Structures, processes, and clustering of Complex Networks
Doktorarbeit in Physik
31.10.2007

Andreas Krueger*

University of Bielefeld, Mathematical Physics

*Electronic address: akrueger@physik.uni-bielefeld.de,networks@AndreasKrueger.de

Contents

A thousand years old mycel
shows itself by the yearly mushrooms[1].

1. INTRODUCTION

1.1. Where in science are we?

Networks (or **graphs**) are based in mathematical Graph Theory [21], they are a widely used tool of complexity science [45]. Networks are built out of the *object* **entity**[2] and the *verb* **connect**. We **connect** 2 **entities** if they have *something in common*, something that we want to study, then 2 connected entities are said to have a *link (edge, tie, ...).*

The set of the links contains *details* and the *macro* structures of connectivity, so the old saying "*You do not see the forest because of all the trees*" is ideally changed to "*Using networks, we begin to see the forest **and** the trees*".

Statistical Mechanics, developed for the physical objects domain for 150 years, can be understood as one of the most important *toolboxes* of theoretical physics, and provides us with theory and methods to describe many-body-problems with many>>1, and can give a foundation for Thermodynamics.

The tools of *StatMech* are increasingly used outside the physical domain. The wide field of *econophysics* [18] ist a good example for the transfer of *StatMech*-models, -theories, -concepts, and -solutions to the domain of the capitalist market; and for example in 2002, there was a conference about "SocioPhysics" in Bielefeld [40], with applications of physics methods to a wider range of social questions than the ubiquitous hunt for money.

A standard *StatMech*-model is the **Ising model** [25], a simple dynamic model with phase transitions and a "temperature" parameter to steer from overcritical (unordered) over critical (long-range alignment, fractals) to undercritical (frozen) state, originally developed to model a magnet [23]. This spin± model now even inspires models in sociology, e.g. to describe innovation-islands and -propagation [13].

To model the short range of the spin interaction, the spins in the Ising model are "living" fixed on a regular d-dimensional lattice $\subset \mathbb{Z}^d$, and interaction happens only between nearest neighbours $i \sim j$ on the lattice. A lattice is a graph, e.g. the 2 dim rectangular lattice is an infinite regular graph with degree = 4 (number of neighbours of each spin).

Now imagine the neighbourhood-quality of this lattice to become totally free of location or distance, so each spin *can* have an interaction with *any* of all the other spins, and your Ising model is now living on a complex network. The interaction J_{ij} can now be different from zero for *any* two spins i and j, and J_{ij} could now be called *adjacency matrix* of the spin network.

The huge methodology of Statistical Mechanics could be successfully applied to lattice models, and the translation to complex networks is on the way.

1.2. Motivation for the choice of this field

Networks are a mathematical abstraction from reality that can be applied to a huge range of object classes, in simple words: To anything that allows a identification of elements and carries a

[1] The biggest living being found on earth is a network, a 880 Hektar (8.8 km^2) sized Hallimasch (Armillaria) funghi mycel, with the calculated age of 2400 years a mass of about 600 tons. http://de.wikipedia.org/wiki/Halimasch

[2] entity = object, thing, countable. "An entity is something that has a distinct, separate existence, though it need not be a material existence." http://en.wikipedia.org/wiki/Entity

connectivity concept.

The studies in networks physics of the past 10 years show an impressing range from the very small (protein-protein interaction in humans, gene regulation networks) to the very large (World-WideWeb), from machinery (hubs on the internet, sourcecode of mySQL) to living systems (food webs, sexual contacts networks, boards of CEOs), from materialistic facts (planes and airports, electricity networks) to pure fiction (figures in the theather play Les miserables, and in Marvel comics) [6], [16], [34], [20].

