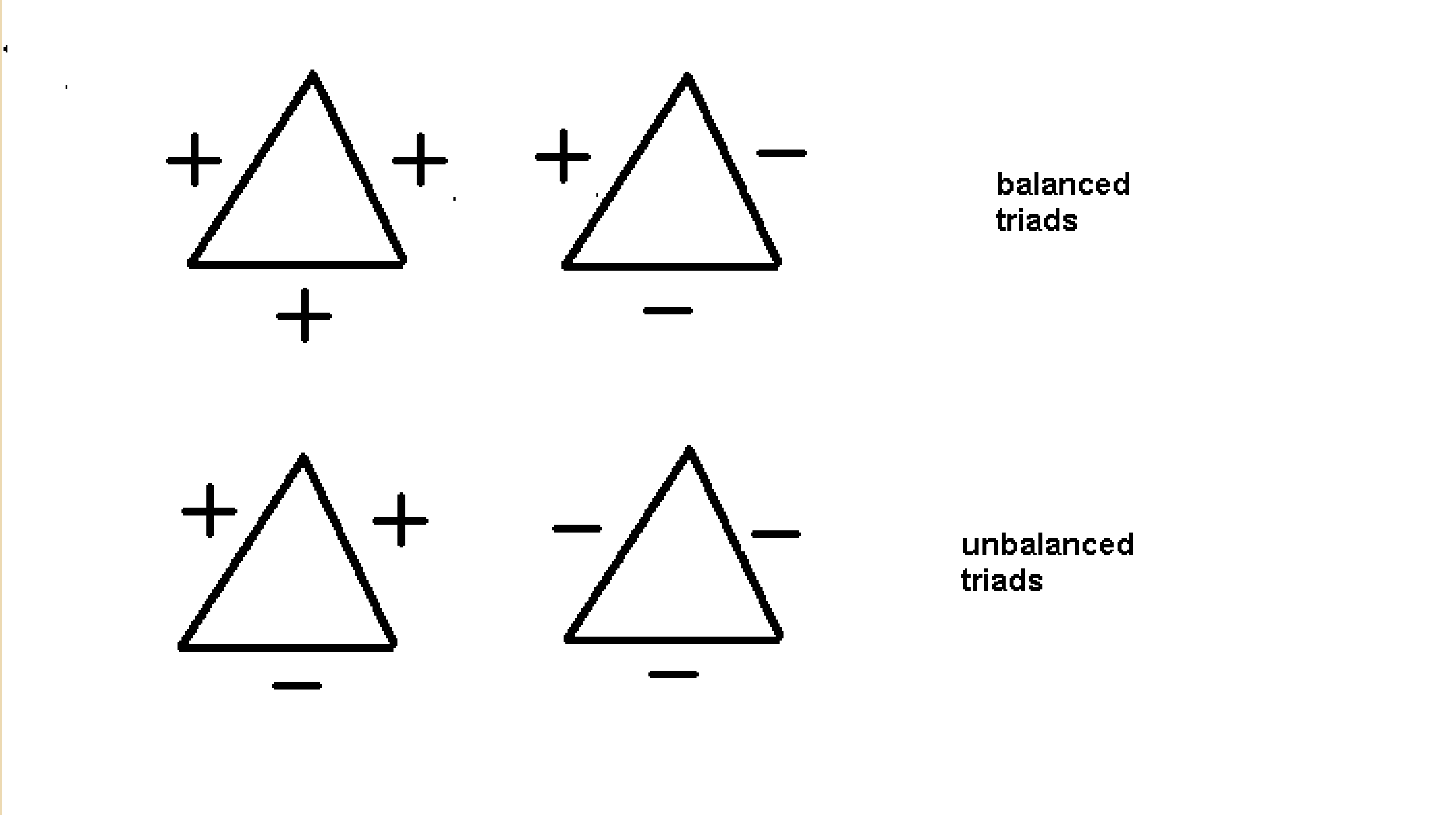
  Friend of my friend is my friend $A(i,k)>0 \ , \ A(k,n)>0 \ \Rightarrow A(i,n)>0$

