Structures, Processes, and Clustering of Complex Networks

Andreas Krüger

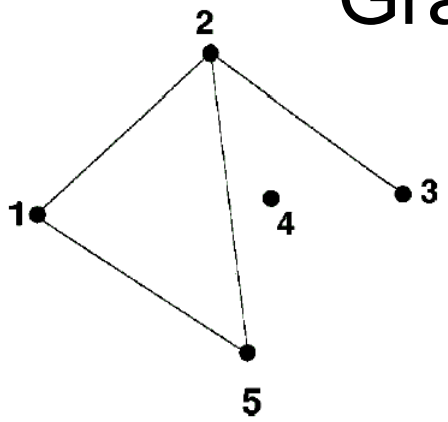
6.2.2008

Dissertation an der Fakultät für
Physik der Uni Bielefeld vom 31.10.2007
→ http://bieson.ub.uni-bielefeld.de/volltexte/2008/1247/ ←

PART 1 of 4:
Quick introduction into

Graphs and the Physics „hype“ about Complex Networks

Graph G=(V,E)



- KNOTEN, node, vertex, actor, point, individual ...
- KANTE, edge, line, bond, tie, connection...

V $\subseteq Z^+$ elements represented by dots

e.g. **V**={1,2,3,4,5}

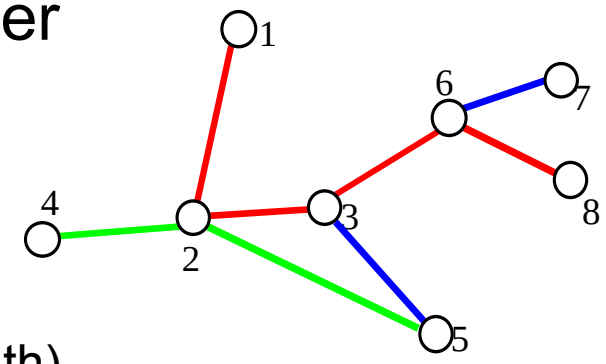
E $\subseteq \binom{V}{2}$, elements represented by lines

e.g. **E**={ (1,2) , (1,5) , (2,3) , (2,5) }

Neighbours:

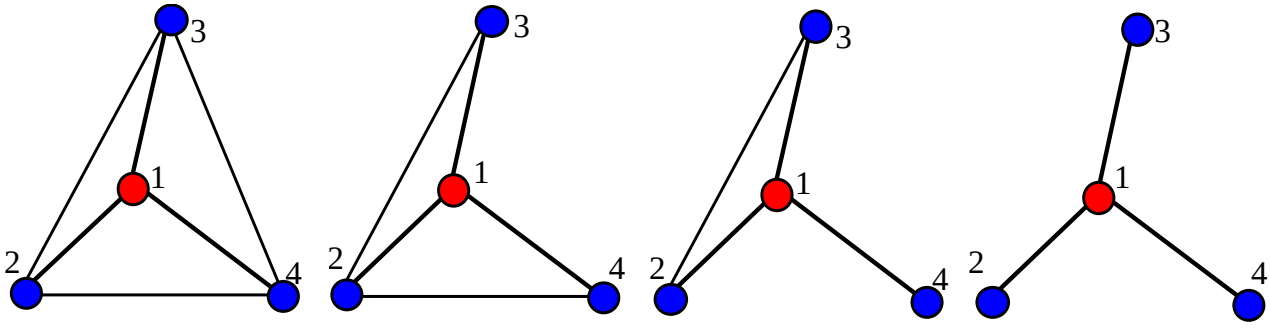
x~y := {$\exists$ **e**$\in$**E** with e=(x,y) or e=(y,x)}

Pfade, Durchmesser



- pathlength (geodesic path)
  - **Shortest** connection between 2 nodes
  - Example pathlength(1,8) = 4
- global graph-properties
  - *Diameter* = **longest** geodesic path (here 4)
  - *characteristic pathlength* = **average** of all paths (i,j)

Cluster-Coefficient, Dreiecks-Zahl



$$C_i = \frac{\#T_i}{k_i(k_i - 1)/2}$$

$$\underset{\text{(global)}}{C} = \frac{1}{N}\sum_i C_i$$

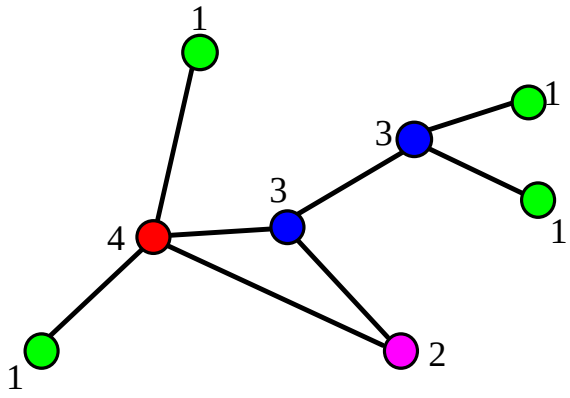
$\#T_i$ = Number of Triangles around vertex i

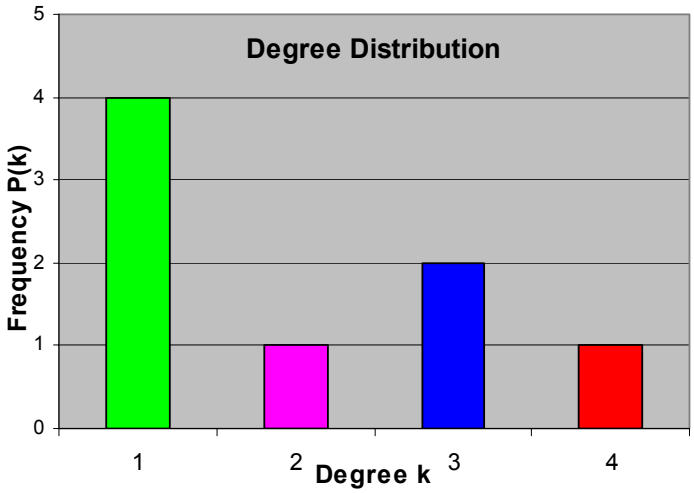
C_i: Estimator for **local density of connections**, "*how many of my friends are friends to each other?*"

degree of a node

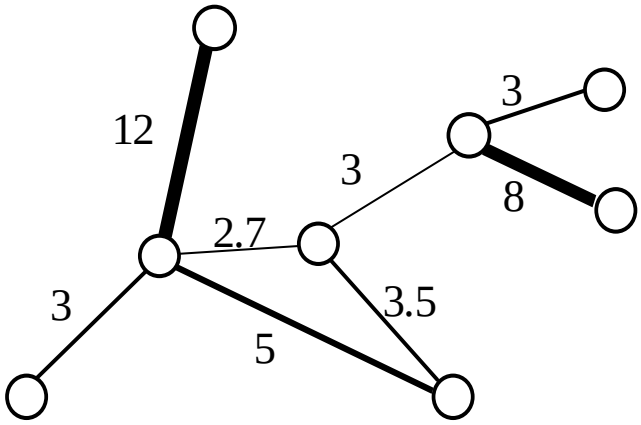
k_x = deg(x) = $|N_1(x)|$
= number of N_1-Neighbours of node x

P(k) = *Degree-Distribution (frequency)*
= **number of nodes** with deg=k



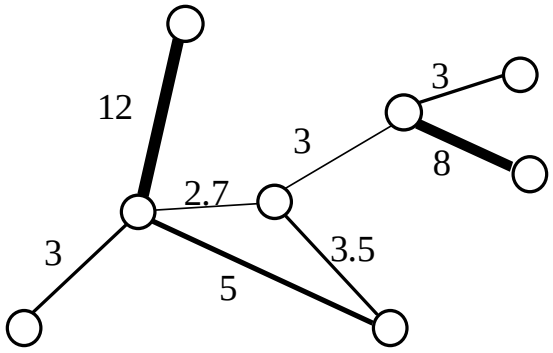


Gewichtete Kanten, gewichteter Graph



Funktion g(e) mit g: E →|R
Gibt jeder Kante eine Stärke, Intensität, etc.
(*ungewichtet*: g(e)=1 für alle e)

Adjacency Matrix



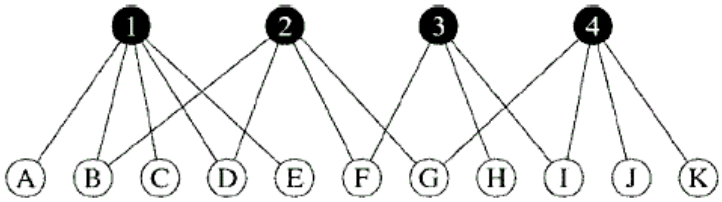
	1	2	3	4	5	6	7	8
1		3	0	0	0	0	0	0
2	3		12	2.7	5	0	0	0
3	0	12		0	0	0	0	0
4	0	2.7	0		3.5	3	0	0
5	0	5	0	3.5		0	0	0
6	0	0	0	3	0		3	8
7	0	0	0	0	0	3		0
8	0	0	0	0	0	8	0	

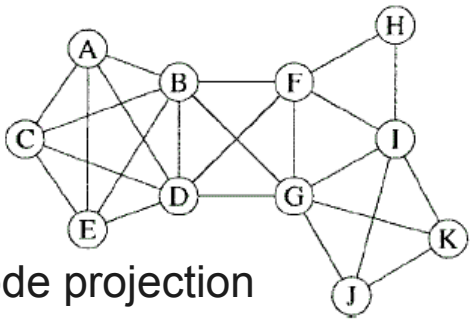
Bipartite Graphs

Up to now we have only seen so called *1-mode graphs*, i.e. there is **one** type of vertices

Now imagine for example 4 **films** (black) and 11 playing **Actors** (white).

From the 2-mode graph we can generate a 1-mode graph by projection (under information loss)





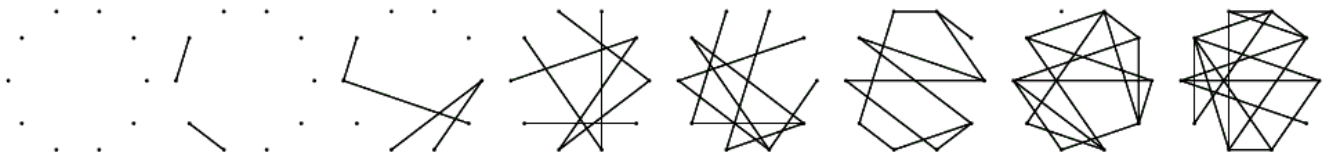
1-mode projection

FIG. 14. A schematic representation of a bipartite graph, such as the graph of movies and the actors who have appeared in them. In this small graph we have four movies, labeled 1 to 4, and eleven actors, labeled A to K, with edges joining each movie to the actors in its cast. The bottom figure shows the one-mode projection of the graph for the eleven actors. After Newman, Strogatz, and Watts (2001).

THE *random* graph model: Erdös Renyi RandomGraph (~19**60**)

http://www.nd.edu/~networks/linked/newfile6.htm

- G(N,p) random graph
- N vertices → # possible edges: $M_{max} = \frac{N(N-1)}{2}$
- p independent probability for each edge (Bernoulli-process) $p \in [0,1]$



p=0	p=0.05	p=0.1	p=0.15	p=0.2	p=0.25	p=0.3	p=0.35
B=0	B=2	B=4	B=7	B=9	B=11	B=14	B=16
NOC=10	NOC=8	NOC=6	NOC=3	NOC=1	NOC=1	NOC=2	NOC=1

Degree Distribution of ER G(N,p) is ~ Poisson

The average is good estimator for the whole distribution (bellshaped)

$$\langle k \rangle = (N-1)p$$

$$= (N-1)\frac{M}{N(N-1)/2} = \frac{2M}{N}$$

$$= \mu$$

The degree has a binomial distribution. For N>>1 it becomes Poissonian:

$$P(k) = e^{-\mu}\frac{\mu^k}{k!}$$

with an exponential tail for large k

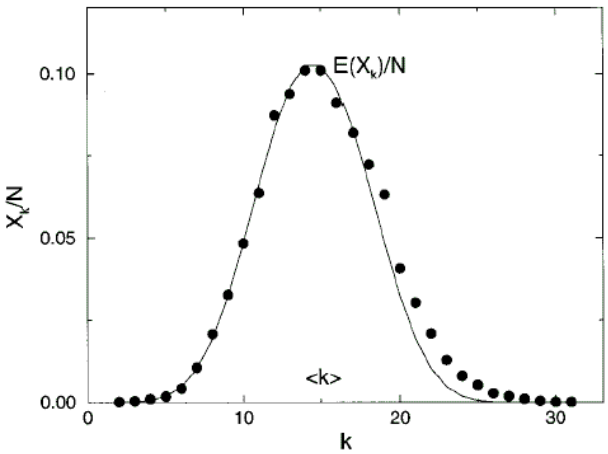


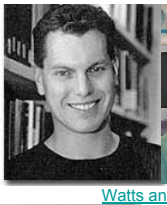
FIG. 7. The degree distribution that results from the numerical simulation of a random graph. We generated a single random graph with $N=10\,000$ nodes and connection probability $p=0.0015$, and calculated the number of nodes with degree k, X_k. The plot compares X_k/N with the expectation value of the Poisson distribution (13), $E(X_k)/N=P(k_i=k)$, and we can see that the deviation is small.

paradigm shift 1998/99
static random → grown networks

Starting Papers

- **Watts**, D. J. and **Strogatz** S. H.,
  Collective dynamics of small-world networks,
  1998.06.04 Nature, 393, 440.
- **Barabasi**, A.-L. and **Albert**, R.,
  Emergence of scaling in random networks,
  1999, Science 286, 509–512 .
- Albert, R., **Jeong**, H. and Barabasi, A.-L.,
  The diameter of the world-wide web,
  1999, Nature (London) 401, 130-131; cond-mat/9907038.
- Barabasi, A.-L., Albert, R., and Jeong, H.,
  Mean-field theory for scale-free random networks,
  1999, Physica A 272, 173–187.

- Barabasi, A.-L.,
  Linked: The New Science of Networks,
  Perseus, Cambridge, MA (2002).



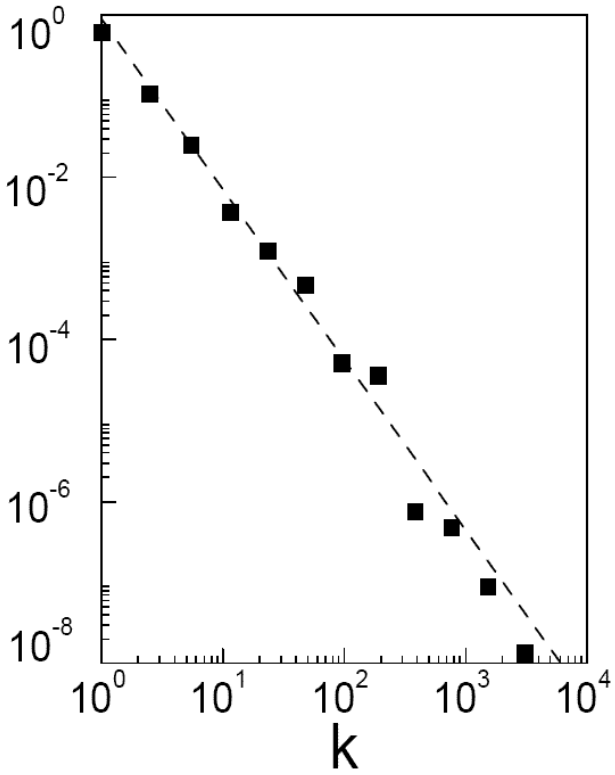
Watts and Strogatz


Reka Albert


Albert-László Barabási


Hawoong Jeong

Empirical Property 1: scale free



WWW-Ausschnitt N=325729
kmean=5.46 gamma=2.1 (condmat9910332)

In MEASURED networks, the degree distribution is **not** Poissonian (with exponential tail) for large k

but "fat tail"
→ falling **power-law**

$$P(k) = \frac{1}{k^{\gamma}}$$

$$\gamma \sim 2.5$$

An average <k> doesn't really make sense here
= no *built-in scale*

→ „scale-free"

Property 1: scale free –
Some objects that seem to have a scale-free degree

- WWW
- Internet-Routing
- Protein-Protein-docking
- citations
- collaborations
  - publications
  - Movie-Actor-Network
- Human Sexuality Networks
- Telefone calls
- brains
  - Caenorhabditis elegans
  - Humans
- computer code
- The Word Web of language

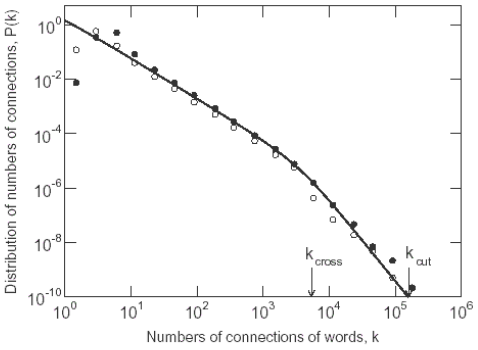


FIG. 9. The distribution of the numbers of connections (degrees) of words in the word web in a log-log scale [126]. Empty and filled circles show the distributions of the number of connections obtained in Ref. [126] for two different methods of the construction of the Word Web. The solid line is the result of theory of Ref. [127] (see Sec. IX.J where the parameters of the Word Web, namely, the size $t \approx 470\,000$ and the average number of connections of a node, $\overline{k}(t) \approx 72$, were used. The arrows indicate the theoretically obtained point of crossover, k_{cross} between the regions with different power laws, and the cutoff k_{cut} due to the size effect. For a better comparison, the theoretical curve is displaced upward to exclude two experimental points with the smallest k (note that the comparison is impossible in the region of the smallest k where the empirical distribution essentially depends on the definition of the Word Web).

scale free: Sicher gegen zufälligen Ausfall, aber empfindlich gegen hub-Angriffe

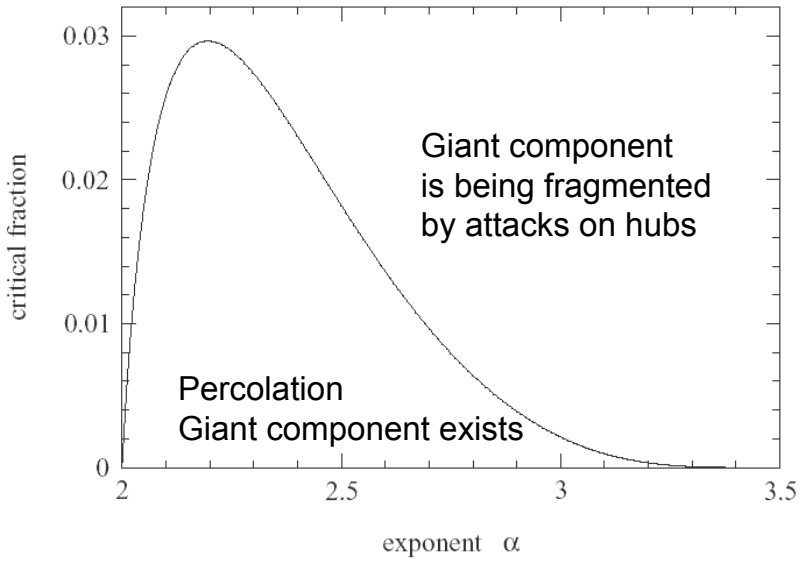


FIG. 14 The fraction of vertices that must be removed from a network to destroy the giant component, if the network has the form of a configuration model with a power-law degree distribution of exponent α, and vertices are removed in decreasing order of their degrees.