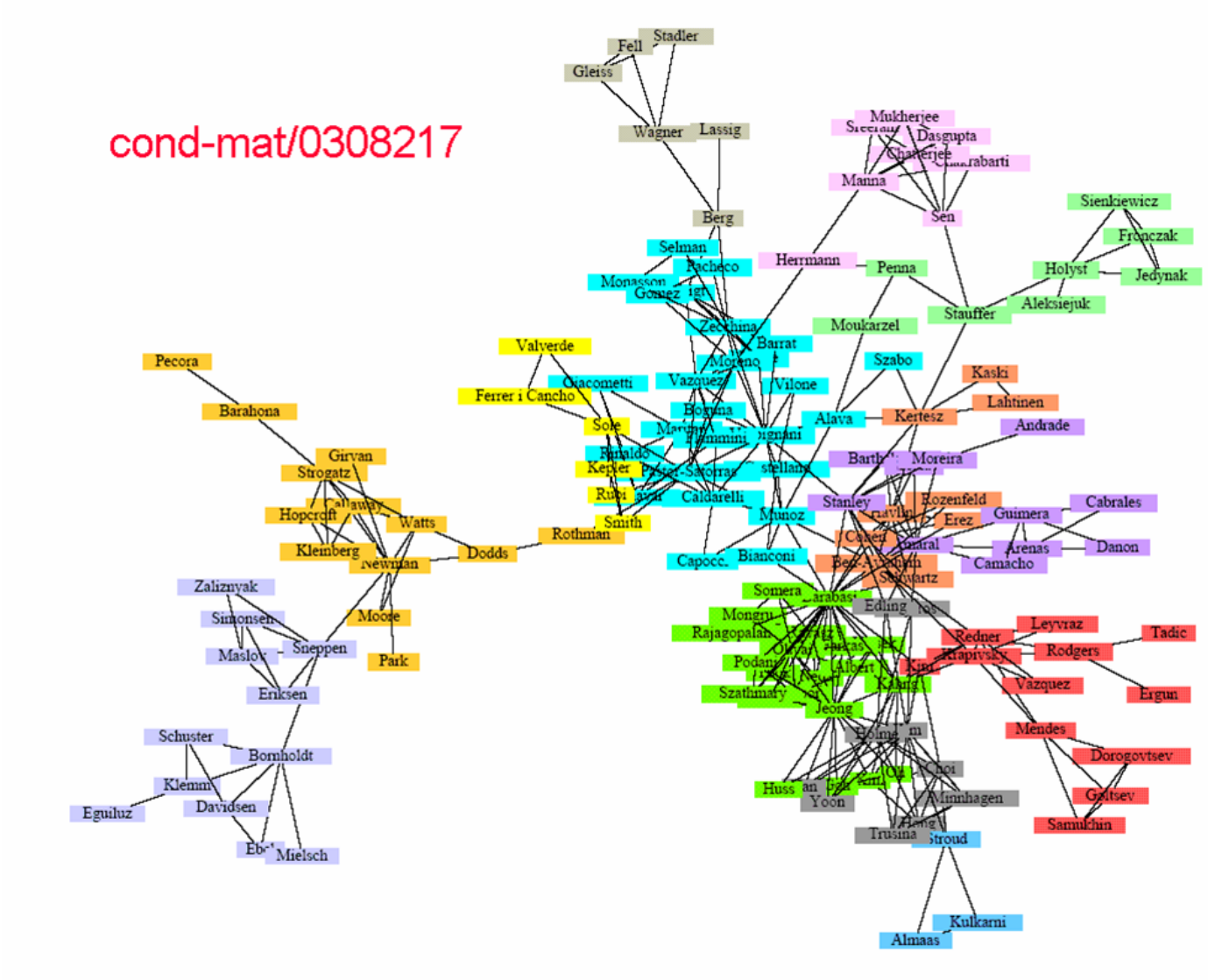


FIG. 1: The communities in the co-authorship network of the new field of network physics after 5 years [34].

1.3. History of and overview about this work

In this dissertation I have studied several important basic aspects of networks. The class of empirical networks that started the "hype" in physics in 1998 and 1999 can be described as

scale-free, small-world networks, so my first experiments started with the programming of an Albert-Barabasi-Model, in which "preferential attachment" creates new links to existing nodes with a probability proportional to the degree.