  Enemy of my friend is my enemy $A(i,k)<0 \ , \ A(k,n)>0 \ \Rightarrow A(i,n)<0$

  Friend of my enemy is my enemy $A(i,k)>0 \ , \ A(k,n)<0 \ \Rightarrow A(i,n)<0$

  Enemy of my enemy is my friend $A(i,k)<0 \ , \ A(k,n)<0 \ \Rightarrow A(i,n)>0$

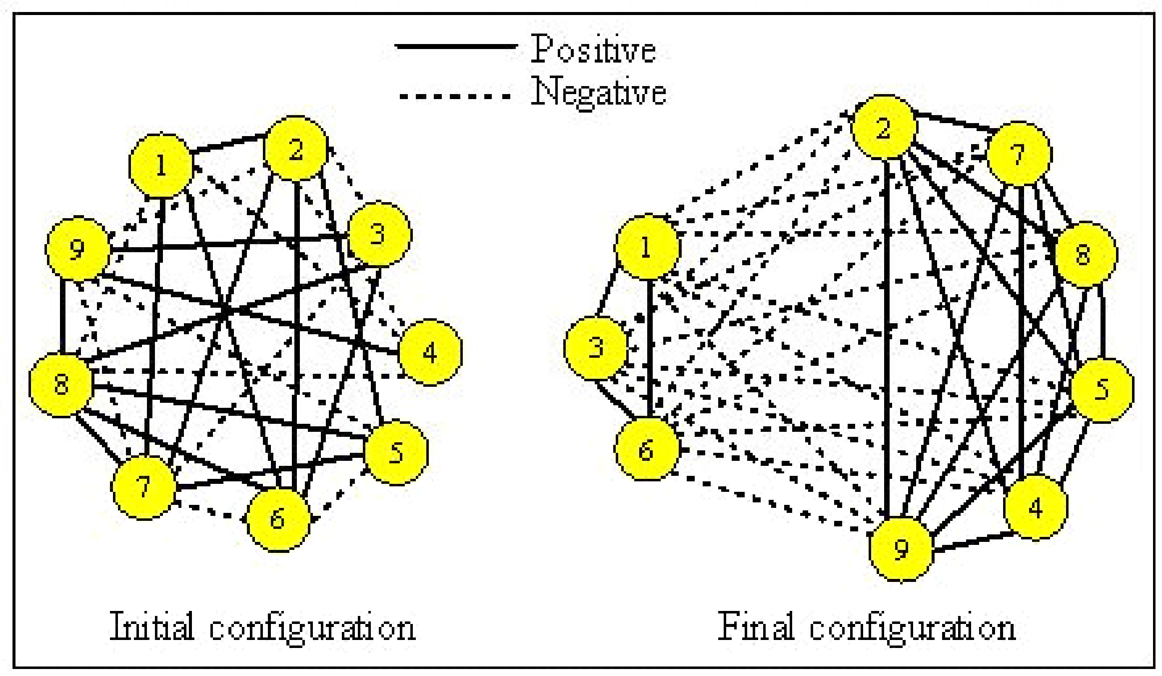
- If the state is not balanced, some members suffer the cognitive dissonance.



•They try to get the balanced state, changing some of their relations.

The Harary theorem

- At the balanced state, the group is divided into two mutually hostile subgroups. The relations within both subgroups are friendly.



[Z. Wang, W. Thorngate, *Sentiment and social mitosis: Implications of Heider's Balance Theory*, J. of Artificial Society and Social Simulation vol. 6, no. 3 (2003)]

Problems with the dynamics

- Sociological:

  Relations are more subtle than just friendly or hostile; more possibilities should be included.

- Computational:

  The final balanced state can depend on the order of modified relations.

A continuous dynamics

- The relations are described by real numbers $A(i,k) \in (-R,R)$
- Their time evolution is determined by differential equations

$$\frac{dA(i,k)}{dt} = \left[1 - \frac{A(i,k)^2}{R^2}\right] \sum_n A(i,n)A(n,k)$$

The results agree with the best six methods, reported by Freeman.

[More details in KK, *Some recent attempts to simulate the Heider balance problem*, Computing in Science and Engineering, July/August 2007, p.80.]

- The task
- A case study – the Southern Women
- The modularity after Newman
- A social problem – the Heider balance
- The continuous approach

Looking for communities in general– - the continuous approach

- In the Heider balance problem, the differential equations were motivated by the specific nature of the cognitive dissonance.
- In a general approach, for each pair we ask: do other connections confirm that this pair belongs to the same set?

The continuous approach

- The differential equations are

$$\frac{dA(i,k)}{dt} = \Theta(A(i,k))\Theta(1-A(i,k))\sum_{n}[A(i,n)A(n,k)-b]$$

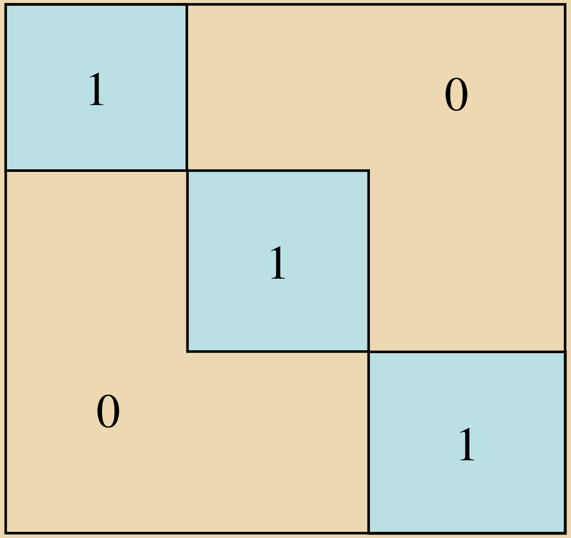
The product of the step functions ensures that once $A(i,k)$=0 or 1, it does not change any more.