From Eqs. (151) and (152), one can easily obtain $k_{cut}(\gamma)$ and $f_c(\gamma)$ (see Fig. 41 [205]). Note that $f_c > 0$ only in the range $2 < \gamma < 3.479\ldots$. When $\gamma < 2$, a finite number of vertices keep a finite fraction of all connections, so their removal should have a striking effect on the network. For $\gamma > 3.479\ldots$, the giant connected component is absent even before the attack. Indeed, in the undamaged network, the giant connected component exists if $\sum_{k=1}^{\infty}(k^2 - 2k)k^{-\gamma} = \zeta(\gamma - 2) - 2\zeta(\gamma - 1) > 0$ [see Eq. (116)]. This corresponds to $\gamma < 3.479\ldots$.

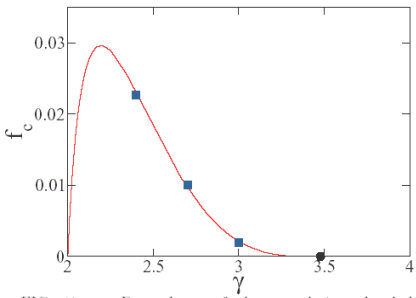


FIG. 41. Dependence of the percolation threshold $f_c = 1 - p_c$ on the value of the γ exponent of the large scale-free network for intentional damage (attack) [205]. Such a dependence was originally obtained in the framework of a continuum approach [63]. Here we present the exact curve. f is the fraction of removed vertices with the largest numbers of connections. The degree distribution of the network before the attack is $P(k) \propto k^{-\gamma}$ for $k \geq 1$, $P(0) = 0$. The circle indicates the point $\gamma = 3.479\ldots$ above which $f_c = 0$. The squares represent the results of calculations and simulation in Ref. [62].

First Model
Albert Barabasi: ***Preferential-Attachment***

1. Growth ! Not in static systems...
   - Per time step **1 new node** and **m new edges**
2. Preferential Attachment
   - **y~x**: new node y, an old node x, but which one?
   - Probability to choose x linearly proportional to current degree of x:

P(y~x with deg(x,t)=$k_{i,t}$) ~ $k_{i,t}$

„the rich get richer "

YES → scale-free, exponent gamma=3
YES → small world property
NO → high clustering

Three Properties of empirical networks

- Scale free degree distribution (→ hubs)

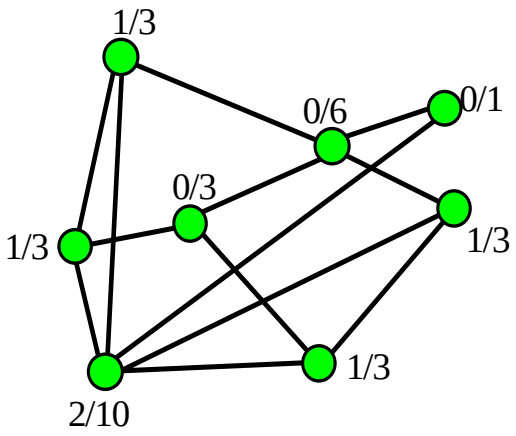
$$P(k) \sim k^{-\gamma}$$

- High clustering = many more triangles than in G(n,p)
  = friends of my friends are friends to each other

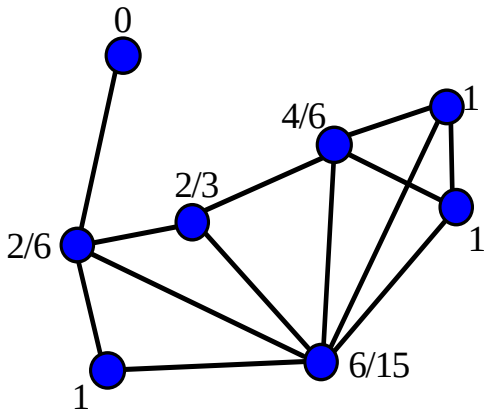
- Small-World:
  - Diameter and average pathlength ~ log N
  - (e.g. *lattice* diameter ~ $N^{1/dim}$ = no small world)

high clustering

$$C_i = \frac{\#T_i}{k_i(k_i-1)/2}$$



C=0.1917



C=0.6333

In both cases M=13 and N=8, but in the *right* picture many more friends are themselves direct friends to each other

! "Empirical Networks" have a significantly **higher clustering-coefficient** than ErdosRenyi-RandomGraphs !

small world

http://www.nd.edu/~networks/linked/newfile8.htm

1998: Watts-Strogatz random rewiring

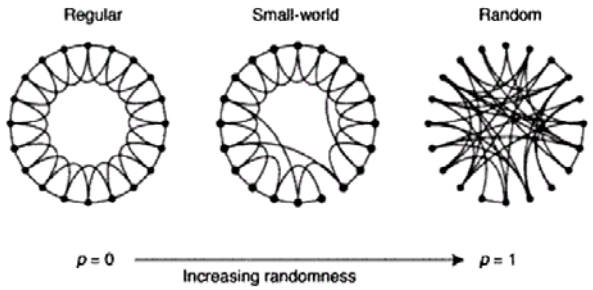


FIG. 15. The random rewiring procedure of the Watts-Strogatz model, which interpolates between a regular ring lattice and a random network without altering the number of nodes or edges) nodes, each connected to its four neare) the original ring is unchanged; as p inc comes increasingly disordered until for p ed randomly. After Watts and Strogatz,

1967: Milgram "6 degrees of separation"

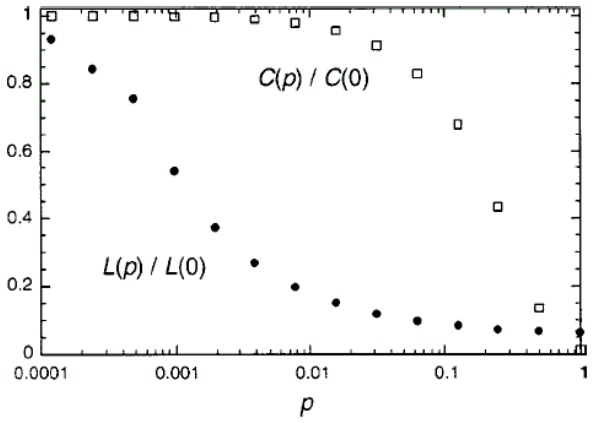


FIG. 16. Characteristic path length $\ell(p)$ and clustering coefficient $C(p)$ for the Watts-Strogatz model. The data are normalized by the values $\ell(0)$ and $C(0)$ for a regular lattice. A logarithmic horizontal scale resolves the rapid drop in $\ell(p)$, corresponding to the onset of the small-world phenomenon. During this drop $C(p)$ remains almost constant, indicating that the transition to a small world is almost undetectable at the local level. After Watts and Strogatz, 1998.

L ~ log N

Paper 1

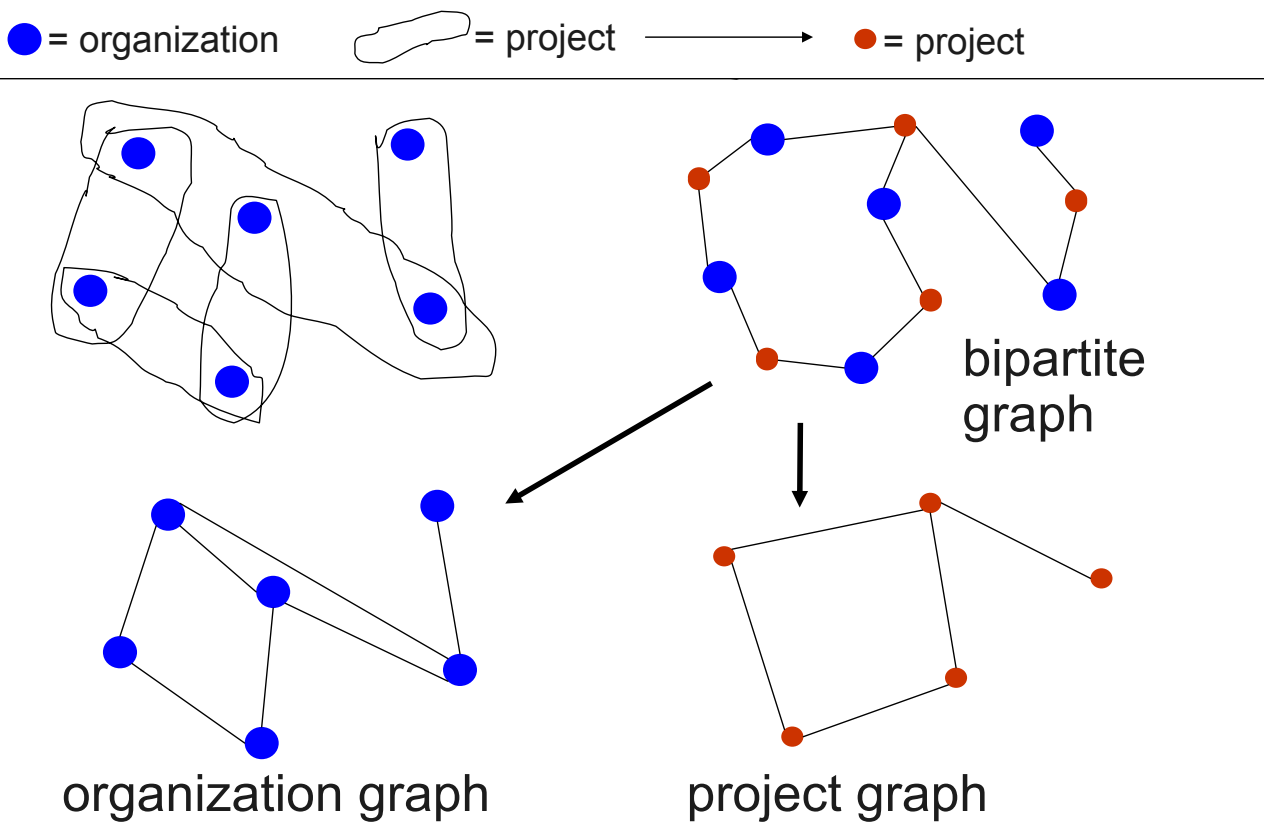
The Network of EU-funded Collaborative R&D Projects

Phys Rev E 73, 036132 (2006)

Projekt-Netzwerke

- Europäische Kooperation zur Datenanalyse mithilfe von Netzwerk-Methoden
  - T.Krüger (B), Th.Roediger-Schluga (W), A. Krüger (BI), M.Barber (PT), und Ph.Blanchard (BI)
- Datensatz beim ARC systems research GmbH, Seibersdorf bei Wien
  - ca. 40,000 EU-geförderte Forschungsprojekte 1984-2004
  - Rahmenprogramm der EU dauern 4 Jahre; mittlerweile 2stellige Milliardenbeträge für Forschungs- und Entwicklungsförderung, Aber "nur" ca. 6% aller Forschungsprojekte (Rest zB nationale Förderung)
  - Quelle: CORDIS-Datenbank der EU, Download, Standardisierung
- Was wissen wir über ein Projekt
  - Beteiligte Institutionen
  - Titel, Themen
  - Beginn, Dauer
  - Verantwortliche, Anschrift, Department/Faculty, …
- **Grundidee:** DB → Netzwerk Es ist ein bipartiter Graph: Projekte ←→ Organisationen
  Dann unimodale Projektion: Projekte werden im Graph verbunden, wenn nicht-leere Schnittmenge von Organisationen
  - Projektion auf Projektgraph
  - Projektion auf Organisations-Graph
  - Frühes Ergebnis: ***skalenfreie*** *Degree-Verteilung* auf Organisationsgraph !!!

Set Graphs



graph characteristic	FP1	FP2	FP3	FP4
# vertices: N	2500	6135	9615	20873
(N for larg. comp.)	(2038)	(5875)	(8920)	(20130)
N outside larg.comp.	462	260	695	743
# edges: M	9557	64300	113693	199965
(# edges M larg.comp.)	(9410)	(64162)	(113219)	(199182)
mean degree: $\bar{d}$	7.65	20.96	23.65	19.16
($\bar{d}$ larg.comp.)	(9.23)	(21.84)	(25.39)	(19.79)
maximal degree: $d_{\max}$	140	386	648	649
mean triangles per vertex: $\triangle$	22.90	169.70	244.91	146.04
($\triangle$ larg.comp.)	(27.97)	177.16	263.84	151.26
maximal triangle-number	966	5295	15128	10730
cluster coefficient: $\bar{C}$	0.57	0.72	0.72	0.79
($\bar{C}$ larg. comp.)	(0.67)	(0.74)	(0.75)	(0.81)
number of components	369	183	455	467
diameter of largest component	9	7	9	10
mean path length: λ of l.c.	3.70	3.27	3.32	3.59
exponent of degree distribution	-2.1	-2.0	-2.0	-2.1
variance of degree exponent	0.4	0.3	0.3	0.3
exponent of org-size distr.	-2.1	-1.9	-1.7	-1.8
variance of size exponent	0.5	0.3	0.5	0.3
mean # projects per org: $\mathbb{E}(\lvert O\rvert)$	2.40	4. 87	5.6	6.24
maximal size ($\max \lvert O\rvert$)	130	82	138	172

TABLE II: Basic network properties of FP1–4 organizations projection.

O-graph

Organisations Projection

graph characteristic	FP1	FP2	FP3	FP4
# vertices: N	3283	3884	5528	9087
(N for larg. comp.)	(2764)	(3662)	(5027)	(8566)
N outside larg.comp.	519	222	501	521
# edges: M	51217	94527	202358	348542
(# edges M larg.comp.)	(50940)	(94471)	(202306)	(348474)
mean degree: $\bar{d}$	31.20	48.68	73.20	76.71
($\bar{d}$ larg. comp.)	(36.86)	(51.60)	(80.49)	(81.36)
maximal degree: $d_{\max}$	282	387	917	771
mean triangles per vertex: $\triangle$	774.41	871.19	1970.30	2034.31
($\triangle$ larg.comp.)	919.53	923.98	2167.05	2158.03
maximal triangle-number	12903	11125	37247	41141
cluster coefficient: $\bar{C}$	0.67	0.54	0.44	0.47
($\bar{C}$ larg.comp.)	(0.75)	(0.57)	(0.48)	(0.50)
number of components	369	183	455	467
diameter of largest component	9	7	10	9
mean path length: λ of l.c.	3.24	2.80	2.72	2.80
exponent of degree distribution	(-0.8, -3.4)	(-0.7, -3.3)	(-0.6, -3.7)	(-0.3, -2.2)
variance of degree exponent	(0.4, 3.6)	(0.3, 1.7)	(0.3, 1.4)	(0.2, 0.6)
exponent of proj-size distr.	-3.59	-2.9	-3.2	-4.1
variance of size exponent	0.6	0.4	0.2	0.3
mean # orgs per project: $\mathbb{E}(\lvert P\rvert)$	3.15	3.08	3.22	2.71
maximal size ($\max \lvert P\rvert$)	20	44	73	54

TABLE III: Basic network properties of FP1–4 projects projection.