When the Mathematical Physics in Bielefeld were asked to help the ARC in Vienna with the data analysis of the CORDIS database, the first question was: Are these networks such of our kind? When we actually **found these three properties** in the project-project networks and the organisation-organisation-networks of EU-funded R&D projects, we were optimistic to be able to help these economists. A fruitful, ongoing cooperation started, which recently succeeded in getting the EU-funded research project "NEMO", in which we, the University of Bielefeld, are an important actor [32].

Paper 1 contains the first phase of our collaborative work on that topic, a study of **global network observables**, and a simple **first model** to reproduce the scale-free degree findings by a static random pairing process with given degree distribution, not by an iterative growth process like in the BA model. We found that a lot of information is stored in the degree distribution; with most of the network observables at hand at that time, we could not easily distinguish between the empirical and the synthetic networks. Only the excess multiplicity was higher in the *empirical Europe*, so here was a first finding that there are stable groups of actors who are collaborating more often than our random set graph model would create.

Paper 2 covers a second important network topic, on a completely different level: **Processes on networks.** Given is a *static* network, so that all the links between entities remain constant. Variable now becomes a function on the nodes, so each node carries some number, in our easiest case a binary number, describing the corruption state of this network node, comparable to the spin state of an Ising model.

In my diploma thesis [26], I had studied continuum percolation. What I have implemented here in *paper 2*, is a kind of percolating process ("When does the *whole* population suddenly become corrupt?"), but with a more complex infection process - and a richer network than the overlapping spheres in Euclidean spaces, or lattices of standard percolation, both extensions with the aim of getting closer to realistic models for complex society and nature.

We name these processes GEP generalized epidemic processes, they contain a usual classical epidemic process term, but additionally a cleaning- and a mean-field term, and most importantly: a jump-function to a strong infection probability, once the number of infected neighbours exceeds a threshold - which is the reason why we call this GEP model a corruption model, because it tries to embrace the most important epidemic aspects of the hard-to-study social disease corruption.

The future of our GEP models will probably be to represent a vehicle for transporting the different aspects of *knowledge transfer in research*, so we will create an adapted GEP model for the NEMO project. At the end of that project we hope to be able to create networks in a growth-GEP coupled manner, so that the GEP state of the network influences the growth of additional links in the network. *Paper 2* is the ground work for that future work, because at first we had to understand better the properties of the GEP process, before it may influence the change of the underlying network.

Paper 3 describes a new clustering algorithm. It is the first output of a gene-tumor analysis cooperation with the University of Marseille [14]. The idea is to transfer our organisations-and-projects methods to a gene-and-tumor network, in which microarray data gives a hint how often a gene is switched on with a certain tumor. As we had never worked on weighted networks before, some ground work for weighted projections had to be done, then the method actually already started to shine light into the gene-gene network and the tumor-tumor network. The work will be continued soon, then with clustering and bipartite clustering of the genes and tumors, which will hopefully help for better tumor treatment and tumor prevention by means of genetic identification.

Clustering was thus the most important next topic. For example, in our 96 tumors, or in our

>20000 EU research organisations, there are nodes that are more strongly connected to each other than to the whole rest of the network. In most networks, groups of nodes have this property of link concentration to "within" and link sparsity to "outside", and this gives rise to the hope that networks can be *clustered* into such partitions. This clustering (partitioning, dividing) sorts similiar nodes together into one group, for example the 96 tumors seem to fall into 14 genetically different tumor types, so we have *14 clusters of tumors*.

There are dozens of clustering methods available now, some using relaxation of Potts models on the network [39], others the elimination of most central actors to cut between the clusters [19], and a third class of methods use the eigenvalues of the adjacency matrix, to sort into the subspaces of the space spanned by the eigenvectors of the adjacency matrix [36].

While I was working on the gene and tumor projections to create networks out of microarray data, I realized the strength of the matrix representation of networks, and found an idea how to cluster networks in a way that has not yet been reported to the physics community of network scientists. I decided to start the project of creating a proof-of-concept implementation of that idea: *to sort the adjacency matrix into a block structure*. Until this program now works to my content, it became a long task of solving numerous obstacles, and most of all, deciding among many different possible paths.