Once the product $A(i,n)A(n,k)>b$, the connection of sites *i,k* via site *n* proves that *i,n* belong to the same cluster and so do *n,k*. Then *A(i.k)* should increase.

(more details in M. Krawczyk and K. K., arXiv:0709.0923)

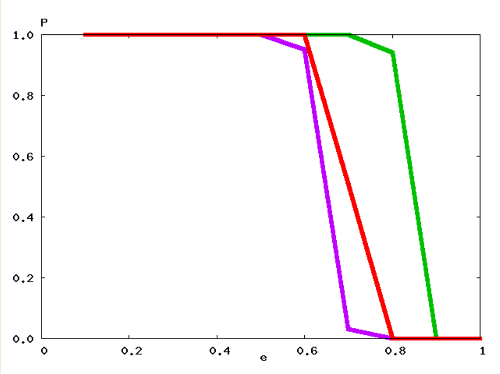
The continuous approach – a test

- For separated and fully connected clusters, the connectivity matrix has a block form

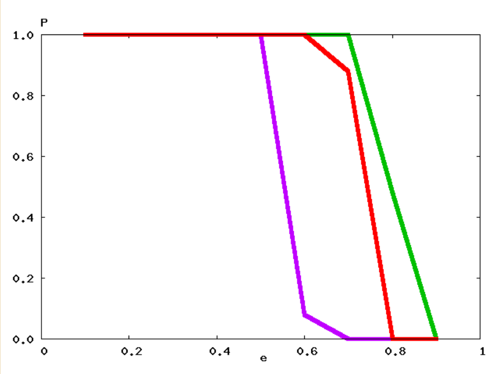


We add a noise: random numbers with the uniform distribution (0,*e*) are added to zeros and subtracted from units. The task is to reproduce the original structure.

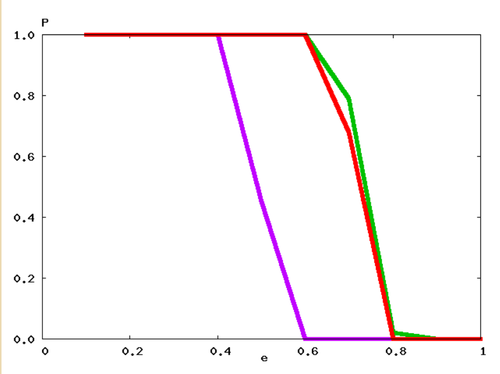
The continuous approach – a test



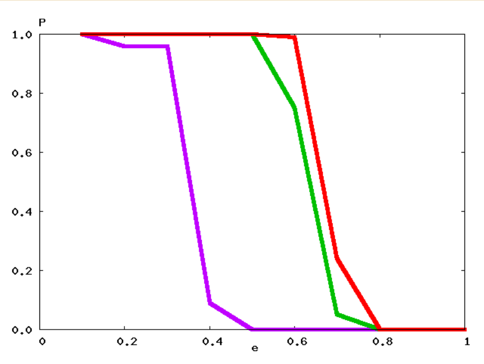
R=2



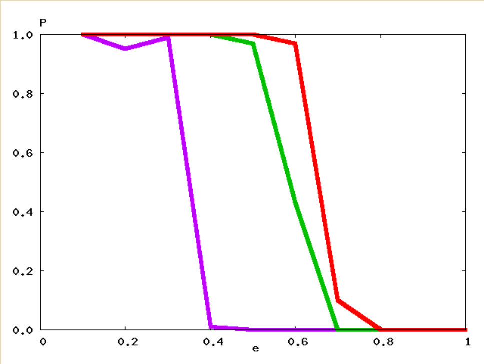
R=3



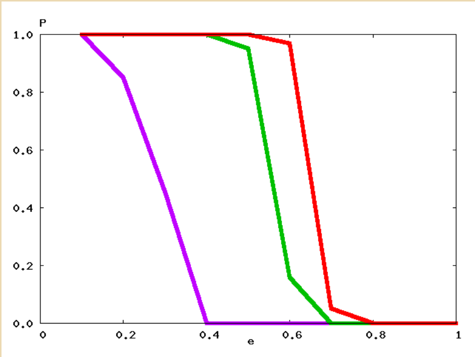
R=4



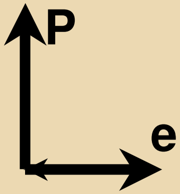
R=6



R=7



R=8



P – probability that the initial structure is reproduced
e – upper limit of noise

NR=60÷72, #=100, b=1/4, — modularity — random walk — continuous

Final remarks

- The method is deterministic. It produces some results even for a purely random graph.

- Direct comparisons with social experiments (Southern women, Zachary karate club) can be dubious.

Thank you

Control questions

- To calculate the modularity, one takes into account:
  A. Only existing connections
  B. Only a proposed division into clusters
  C. Both A and B

- Differential equations raise difficulties because
  A. high numerical precision is needed
  B. the order of modified relations does matter