P-graph

Projects Projection

Projekt-Netzwerke

- Erste geteste Hypothesen zur Struktur-Enstehung:
  - Basismodell: ***RandomSets*** von rein zufällig zusammengewürfelten Mengen von Organisationen **sind Projekte** (Projektgrößen gegeben)
  - bipartites Configuration Modell (ähnlich MolloyReed 1995/1998): Projektgrößen UND Organisationsgrößen gegeben aus Empirie
- Nur einige Fragen, die uns leiten:
  - Rekonstruierbarkeit durch Simulation?
  - Wesentliche Eigenschaften des empirischen Netzwerkes
  - Nützliche neue Netzwerk-Observablen entwickeln
    - meso-Skala zwischen lokal und global -> Clustering in Communities
    - Dreieckszahl/Clusterkoeff. auf gewichteten Graphen -> noch ununtersucht
  - nach unterschiedlichen Rollen im Netzwerk untersuchen
    - Hubs, zentralste Knoten... (topologische Rollen)
    - stabile Akteurs-Konfigurationen
    - Knoten/Kanten mit Eigenschaften (Industrie vs. Wissenschaft, Agrar vs. Telekommunikation, Kontraktformen, etc.)
  - welche Partitionierung in Communities ergibt instruktive Einsichten?

Algorithmus 1: RandomSetModel

1. Zielgröße: P Projekte, O Organisationen
2. Wähle eine Projektgrösse $|P_i|$ aus
   - <-> tatsächliche Projekt-Größenverteilung aus empirischer Untersuchung
3. Für das Projekt P_i wähle $|P_i|$ zufällige Organisationen aus dem Organisations-Pool
4. Wiederhole 2. und 3. bis P Projekte kreiert wurden
5. Identifiziere eventuell unbenutzte Organisationen, lösche sie (d.h. Pool muss vorher etwas größer sein)
6. Speichere Netzwerk als bipartiten Graph zur Weiterverarbeitung

Algorithmus 2: bipartites configuration modell (~MolloyReed)

Methode wird bei Paper 2 genauer erklärt…

Haupt-Eigenschaft:

Die insgesamt P Projekte, O Organisationen haben asymptotisch *dieselben* Projekt-Größen $|P_i|$ und Organisations-Größen $|O_j|$ *Verteilungen* wie die empirischen Netzwerke

Um empirische und generierte Graphen zu vergleichen …

… sehen wir immer *3 Diagramme zusammen*

Empirical Network: „FP3“

3rd framework programme

O=9615 P=5529 M=31380

Project-Sizes: min=1 max=73 mean=5.6
Orgas-Sizes: min=1 max=138 mean=3.2

RandomSet Network

~ same #orgs, #projs as FP3

~ same project sizes as FP3

bipMolloyReed Network

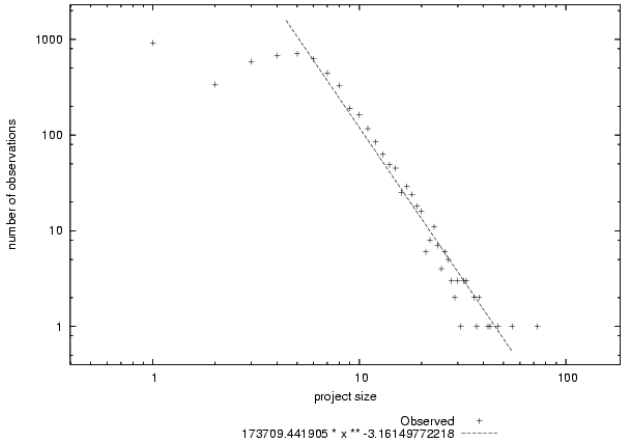
~ same #orgs, #projs as FP3

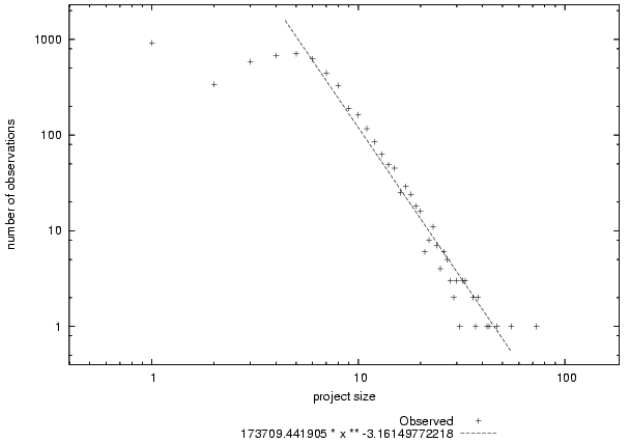
~ same project sizes as FP3
~ same organization sizes as FP3

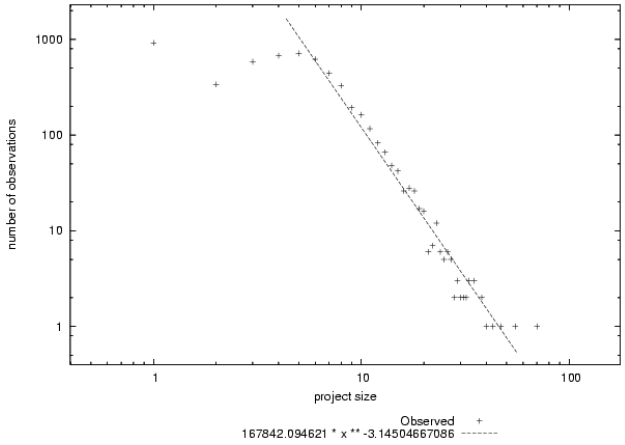
Project **Sizes**:

identical

Because the empirical project sizes are the **inputs** for both simulations



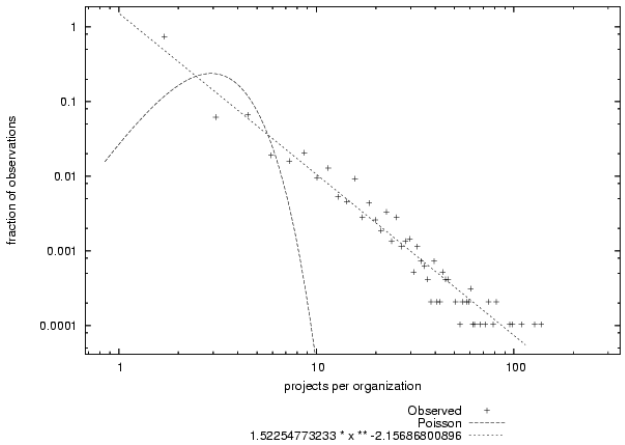


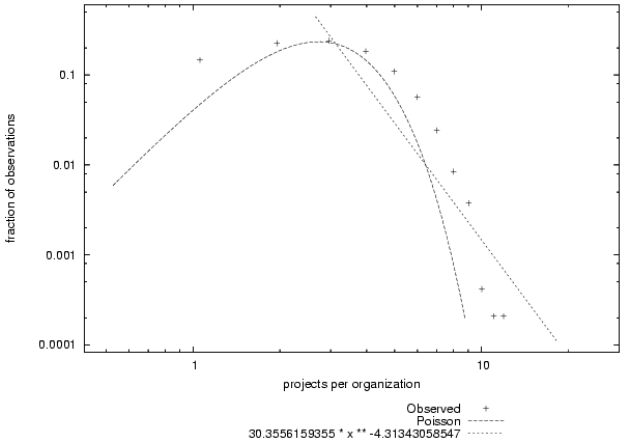


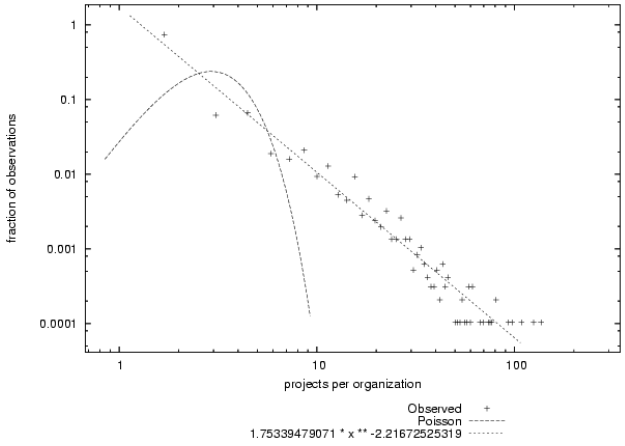
Organization **Sizes**:

RandomSetModel:
Exponential Decay ~ Poisson Distrib.
predicted by theory!

FP3 <-> MolloyReed
identical because input



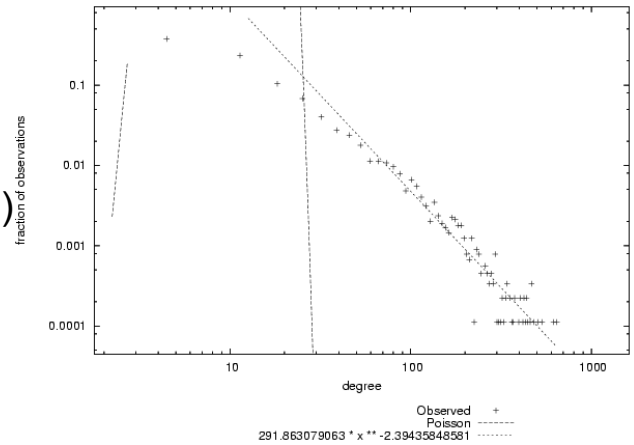


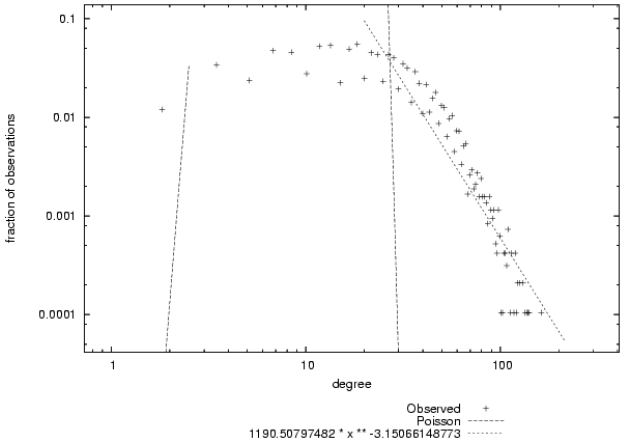


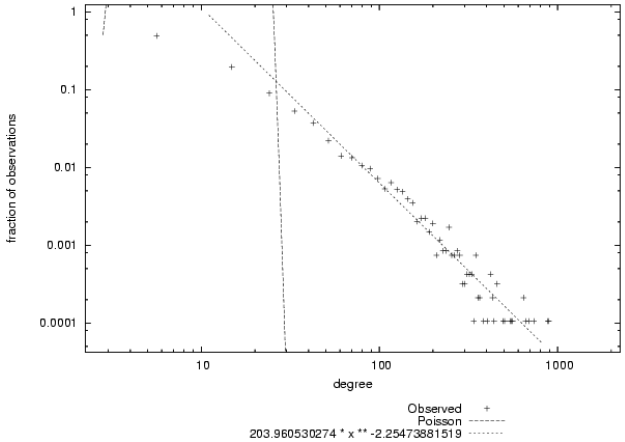
Organization Graph:
Projection onto organisations

RandomSet:
fat tail *but* much steeper (exponent >3.2)

MolloyReed:
similiar to empirical FP3;
both are ~ on a straight power law
with exponent 2.3 - 2.4





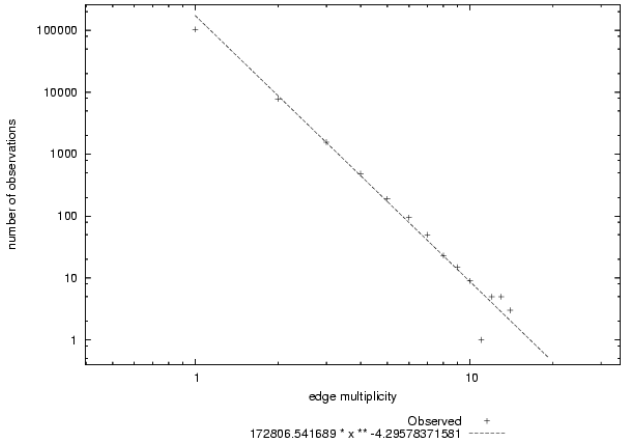


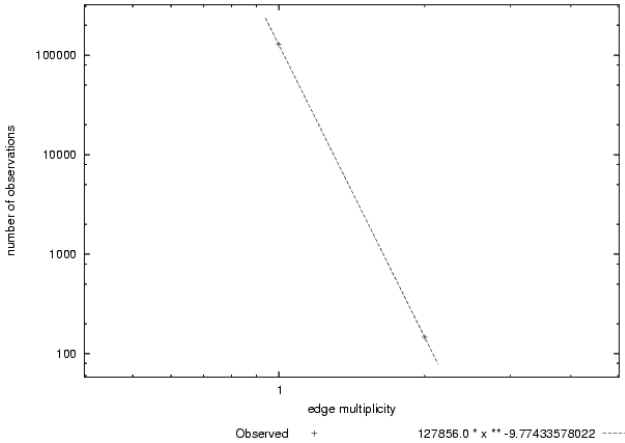
Organization Graph:
Edge Multiplicities

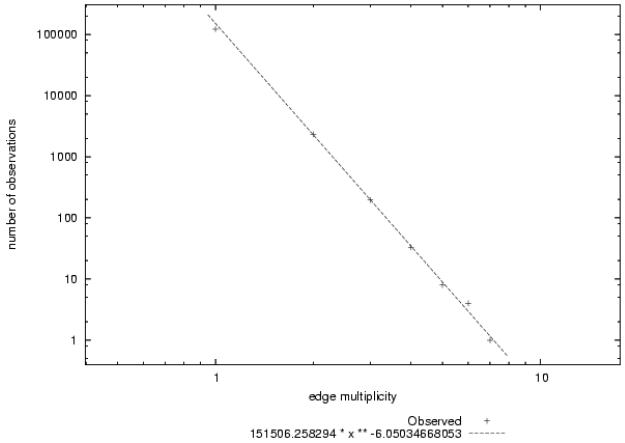
Empirical FP3:
highest multiplicity 14

RandomSet:
only 1 and 2

MolloyReed:
highest multiplicity only 7



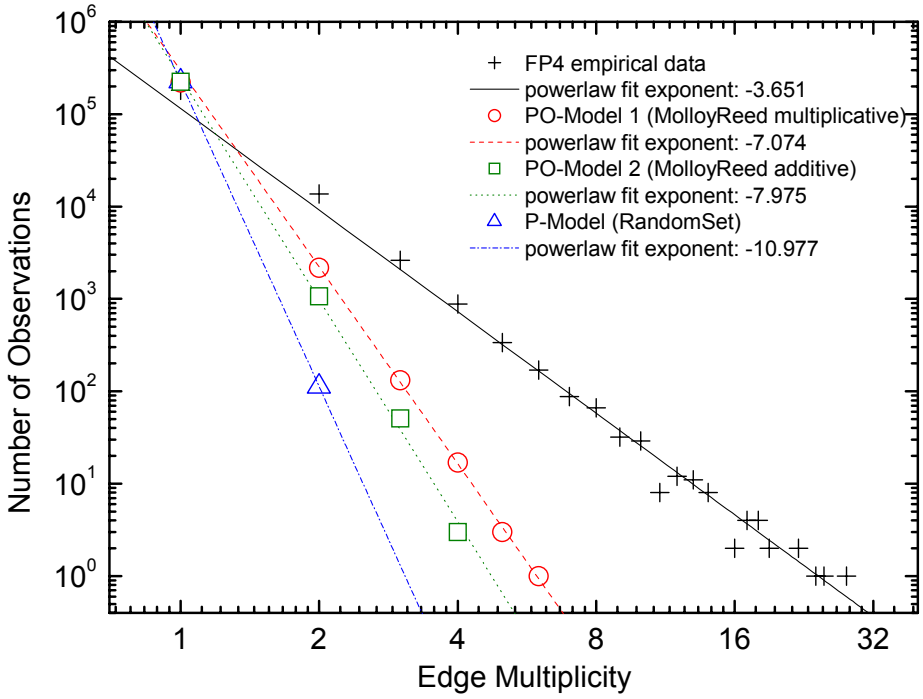




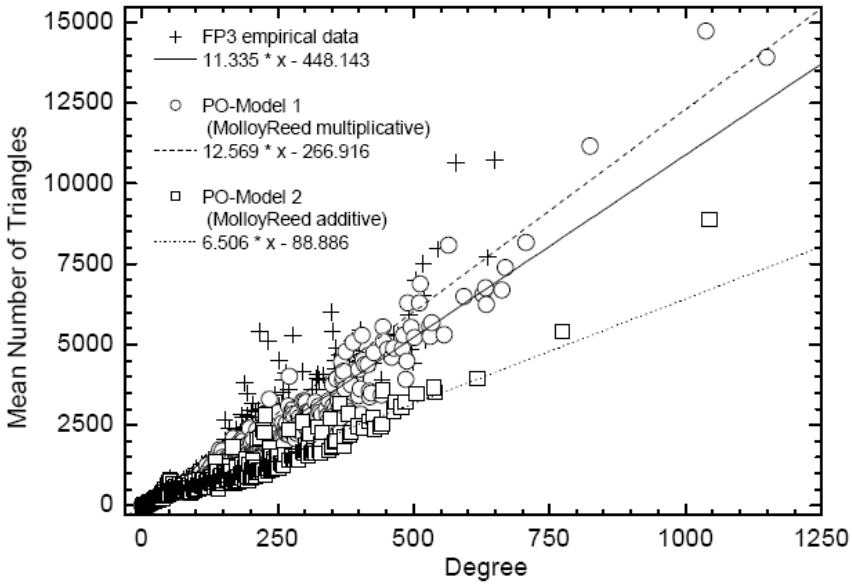
Degree-Korrelation

- Molloy Reed: multiplicative vs additive
- Modell wird bei Paper 2 genauer erklärt
- Schematisch:
  Wahrscheinlichkeit für Kante zwischen zwei Knoten mit gegebenem Degree ist
  - Multiplikativ:
    proportional zum Produkt der Degrees
  - Additiv:
    proportional zur Summe der Degrees