This new clustering algorithm is presented with clustered examples, the sourcecode will be published soon. Compared solely on the time complexity level, the CAMBO algorithm is not worth mentioning, because there are already several faster algorithms than $O(N^3)$, however the results are good and relatively easy to interpret, and the algorithm has tuning parameters for most important clustering aspects like edges, triangles and structural equivalence. I suggest to cluster smaller networks with this algorithm, up to ~500 nodes is no problem.

So the millions of nodes that are clustered by faster algorithms are unatainable due to the $O(N^3)$ of the strategy. There are still some aspects of the CAMBO algorithm that can be improved, and the approach of sorting the adjacency matrix might quite as well inspire other, faster ideas.

1.3.1. The order of this dissertation

This introduction
Technical terms of graphs and networks
The beginning of this area of physics.
About paper 1: *The Network of EU-Funded Collaborative R&D Projects*
About paper 2: *Corruption as a generalized epidemic process (GEP)*
About paper 3: *Clustering by Adjacency Matrix Block Ordering (CAMBO)*
Concluding remarks
Papers 1, 2 and 3

1.4. Graphs / Networks

Let $G = (V, E)$ be graph with $N \in Z^+$ vertices $V \subseteq Z^+$, and $M \in Z^+$ edges; the edge set in the *directed case* is $E \subseteq V \times V$, and in the *undirected* case $E \subseteq \binom{V}{2}$. See figure 2 for an example of an unweighted graph. The cardinality of the edges set E is bound by the full graph $M \leq \frac{N(N-1)}{2}$. A *weighted* graph (figure 3) carries an edge function $w : E \to \mathbb{R}$, $w(e) = w(v_1, v_2)$ with a real number $w(e)$ for each edge, which can e.g. represent a connection or co-ocurrence strength between node v_1 and v_2 (v_i and v_j sometimes simply called nodes i and j). The unweighted case is included, with $w(e) \in \{0, 1\}, \forall e$.

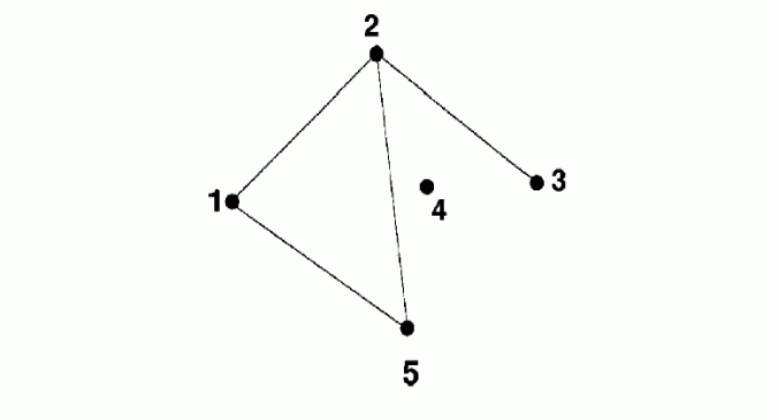
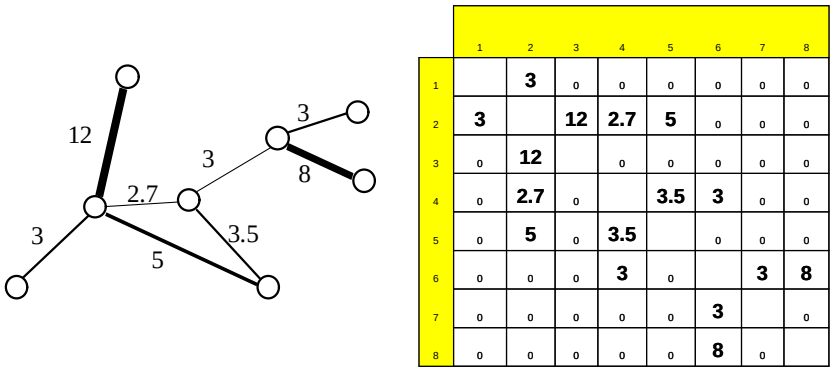


FIG. 2: A drawing of the undirected unweighted graph G=(V,E), V={1,2,3,4,5}, E={(1,2), (1,5), (2,3), (2,5)} with N=5 nodes (individuals, actors, vertices, points, ...) and M=4 edges (links, bonds, ties, lines, connections, ...).



	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	

FIG. 3: Example visualization of a **weighted** graph and its **adjacency matrix.** As the graph is undirected, the matrix is symmetric. The diagonal would mean "self-loops" of one node back to itself, and is seldom used, i.e. set to zero.

1.4.1. Graph measures

The **pathlength** between node i and j is the geodesic, or shortest path between them, i.e. the minimal number of edges to traverse from node i to get to node j. Two global network measures can be created from that: The maximal pathlength in a network is called **diameter**, and the **average pathlength** can be calculated from all (i, j) combinations.

The **degree** $\mathbf{k}_x$ of a node x is the cardinality of its N_1-neighbourhood: How many neighbours has node x? A **degree distribution** is plotted with degree k on the x-axis, and the frequency of that degree $P(x)$ on the y-axis (figure 4).

The **triangle number** of a node is the number of triangles $\#T$ this node is in; C_i compares this to the *possible* maximal number of triangles (figure 5), and C is a global average, called **cluster coefficient**:

$$C_i = \frac{\#T}{k_i(k_i-1)/2} \qquad C = \frac{1}{N}\sum_i C_i \tag{1}$$

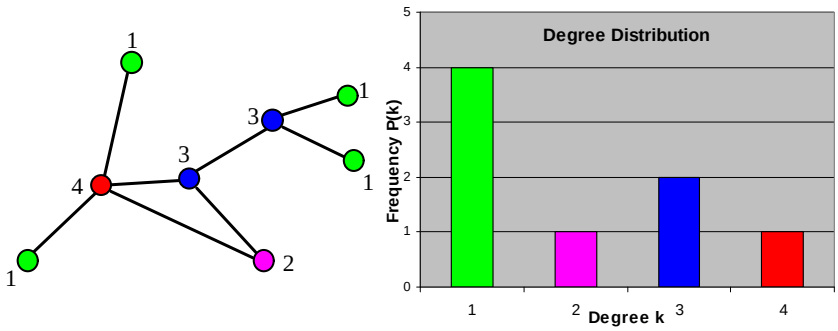


FIG. 4: example for **degree** and **degree distribution**

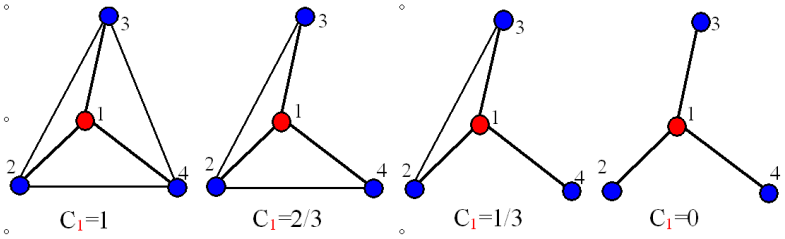


FIG. 5: example for a decreasing *clustering coefficient* (eq. 1)

1.4.2. Erdösz-Renyi (ER) Random Graphs (RGs)

Erdös and Renyi (figure 6) invented Random Graph Models in the 1960s, and they became quasi-standard until the advent of the new network physics around the millennium. Their G(N,p) model with a number of edges between 0 and M_{max}=N(N-1)/2, creates all edges with independent probability p (Bernoulli process).

The degree has a *binomial distribution*, for $p << 1$ and $N >> 1$ *Poissonian* $P(k) = e^{-\mu}\frac{\mu^k}{k!}$ with a mean degree of $\mu = \langle k \rangle = (N-1)p = (N-1)\frac{M}{N(N-1)/2} = 2\frac{M}{N}$

The exponential tail for large k does not allow for very large k far away from $\langle k \rangle$. The distribution is bell-shaped (see figure 7), so $\langle k \rangle$ really makes sense as an "average", because there is a built-in *scale*.