Multiplicative vs. Additive: OrganisationsGraph Edge Multiplicity



Multiplicative vs. Additive: OrganisationsGraph Mean Triangles vs. Org-Degree



Wichtige Aspekte dieses 1.Papers

- Dieses allererste Paper diente vor allem der *empirischen* Datenanalyse:
  - Anwendung des „Netzwerk-Hypes“ der Physik auf den Datensatz, Zusammenarbeit mit dem ARC in Wien
  - Handelt es sich um skalenfreie Netze?
  - Welche Kenngrößen erheben wir?
  - etc.
- Meine Aufgaben:
  - Wissenstransfer nach Wien
  - Python erlernen (Michael Barber)
  - Netzwerke programmieren lernen
  - Zufalls-Netzwerke als Vergleichsobjekte mit der Empirie: VIEL INFORMATION LIEGT IN DEN GRÖSSENVERTEILUNGEN !
- Konsequenzen bis heute:
  - NEMO www.nemo-net.eu
  - „Network Models, Governance and R&D Collaboration Networks“
  - EU project ~ 1.6 Mio € (3 Jahre) zur Erforschung des Datensatzes

GEP = Generalized Epidemic Processes

Corruption as a GEP

Philippe Blanchard
Andreas Krueger
Tyll Krueger
Peter Martin

arxiv:physics/0505031

Corruption?
Imagine any contagion process with

1. Neighbour infection
   - **Threshold** contagion, i.e. local infection only if „level of corruption of my neighbours exceeds Δ“
   - plus small infection probability if less than Δ
2. Mean field infection
3. Mean field desinfection

e.g.:

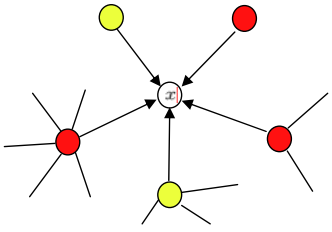
- opinions, fashions, …
- waves of scientific hypes, discussed topics…
- innovation processes
- ***Corruption…***

Corruption state variables

ω(x, t) in [0,1]
= node x is corrupt/non-corrupt at time t

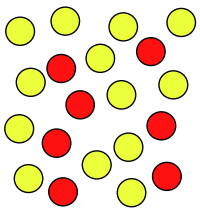
Ω(x, t)
= number of infected neighbours

$$= \sum_{y \sim x} \omega(y,t)$$



b_t = total prevalence of corruption at time t

$$b_t = \frac{1}{|V|} \sum_{y \in V} \omega(y,t)$$



7 / 20 = 0.35

Implemented LOCAL Processes:

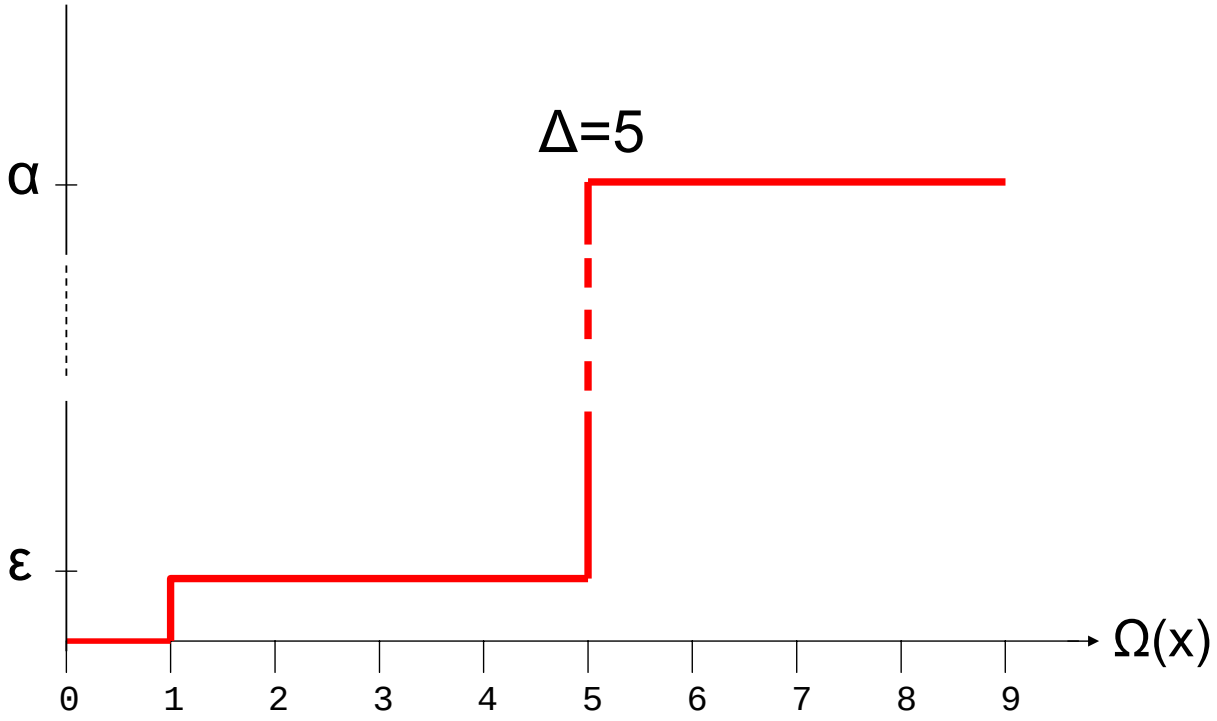
α-process: "if enough neighbours…"

- The **local** transmission probability for # of corrupt neighbours ≥ Δ
- Typical value: α>>ε,β,γ
- Possible translations: **Influenceability by others, Decisiveness**

ε-process: "if at least one neighbour …"

- (similar to classical) **local** epidemic probability for # of corrupt neighbours < Δ
- Typical value: ε<<α,β,γ (very small)
- Possible translations: **Naiveity**

Local infection probabilities, (absolute Δ=5 threshold)



Implemented GLOBAL Processes:

β-process: "infection through public opinion"

- The mean field transmission process due to the total (believed) prevalence of corruption
- Typical value: ε<β<γ
- Possible translations:
  "**Random" infection: How informed are you?**
  How much do you belief in mass media?

$$\Pr_{\beta}(\omega_{t+1}=1 \mid \omega_t=0)=\beta(b_t)(1-(1-b_t))=\beta(b_t)^2$$

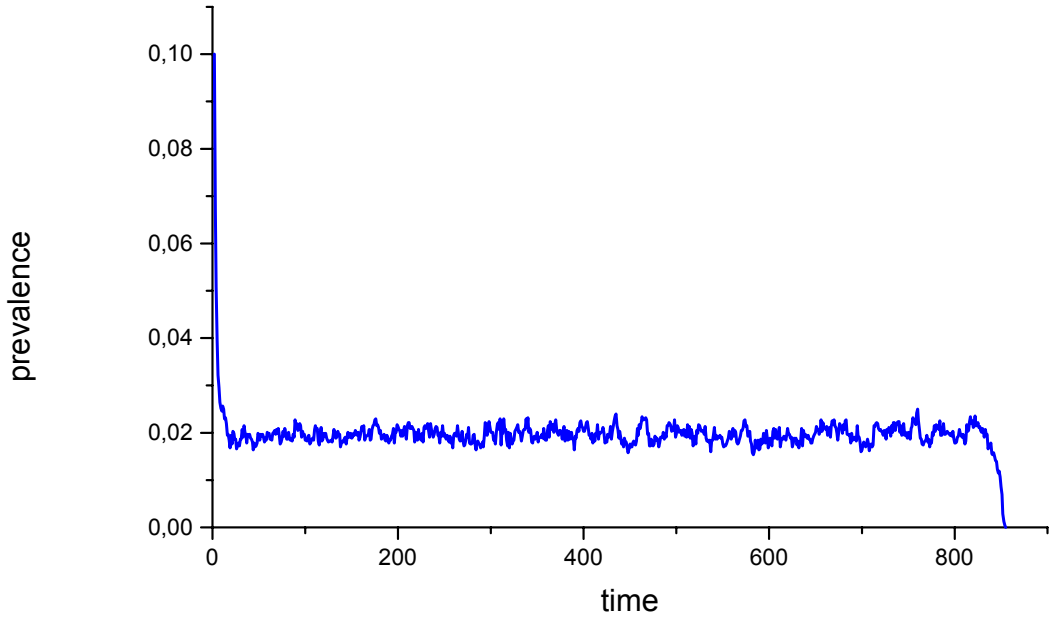
γ-process: "(only) the healthy can cure others"

- The (mean field!) corruption recover process due to the fight of the (healthy) society against corruption
- Typical value: β<γ<α
- Possible translations:
  random resistance / recovering / cleaning

$$\Pr_{\gamma}(\omega_{t+1}=0 \mid \omega_t=1)=\gamma(1-b_t)$$

Single run until stagnation 1

Low semistable prevalence in a real collaboration network

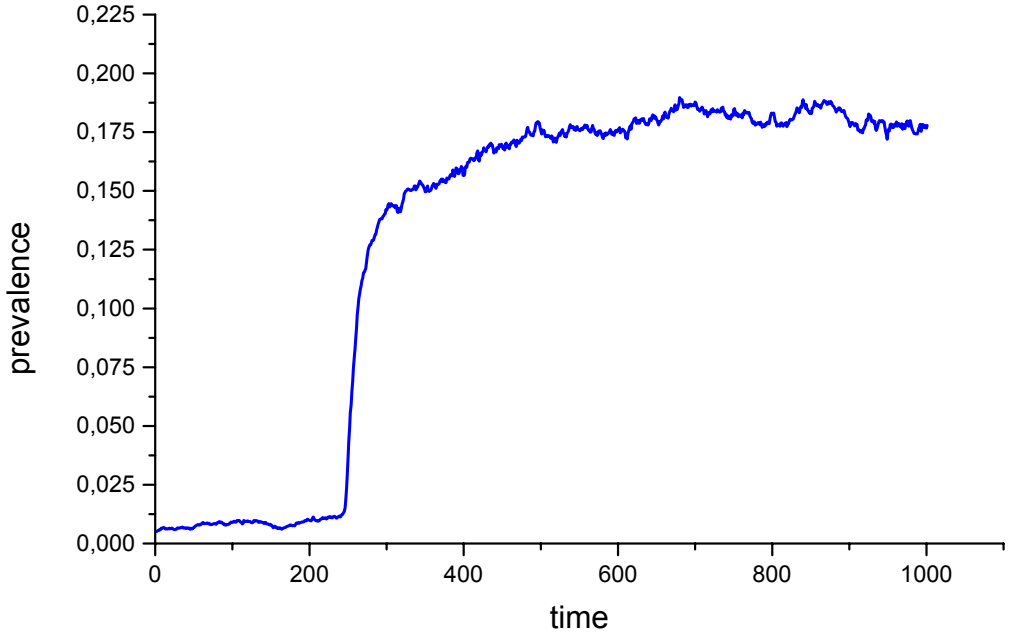


Δ=30 ε=0 α=0.99 β=0.09 γ=0.545 b_0=0.10

Network FP2: N=4879 M=57633 mean degree=23.6 mean triangles=256.9

Single run until stagnation 2

Jump from very low prevalence to low prevalence

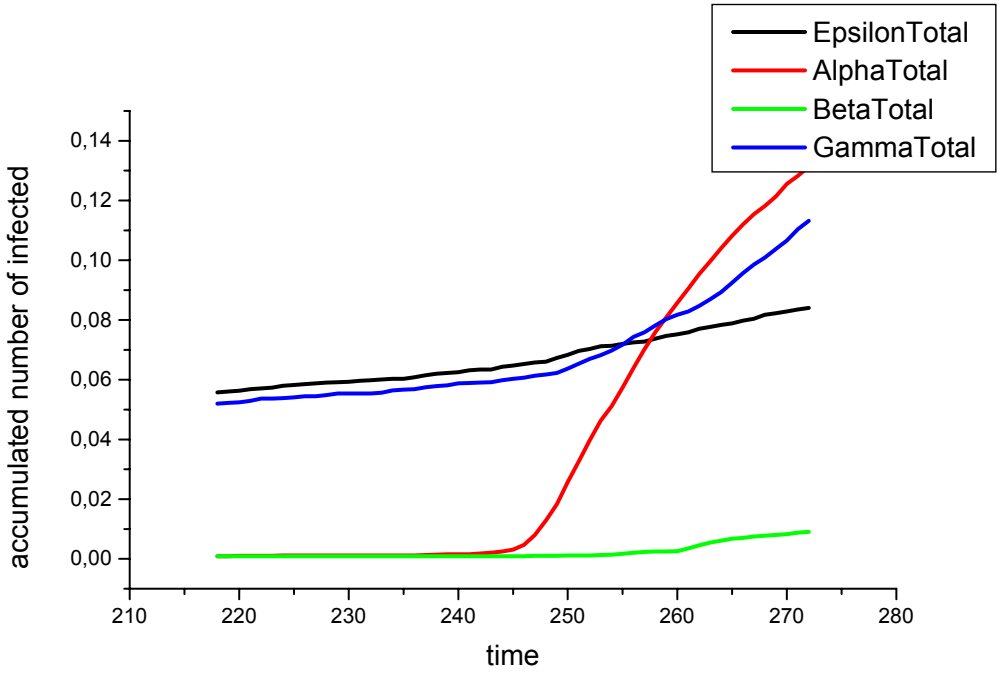


Δ=25 ε=0.001 α=0.20 β=0.04 γ=0.03 b_0=0.005

Network FP3: N=7710 M=93852 mean degree=24.3 mean triangles=418.1

Single run until stagnation 2

Contribution of the 4 (des)infection paths for the “jump”



Structure of the Python Program

Initial infection with b_0 corrupt nodes (random/ball)

→update one vertex

→all vertices, sync update

→do until stagnation

Example Legend

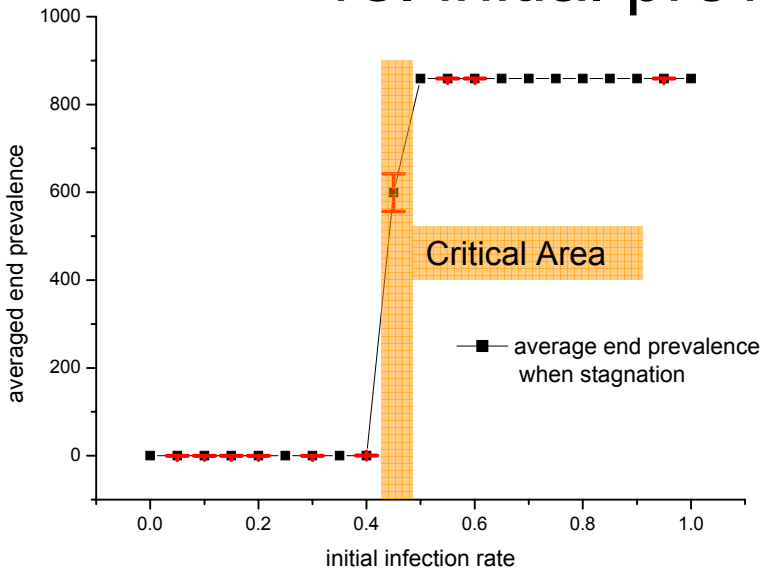
get: end prevalence (usually ~0 or ~1)

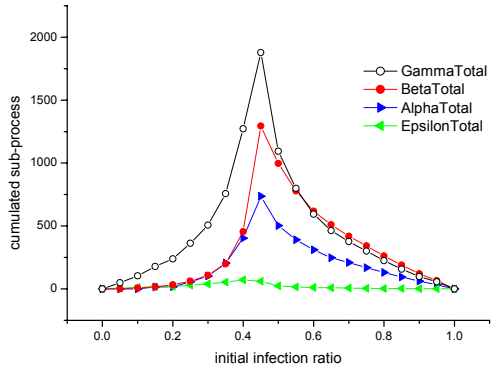
→ many runs to get *average* end prevalence