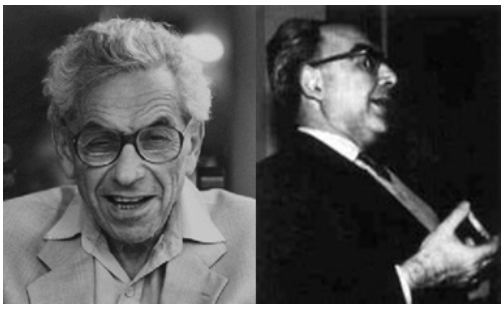


FIG. 6: Portraits of Paul Erdös and Alfréd Rényi, taken from http://www.nd.edu/~networks/linked/newfile6.htm

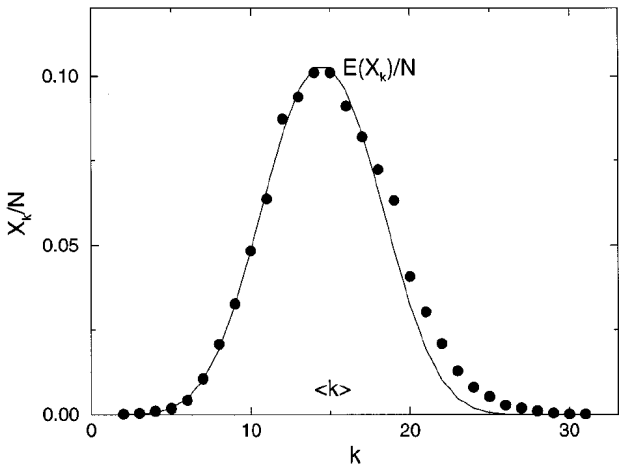


FIG. 7: The degree distribution that results from the numerical simulation of a random graph. We generated a single random graph with $N = 10\ 000$ nodes and connection probability $p = 0.0015$, and calculated the number of nodes with degree k, X_k. The plot compares X_k/N with the expectation value of the Poisson $P(X_k = r) = e^{-\lambda_k}\frac{\lambda_k^r}{r!}$ distribution, $E(X_k)/N = P(k_i = k)$, and we can see that the deviation is small. [figure and caption taken from [6]]

1.4.3. Paradigm shift 1998/1999: static → grown networks

Measuring the connectivity of 300,000 webpages, Barabasi, Albert and Jeong in 1999 discovered a completely different degree statistics [4], [5], shown in figure 8. Apart from finite size effects there is no built-in scale, there are so-called "hubs" of any size ("hubs" have many more connections than other nodes). (Up to $k \sim N^{1/\lambda}$) the degree distribution has a *fat tail* that falls like $P(k) \sim k^{-\lambda}$ with $\lambda \cong 2.5 \pm 1$ for many very different networks. And important to see, the empirical data does not only show a *tendency* to follow that rule and is scattered all around that theoretical curve, but in this case, the empirics closely follow the curve in a law-like manner.

Since then, many networks of that kind have been discovered in the empirical world, and it looks as if there is a new universality class, *the universality class of scale-free, small-world networks*. While the scale-free property was introduced by Barabasi and Albert as mentioned above, Watts and Strogatz [44] called the other two important properties of empirical networks (slightly misleading) "*small-world*" = *high clustering* and *small diameter*.

Small diameter means that the diameter is approximately a logarithmic function of the number of nodes, because there are at least some "short-cuts" between remote parts of the network. As a counterexample, a 2dim 100x100 lattice is no small-world, because the longest path between two nodes is 198, while a usual complex network with as many nodes (N=10000) might only have a diameter of 4 or 5. The 1967 Milgram experiment [31] coined the urban legend term "Six Degrees of Separation" to express that the mean number of connecting steps between any 2 people in the US was estimated to be 6, by a mail experiment with 100 letters, of which about 20 letters returned - and these 20 only needed such a low number of hops to get from the source to the target person without using the mail system.