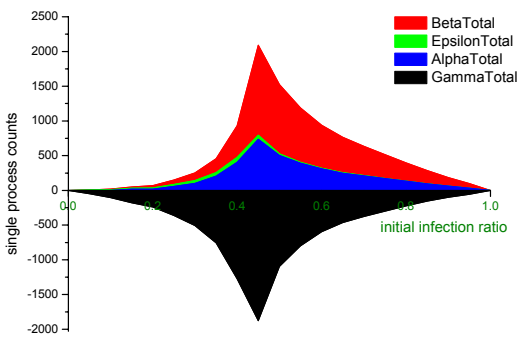
→Transition finder: vary b_0 to locate $b_{undercrit}$ and $b_{overcrit}$ and get (mean value) b_{crit}

→ sweep (network property) X=N, M, T or λ to plot b_{crit} over X

end prevalence vs. initial prevalence



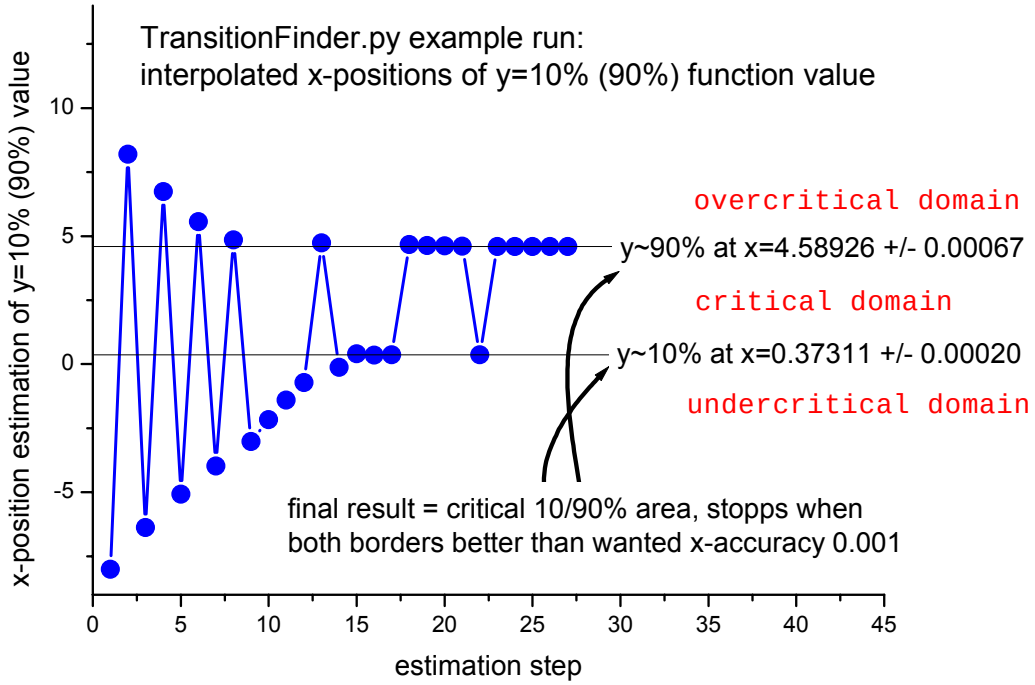




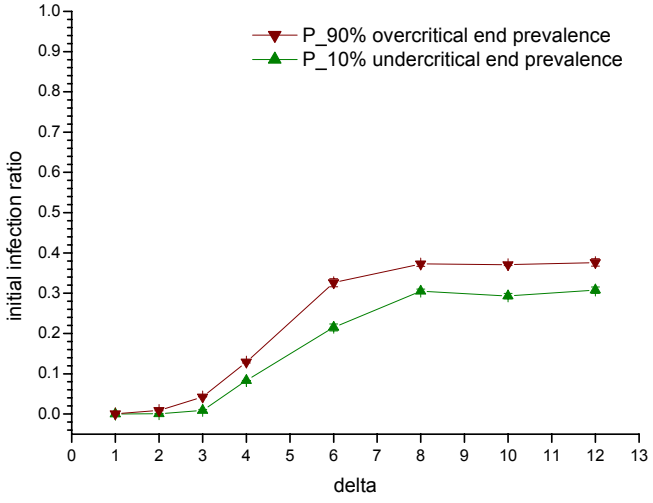
Δ=8 ε=0.001 α=0.20 β=0.08 γ=0.06

N=859 M=3368 by Random Set Model (O=1000 P=500)
degree: min=2 max=33 mean=7.8
44% nodes have degree ≥ 8

Find x_{crit} but avoid critical area by linear interpolation of 10% / 90% y-value

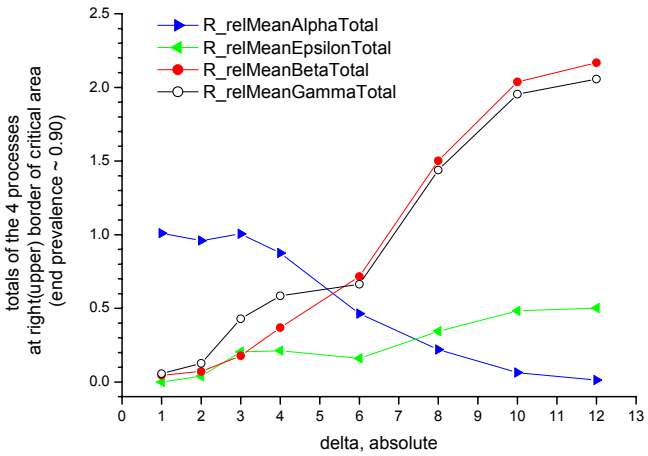


Critical Density over Δ = neighbour infection threshold



Lower and upper bounds for b_{crit}
as a function of neighbour infection threshold Δ

ε=0.005 α=0.35 β=0.08 γ=0.04
GNM with N=1500 M=5000
→ mean degree ~ 6.7

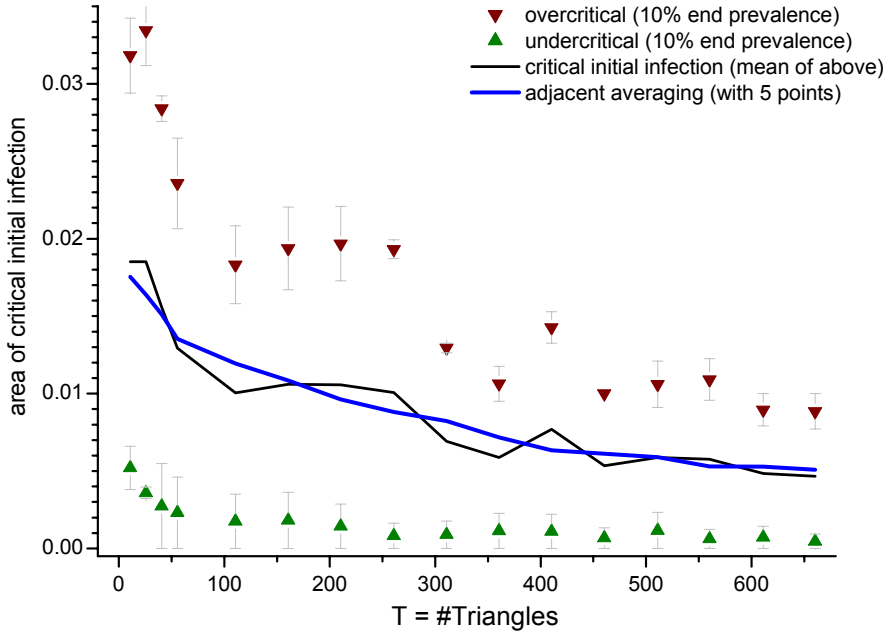


Total number of state changes
splitted into the 4 different sub-processes

Clustering helps corruption

- In classical epidemics, local clique-clustering slows down the disease spreading because of **re-infection** instead of the infection of healthy
- Here, though, the highly clustered, medium-degree vertices are especially well-suited for the spread of corruption, because a threshold Δ of neighbours has to be corrupt to trigger the α-process

Critical Density over T = #triangles



Δ=2 ε=0.005 α=0.30 β=0.08 γ=0.04 (20 runs averages)

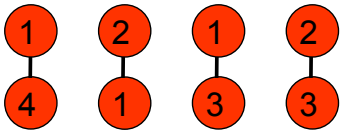
Network GNTM: N=1000 M=2000

Configuration Modell Graph Generator
(~ ”Molloy Reed”)

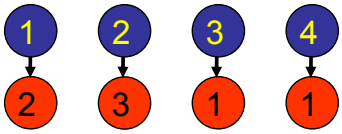
1) ***Wanted* Degree Distribution**, e.g. once k=3, twice k=2, once k=1

2) For each node draw k from distribution and create k clones (“virtual nodes”)

3) Random Pairing



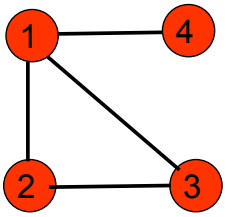
Our modification:
For **addditive** (instead of **multiplicative**)
DegreeDegree Correlation:
3b) Pairing with ~equal “Out-Degree” for each node



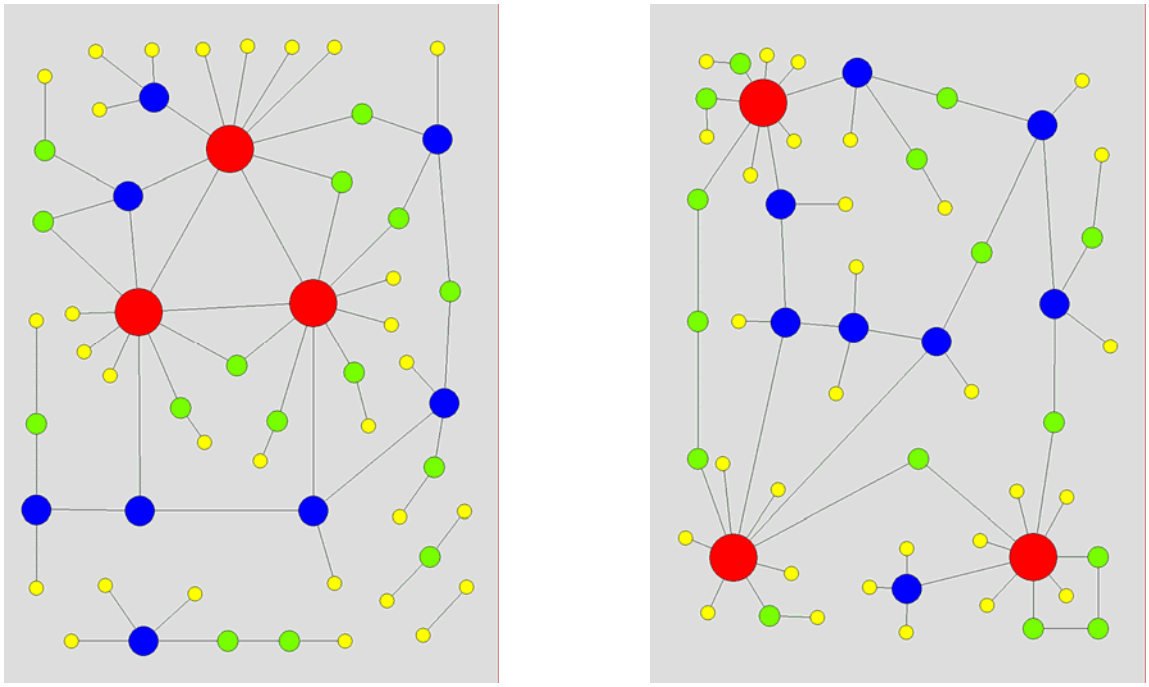
4) Identify again all virtual nodes from same originator

5) Remove double- and self-links

6) The result: A network with the ~wanted degree distribution:

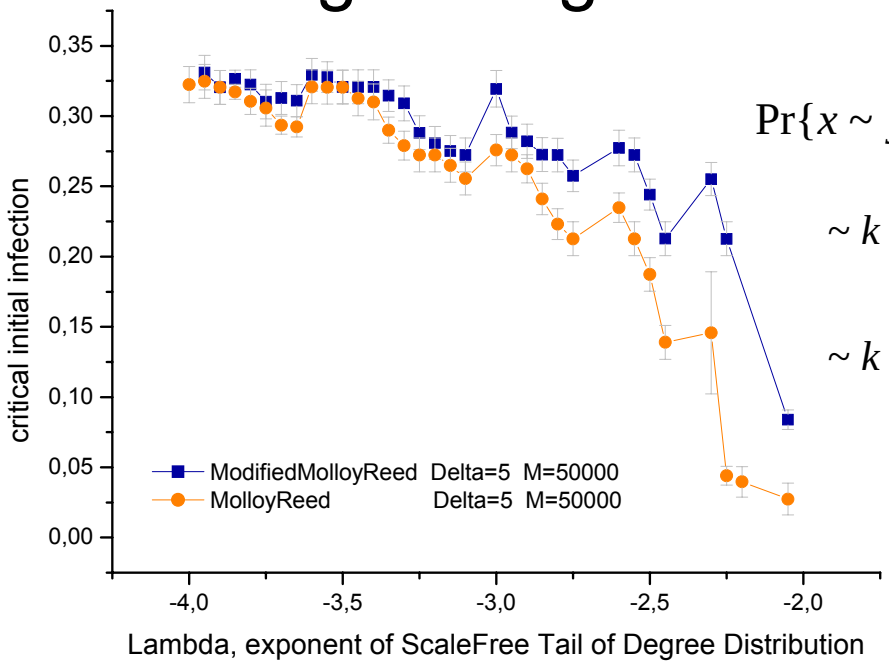


Multiplicative vs. Additive
Degree-*Correlation*



N.B.: identical degree *distribution:*
3 reds (degree 10); 8 blues (degree 4), 15 greens (degree 2), …

Additive vs. Multiplicative DegreeDegree Correlation



$$\Pr\{x \sim y \mid d(x) = k \wedge d(y) = k'\}$$

$\sim k \bullet k'$ multiplicative (normal MolloyReed)

$\sim k + k'$ additive (modified MolloyReed)

DegreeDegreeCorrelation

Δ=5 ε=0 α=0.35 β=0.08 γ=0.04

N=20000 M=50000 generated by (modified) MolloyReed with given Degree-Histogram

Additive vs. Multiplicative DegreeDegree Correlation

- SF-Networks with multiplicative DegreeCorrelation (hierarchical, …) are more easily corrupted than those with additive DegreeCorrelation (polycentric, democratic)
- Especially true for low $\lambda < 3$ (where *very big hubs* exist).

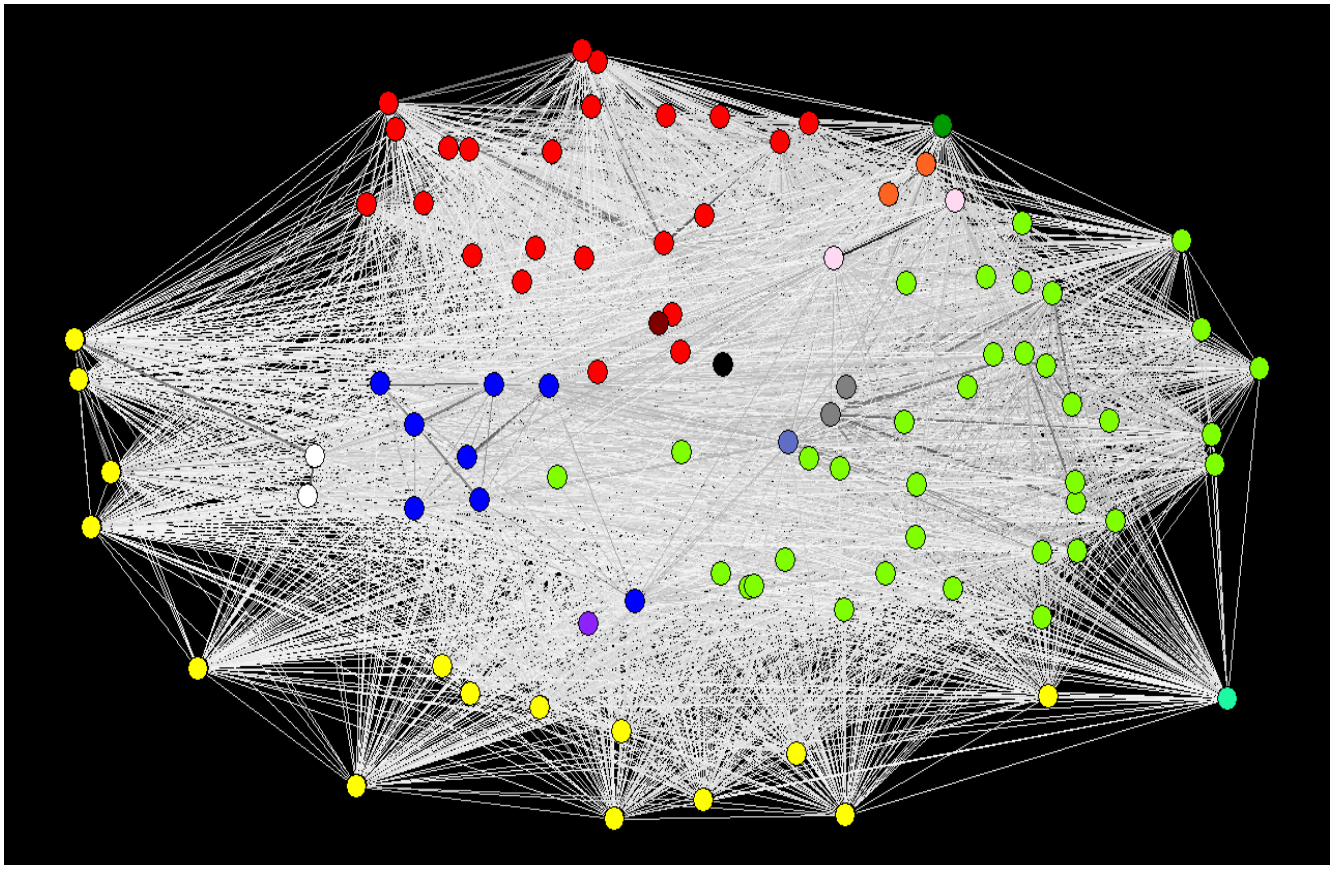
Epidemic Control

- This is an ABSTRACT model!
  Only structural & schematic tendencies!
- Positively correlated to corruption:
  α & ε = strength of influence of others
  β = strength of e.g. mass media
- Negatively correlated to corruption:
  Δ=How many neighbours have to be corrupt?
  γ=How strong does the society fight back?
- avoid high clustering
- “Transparency”: Δ↑ α↓ β↓
- “Police”: β↓(increase of fear), γ↑(uncovering rate)
  but γ> α, β is a “total police state”
- Moral resistance: Δ↑ α↓
- (Hierarchical) Decision Systems should be as flat, independent, polycentric as possible!

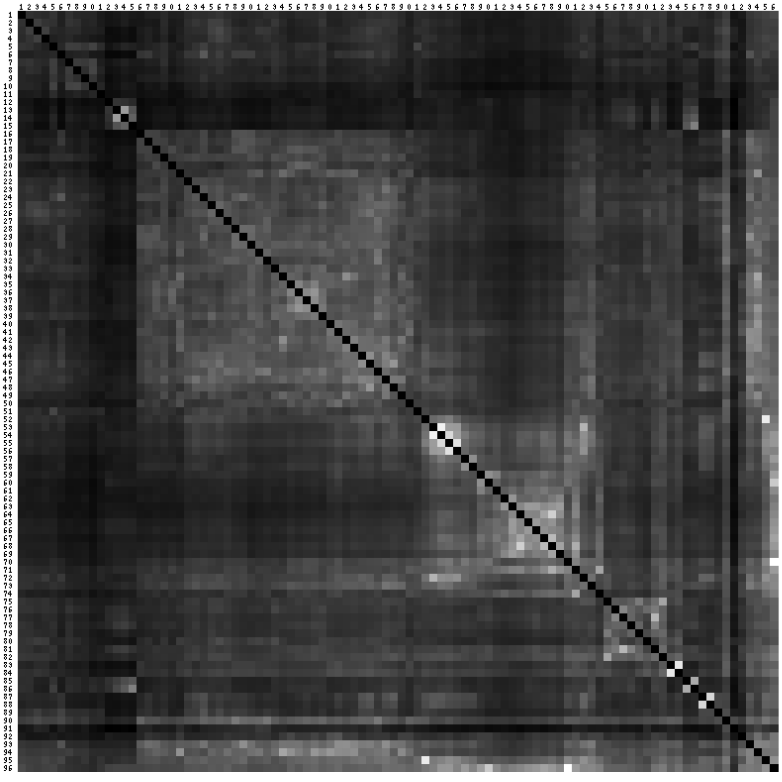
Paper 3

Clustering by Adjacency Matrix Block Ordering (CAMBO)

Lymphoma Tumor Tissue Samples



Lymphoma Tumor Tissue Samples



Adjacency matrix
(dark = low weight)

96 tumor samples

Showing their mutual similarity with respect to 4026 gene log expression levels.

Weighted (+,+) Projection from 96x4026 data table

Best clustering so far with

numCl=14
modularity=0.06654

at b=9 c=5
Mean weight a=0.22

Line Difference Sorting

Idea: Block-sort the matrix

1) Start with any line as the first one

2) add that line as the next one
that has the *lowest line difference*
to *all* previously chosen lines

What to choose as line difference?

Adjacency Matrix and Square

A_{ij} Unweighted:
=0 unconnected
=1 there is a link between node i and j

N_1-neighbourhood of i

Weighted:
How similiar/connected are nodes i and j

$$(A^2)_{ij} = (\sum_k A_{ik} A_{kj})_{ij}$$

~ N_2-neighbourhood of i

Unweighted:
How many paths of length 2 are there between node i and j

Weighted:
How strong are paths-of-length 2 between i and j over all other nodes

$A_{ij} > 0$ and $(A^2)_{ij} > 0$

→ There are TRIANGLES
connecting node i and j

A_{ij} **N_1-neighbourhood of i** $(A^2)_{ij}$ **N_2-neighbourhood of i**

Structural Equivalence

$$B^* = \sum_{k \neq i,j}^{N} | A_{ik} - A_{jk} |$$

How *different* is the **N1**-neighbourhood of nodes i and j

Structural Equivalence in N_2-neighbourhood

$$C^* = \sum_{k \neq i,j}^{N} | (A^2)_{ik} - (A^2)_{jk} |$$

How different is the **N2**-neighbourhood of nodes i and j

LineDifference Matrix L

$$L(\tau,\beta,\gamma) \equiv -1 \cdot \overline{A} - \tau \cdot \overline{A^2} + \beta \cdot B + \gamma \cdot C$$

$$\overline{A} \equiv \frac{1}{\alpha_1} A$$

α_1 : mean non-diagonal A_{ij} element

B & C: Including normalization for standard deviation 1 and for size N

$$L(\tau,\beta,\gamma) = -\overline{A} - \tau \overline{A^2} + \beta B + \gamma C$$ LineDifference Matrix L_{ij}

Rough sketch of the blocksorting:

1) Choose *any* node/line as the starting node/line

2) Choice of next best node:

- Of all previously chosen nodes { x }
- and all not-yet-chosen nodes { y }
- choose that node y with the lowest L_{xy} to *any* of the { x }

3) Store that L_{xy} as the „LineDifferenceStep“ S_y

4) Goto (2) until ready

5) Reorder the original matrix into that order

Blocksort: OK, but: Where are good clusters?

The „Newman Modularity“ Q can be used to quantify the goodness of a clustering as one real number:

The *degree* k_v of a vertex v is defined to be the number of edges incident upon it:

$$k_v = \sum_w A_{vw}. \tag{3}$$

The probability of an edge existing between vertices v and w if connections are made at random but respecting vertex degrees is $k_v k_w/2m$. We define the modularity Q to be

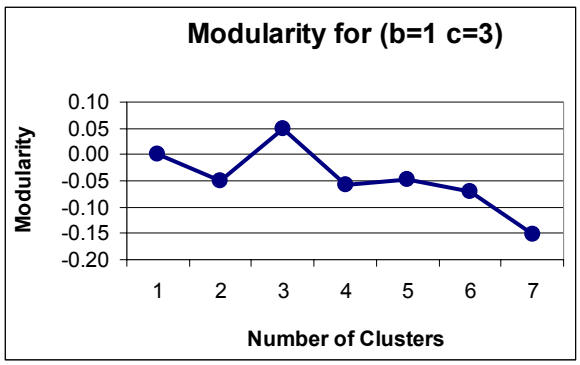
$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w). \tag{4}$$

If the fraction of within-community edges is no different from what we would expect for the randomized network, then this quantity will be zero. Nonzero values represent deviations from randomness, and in practice it is found that a value above about 0.3 is a good indicator of significant community structure in a network.

arXiv:cond-mat/0408187 v2 30 Aug 2004

Blocksort: OK, but: Where are good clusters?

- Sort the stored LineDifferenceSteps S_y in decreasing order, highest is first *cut position*
- Introduce more and more cut positions, and thus clusters, in that order
- Calculate Newman Modularity for each clustering
- Choose the one clustering with the *highest modularity* …



… Choose the one clustering
with the highest modularity …

… as the (locally) best clustering
(for this [tau, beta, gamma] – point)

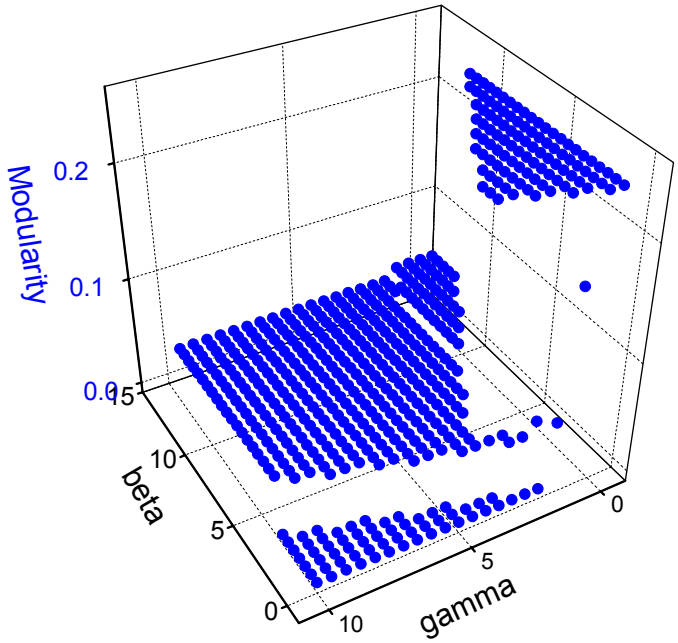
$$L(\tau,\beta,\gamma) = -\overline{A} - \tau\overline{A^2} + \beta B + \gamma C$$

three FREE parameters, mostly set gamma=0

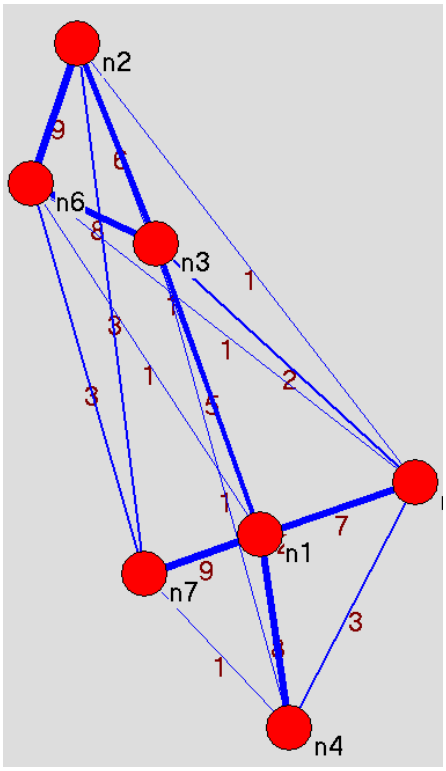
For each [tau, beta, gamma] there is a (locally-) best clustering

Globally best clustering

- Scan the (tau, beta, gamma) manifold for the *overall highest Newman modularity*



Example network



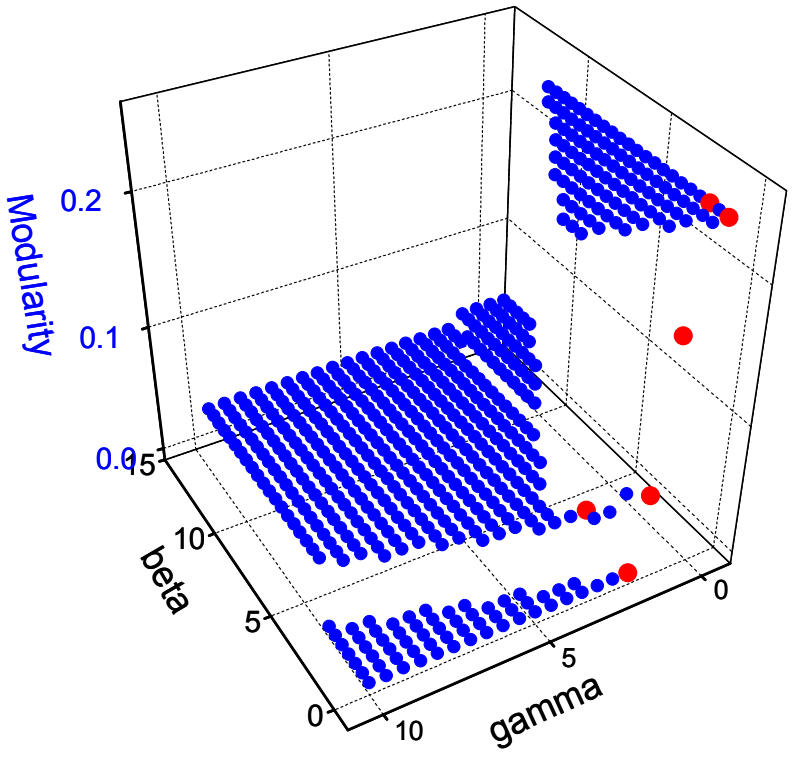
Adjacency matrix A

```
 158719
1  6 193
56  128
8  1  3  1
7123  12
198  1  3
93  123
```

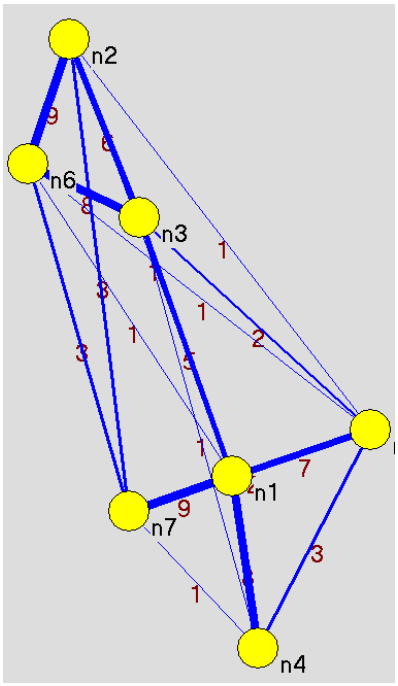
Difference Matrix B* of A

0	29	33	18	15	31	21
29	0	10	25	22	2	22
33	10	0	19	18	12	12
18	25	19	0	5	27	9
15	22	18	5	0	24	10
31	2	12	27	24	0	24
21	22	12	9	10	24	0

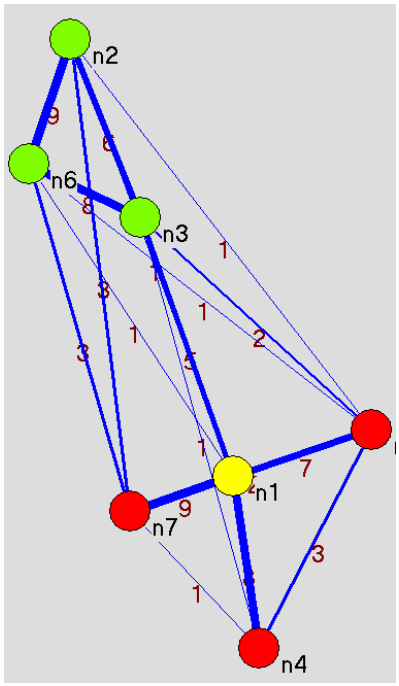
beta, gamma – scan:



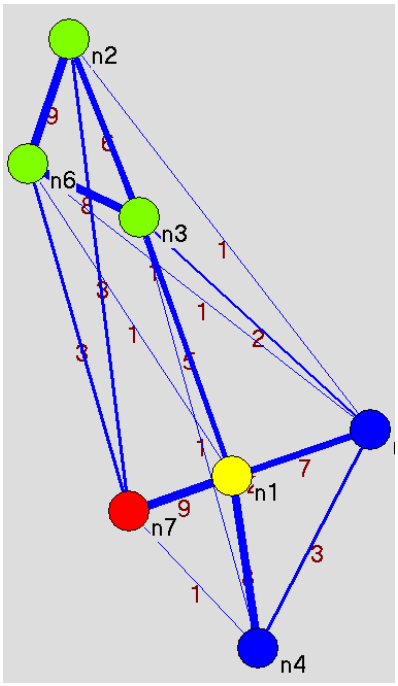
Let‘s look at some of these clusterings …