Networks with a high *clustering coefficient* are called *highly clustered*, so there are much more triangles (and cliques of higher order) than in a purely random (e.g. ER-) model; friends of mine are much more likely to be friends among themselves than if friendship were not a social glue of that kind. High clustering has a lot of influence, for example in the corruption process we studied

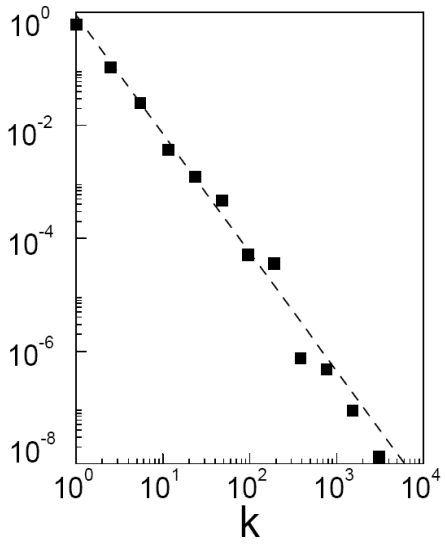


FIG. 8: The degree statistics (frequency P(k) over degree k) of the N=325729 WWW-pages that were examined by Barabasi, Albert and Jeong in 1999. Note the double-logarithmic scale. The *data* is very close to the *fit* $P(k) \sim k^{-2.1}$ (figure taken from [4]).

FIG. 9: Albert, Barabasi, Jeong, Watts and Strogatz started the *network hype* in physics (pictures from the visual companion of *Linked [7]*)

in the second paper.

Many papers have been published in this past decade, many observables have been added to the three mentioned above, nowadays there are many ways not only to see the unifying aspects of all found networks, but also means to distinguish between similiar networks. One interesting example being the (dis)assortativity of the degree-degree-correlation. It was found [33] that social and technical/natural networks might have the same degree distribution, but their *degree correlation* differs. Social networks are often *assortative*, the "big guys" (hubs, high degree) tend to collaborate more with other "big guys" - while in natural or technical networks, hubs are more often connected to nodes with a much smaller degree (disassortative mixing). We study the effects of such a degree correlation difference in the corruption simulation (paper 2).

A very interesting network *observable*, already described earlier by Bollobas [30], has just been rediscovered by Bagrow *etal* [3]; we are now e.g. able to identify graph *isomorphs* (permutations of node labels) more easily. Bagrows *portraits of networks* look at the number B_{lk} of nodes which have a N_k neighbourhood with exact size l, with $k = 1$ containing the usual degree distribution.

They created a binary relation $\Delta(B, B')$ between 2 of these B_{lk} to compare 2 networks, the result is a positive number.

We now have another good observable to compare any two networks with each other, which will among (other new possibilities) enable us to create much better models (e.g. for the EU research network), because now we will really see the differences between the empirical and synthetical world.

2. ETHICAL CONSIDERATIONS

No doubt, the chances of network analysis are huge. World concepts still very often lack truely empirical groundings, and network analysis will help a lot to extract important insights from all the data that is there (for example see figure 14). And world concepts also often lack the important connectivity aspect: We might have understood isolated concepts of anything to an astounding degree, but putting all the reductionist details together into a bigger picture is a hot topic of this "century of complexity" [1]. Solving all the life-threatening environmental problems that modern society has created will force mankind to big advances in sufficiency ("rich is who does not need much"), efficiency ("more gain out of less ressources") and consistency ("nature-like cycles") - and especially in this third and hardest dimension of change, network models of real-world data might prove to be very helpful to create new and optimize all existing systems.

(As usual) such huge chances are accompagnied by rather worrying pitfalls. This chapter wants to open up a few of the *ethical* questions, that arise when your "nodes" are actually human individuals.

2.1. Ethics of Network Studies

As an introduction into this avoided topic, we give a short overview about the still few papers in the SNA (Social Network Analysis) community [8] [9] [10] [24] [46]. Essentially, this chapter is a summary of the posed questions and given answers in these 5 papers.

Most classical aggregated questionaires (AQs) can be analyzed successfully without knowing who the respondents are. Network studies are different, here we need the knowledge about the **identity** of a person, and their local neighborhood in order to construct the network ("name three friends"). *Anonymity of response* is just not possible, when we ask network questions. *Anonymization* can be offered, though - but in the first place, this is a promise, and needs proper study design, data handling, and results reports.