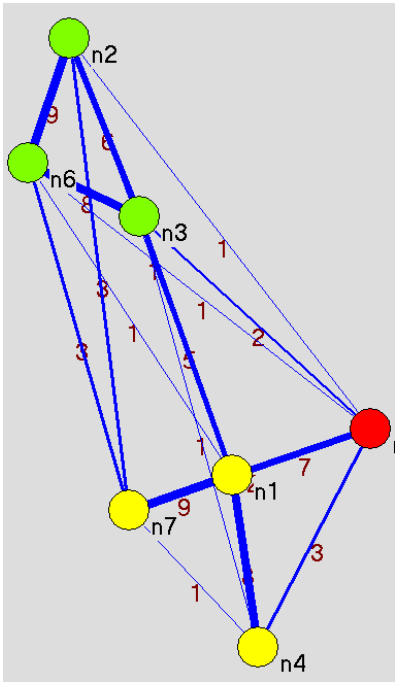
modularity=0
b=0 c=2



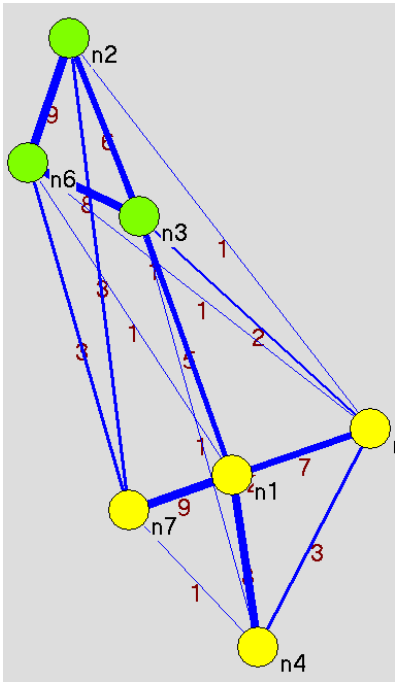
modularity=0.04811
b=1 c=3



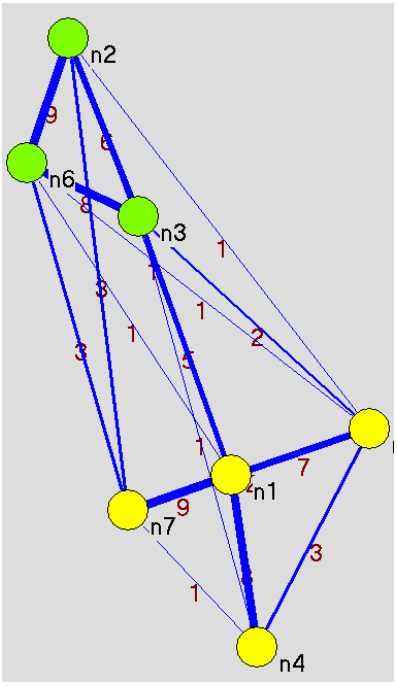
modularity=0.05763
b=0 c=1.5



modularity=0.17100
b=0 c=1



modularity=0.24162
b=1 c=0

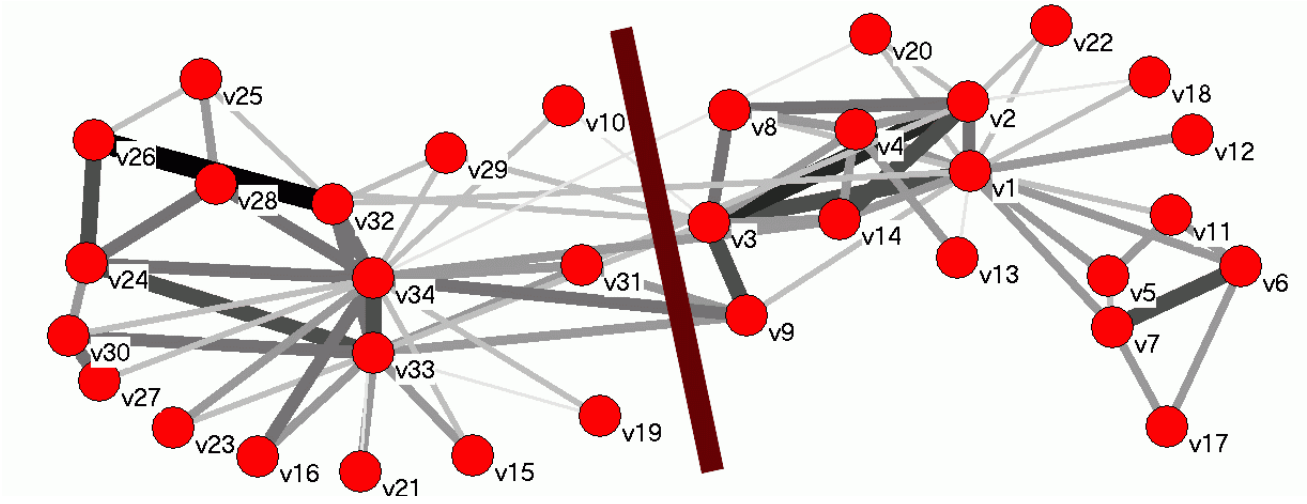


modularity=0.24162
b=0 c=0

ZACHARY KARATE CLUB

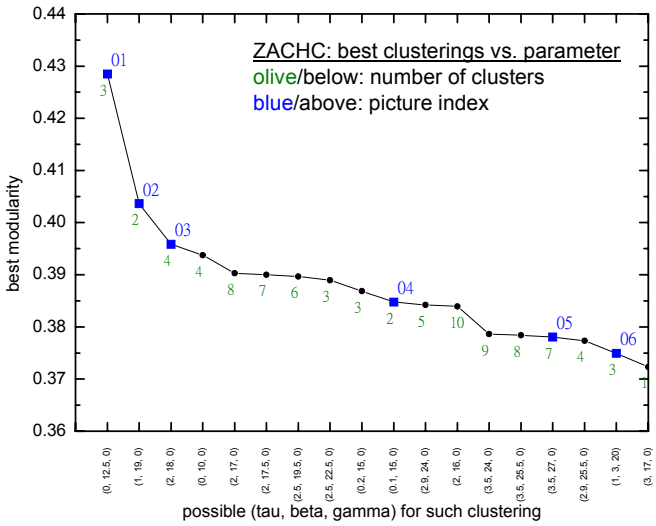
- **DATASET** ZACHARY
- **DESCRIPTION** Two 34×34 matrices.
- ZACHE symmetric, binary.
  ZACHC symmetric, valued.
- **BACKGROUND** These are data collected from the members of a university karate club by Wayne Zachary. The ZACHE matrix represents the presence or absence of ties among the members of the club; the ZACHC matrix indicates the relative strength of the associations (number of situations in and outside the club in which interactions occurred).
- Zachary (1977) used these data and an information flow model of network conflict resolution to explain the split-up of this group following disputes among the members.
- **REFERENCE**
- Zachary W. (1977). An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33, 452-473.

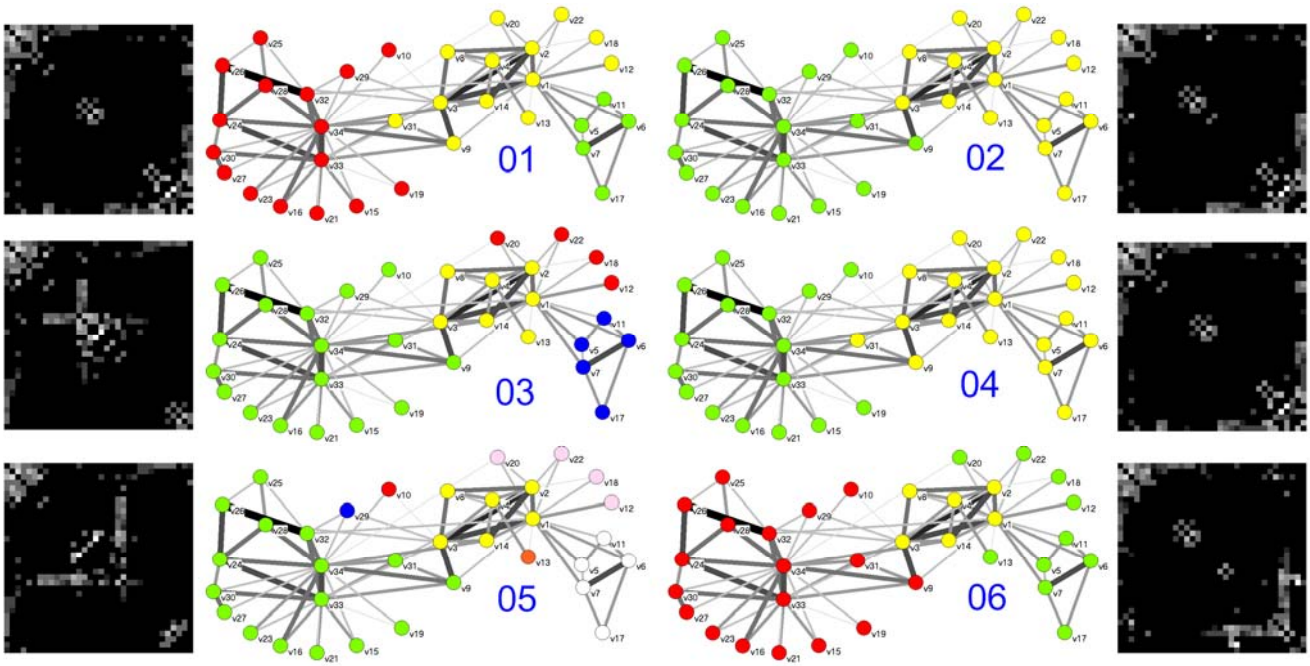
Zachary Karate Club Split



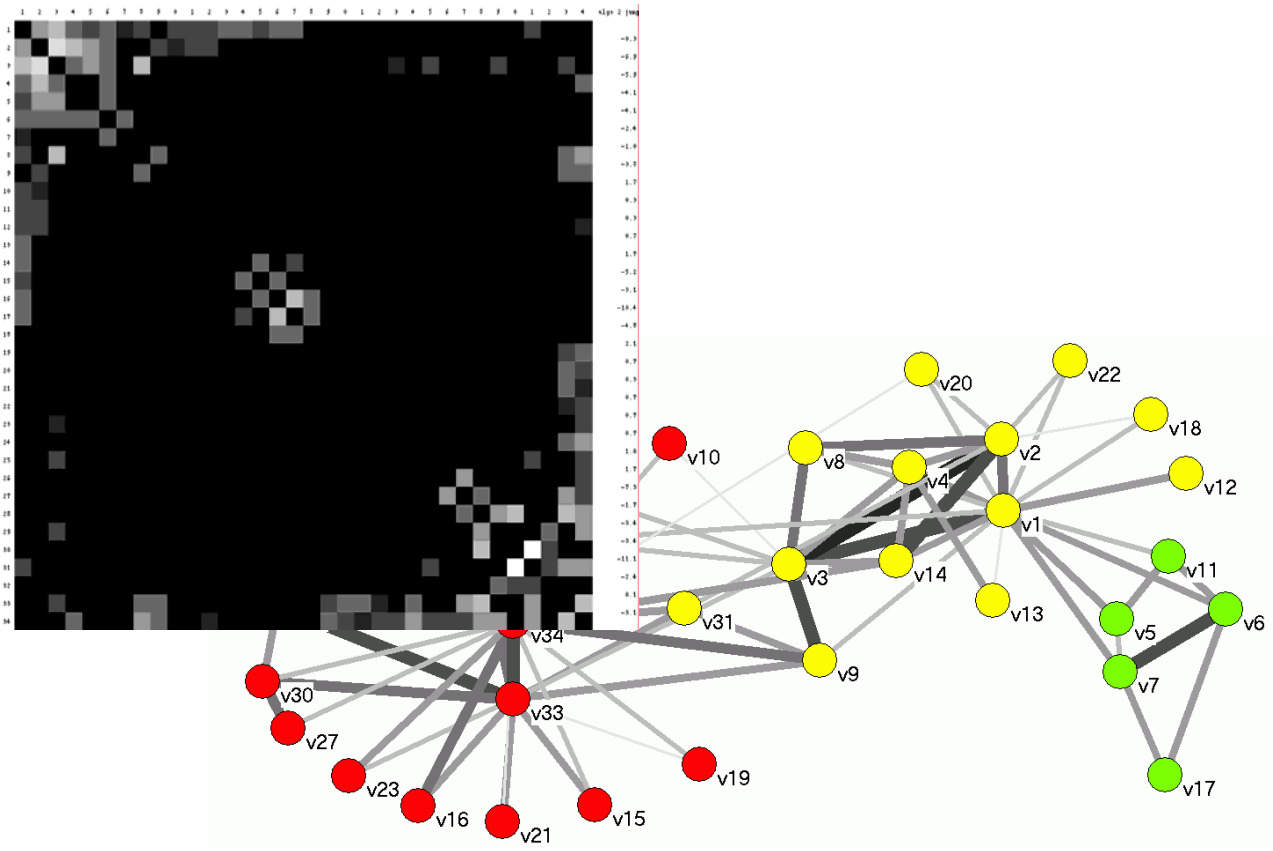
Zachary Karate Club Clustering

pic#	Cluster	Modularity	tau	beta	gamma	rank
01	3	0.42849	0	12.5	0	1
02	2	0.40363	1	19.0	0	2
02	2	0.40363	2	22.5	0	3
02	2	0.40363	1	5.0	10	4
03	4	0.39584	2	18.0	0	5
	4	0.39374	0	10.0	0	6
	8	0.39028	2	17.0	0	7
	7	0.39002	2	17.5	0	8
	6	0.38969	2.5	19.5	0	9
	3	0.38895	2.5	22.5	0	10
	3	0.38686	0.2	15.0	0	11
04	2	0.38474	0.1	15.0	0	12
	5	0.38422	2.9	24.0	0	13
	10	0.38394	2	16.0	0	14
	9	0.37866	3.5	24.0	0	15
	8	0.37840	3.5	25.5	0	16
05	7	0.37807	3.5	27.0	0	17
	4	0.37734	2.9	25.5	0	18
06	3	0.37494	1	3.0	20	19
	11	0.37232	3	17.0	0	20
	10	0.37214	4	14.0	10	21
	9	0.37188	4	15.0	10	22
	6	0.37118	4	16.0	10	23
		...				
07	3	0.22631	3.5	39.0	0	i
08	2	0.22507	1	7.0	20	i+1

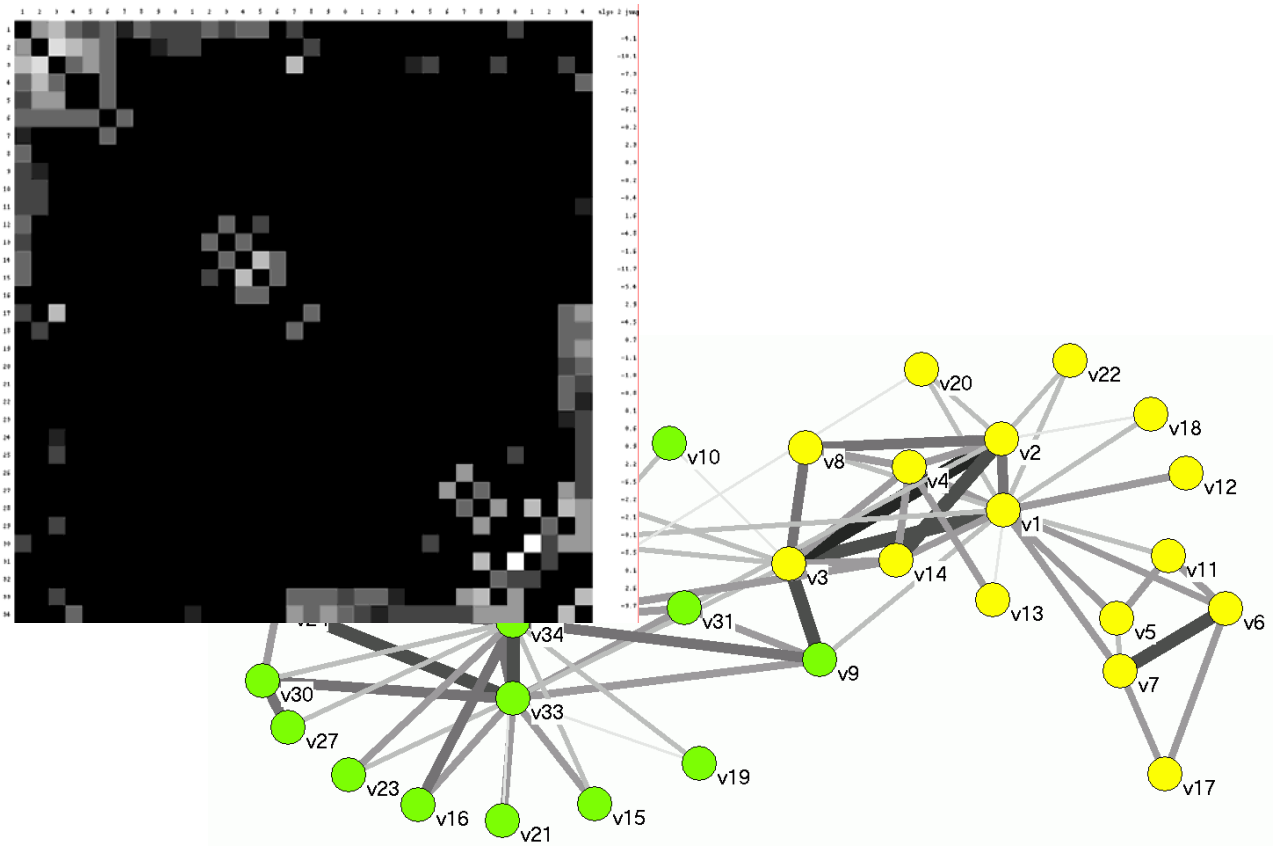




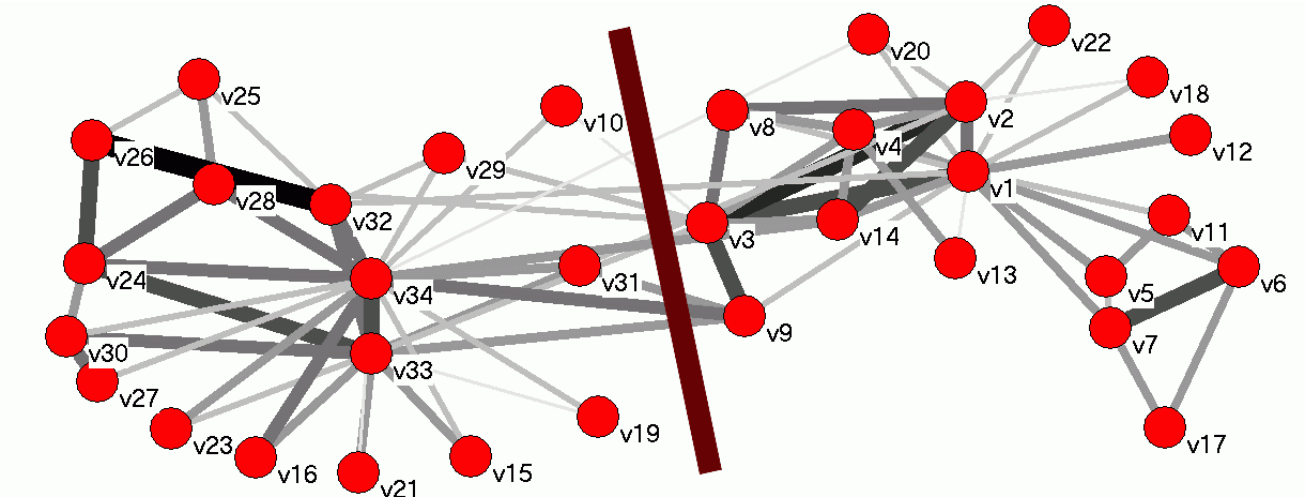
Zachary Karate Club BEST Clustering



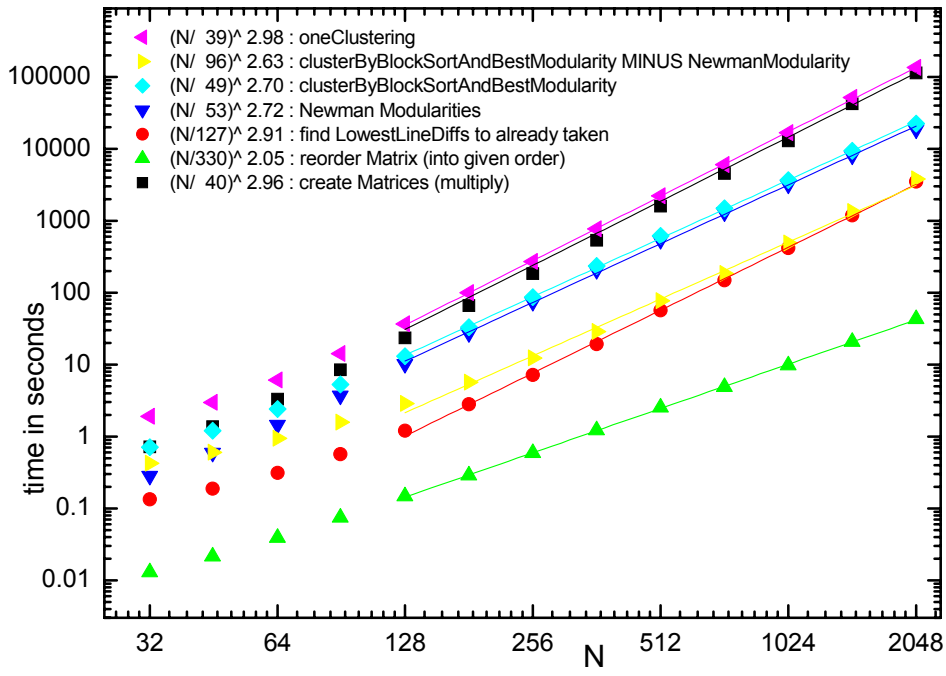
Zachary Karate Club second BEST Clustering



Zachary Karate Club Split



Time Complexity ~ O(N^3)



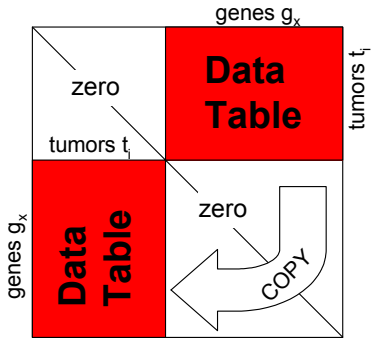
plans

- Instead of scanning the parameter plane, implement a search (hill-climbing)
- Implement portions in faster languages

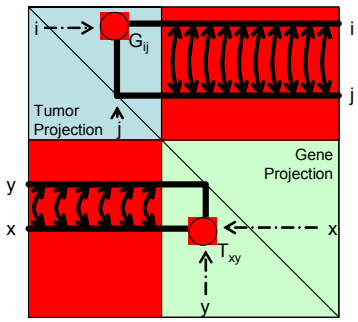
- OR: Abandon the algorithm!

- Once the projections (genes, tumors) are clustered, go back to bipartite picture

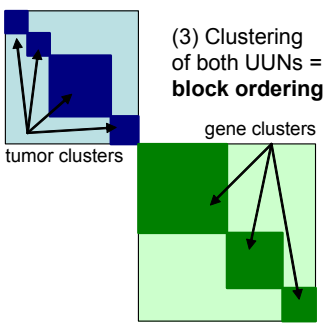
Data Tables = Bipartite Weighted Networks



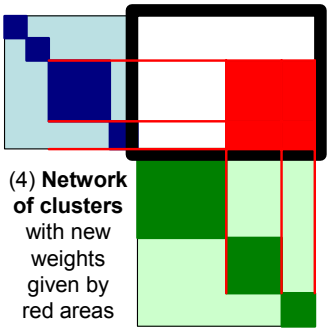
(1) From *any data* table make UBAM
Undirected Bimodal Adjacency Matrix



(2) Weighted Projection in 2 UUNs
Undirected Unimodal Networks



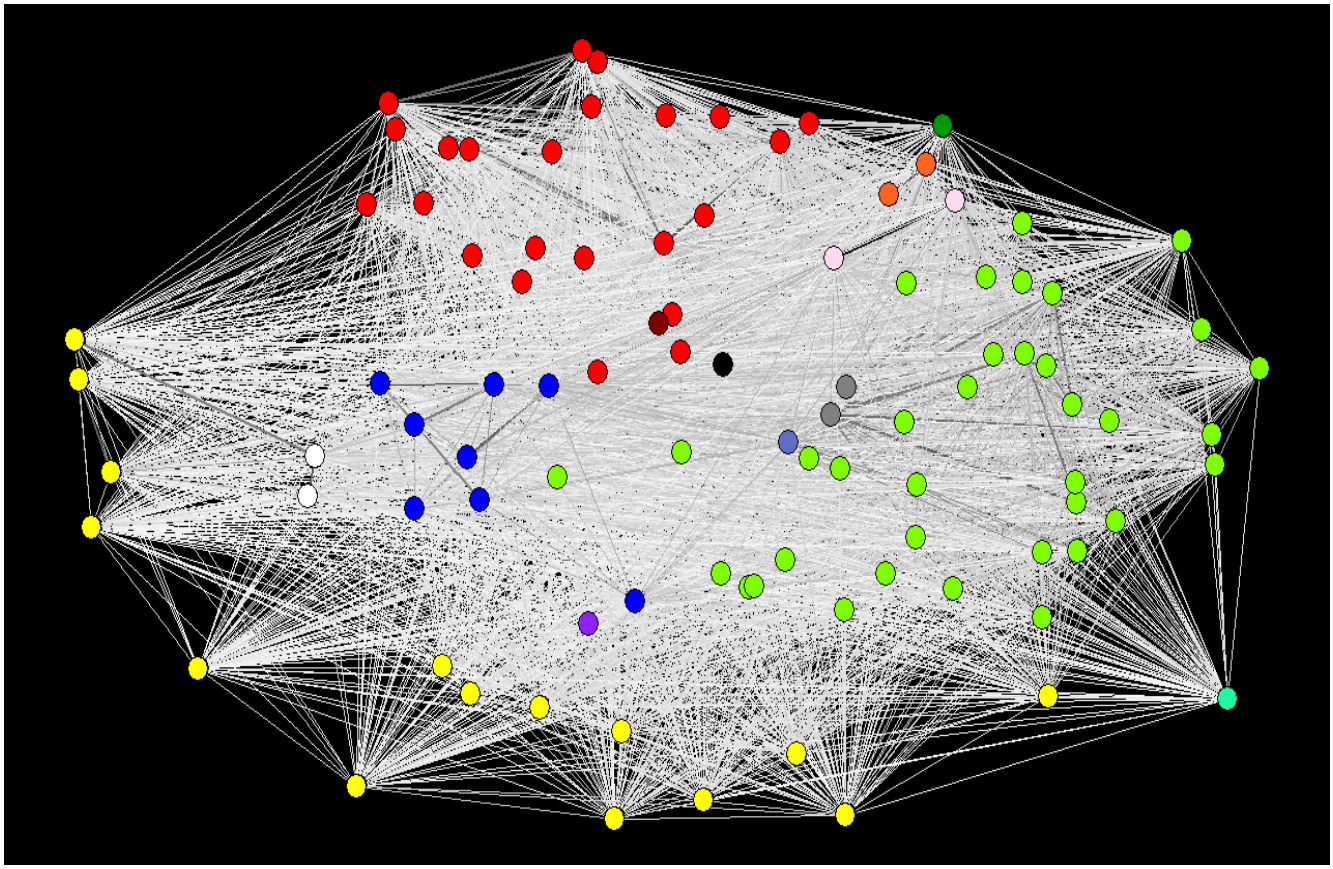
(3) Clustering of both UUNs = **block ordering**



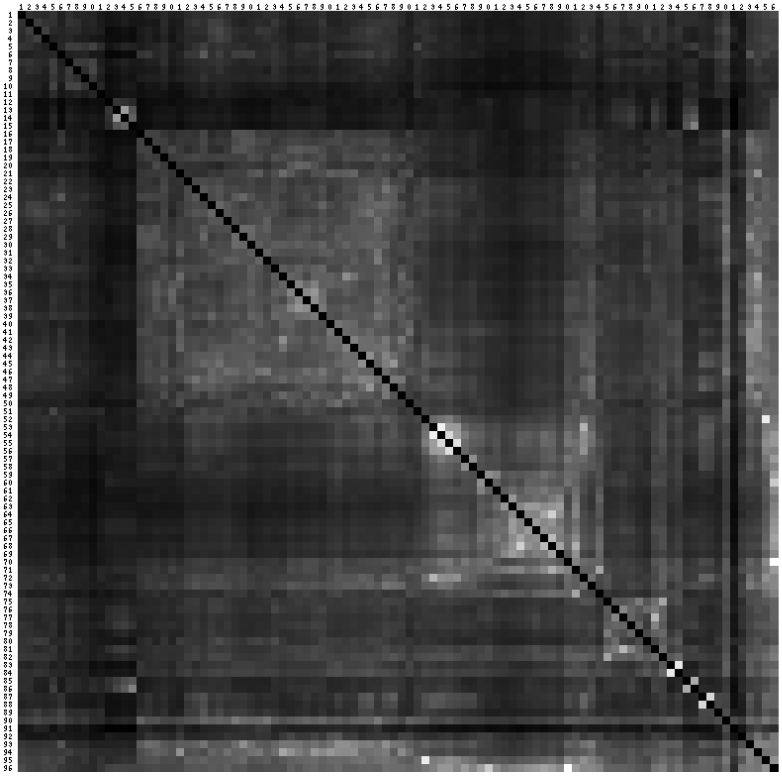
(4) **Network of clusters** with new weights given by red areas

Networks of Clusters

Lymphoma Tumor Tissue Samples



Lymphoma Tumor Tissue Samples



Adjacency matrix
(dark = low weight)

96 tumor samples

Showing their mutual similarity with respect to 4026 gene log expression levels.

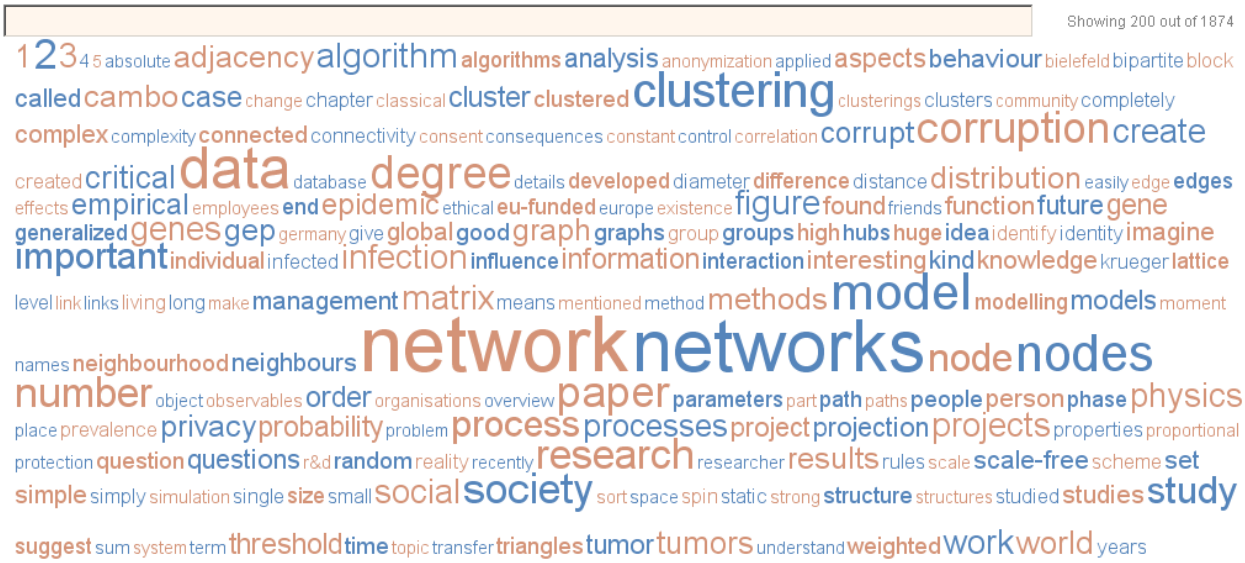
Weighted (+,+) Projection from 96x4026 data table

Best clustering so far with

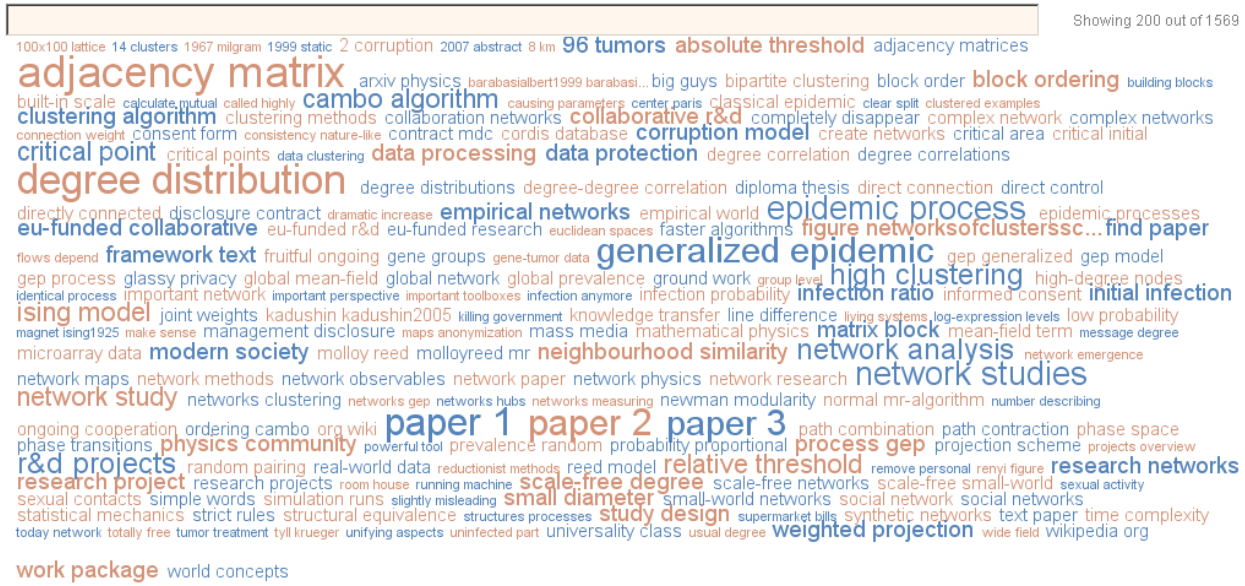
numCl=14
modularity=0.06654

at b=9 c=5
Mean weight a=0.22

Rahmentext: Die häufigsten 1-Wörter



Rahmentext: Die häufigsten 2-Wörter



Structures, Processes, and Clustering of Complex Networks

Andreas Krüger

6.2.2008

Dissertation an der Fakultät für Physik der Uni Bielefeld vom 31.10.2007
→ http://bieson.ub.uni-bielefeld.de/volltexte/2008/1247/ ←