Network surveys are relatively new, so many respondents might still unable to foresee the consequences of their answers, while they usually already have an intuition for classical AQ surveys and possible consequences of disclosing personal information.

As a consequence of this naiveity, Borgatti and Molina call the present the *golden age of social network research*: "How many network audits of this kind can be done before employees learn to fill out the forms strategically?" [8].

2.1.1. Protecting the data

The original data should remain under direct control of the researcher, who might want to use data encryption (no one else can use it), physical separation from the internet (hackers cannot

[1] Stephen Hawking on January 23, 2000

access it), and probably most important: a codebook.

Anonymization requires to remove personal identifiers as soon as possible (e.g. replace all names with id-codes), the use of codebooks (linking files) allows later to recover the identity, e.g. for longitudinal studies. If not even the researcher should be able to link back to the names, a third party can be asked to hold the only codebook. In cases where the data are interesting to the government (subpoena power), it is suggested that such files are stored physically beyond the control of the researchers; outside legal jurisdiction, e.g. in a foreign country.

One should not forget that network study results also allow to *reconstruct the identity*. Information on each person is available in the company's database or is common knowledge. If e.g. there is only one male of age 35-44 within a group of nurses, it is easy to beat all struggles of anonymization.

2.1.2. Organizational Research

The hierarchical structure of organizations poses a challenge to network study design (or to the decision to completely turn down the offer of a study). Without anonymity possible, the employee's participation is more risky than in other situations. Management will see the results; and employees might even get laid-off, if that is not forbidden by the study design. If the researcher hands over the data to management, he cannot simply claim no responsibility over how it is used. Rather management needs to be a party to the consent form. Suggested are two contracts: the truly informed consent (TIC) form, signed by every respondent, and the management disclosure contract (MDC), signed by the organization, and preferably included in the TIC.

Most important in the TIC is honesty. Without anonymity, confidentiality can in principle never be guaranteed, so consent forms should not mislead respondents. They should be given a sample network map so that they understand the very data processing, and they should be promised that these maps will not be shown to others than the management. And whenever possible, management will only see results aggregated to the group level. And: What kinds of conclusions and what consequences might emerge from the study?

The management disclosure contract (MDC) should contain a guarantee that the data will not be used to evaluate individual employees, but will be used in a way to improve the company as a whole. If the data is handed over, and an employee laid-off by using that data, this could be seen as unethical behaviour of the researcher; but if this was forbidden by the MDC, then it would be the company which commited that breach.

Borgatti and Molina [9] give samples for each a TIC and a MDC.

Even if a TIC should allow each employee to decide to opt-out of the study, here is a true dilemma: Opt-out is a special problem in network studies, for if a central person or a bridge between two groups decides to opt-out, eliminating this non-respondent reduces the validity of the the whole analysis - also an ethical problem if a true representation of the network is aimed at, which is the case, normally.

A special question of consenting participants arises in the standard "name three friends" question. While the respondent herself knows that she answers, the named three friends might be out of reach to be contacted if they consent to participate in a network study.

2.1.3. Who benefits

A position taken by Kadushin [24] is that the investigators themselves are the prime beneficiaries (publication, reputation, ...). Questioning who benefits from a network study, might lead to insights that improve the ethical grounds on which the whole project is undertaken. One reason for science,

humanity as a whole should be the beneficiary - e.g. containing a disease like HIV by means of better modelling of the epidemic network. A valid research methodology has to be chosen to ensure research results are of genuine benefit to society.

Probably easiest to think of is the organizational benefit, if the study reveals structures and processes that were previously unknown - and the company embraces them bravely, and includes this sometimes disturbing knowledge into the everyday work. If the network insights are shared with the employees, their workplace can become a better place. If, though, the benefits are only on the side of the management, the ethical situation is to be re-evaluated.

2.1.4. Results presentation

Several aspects of anonymized result presentation have already been mentioned and will not be repeated here. In network maps, anonymization by grouping of nodes is possible by e.g. labelling the nodes by their properties (e.g. department or office in the organization) instead of labelling each node with its true identity.

If people are shown in network maps, or ranked by e.g. centrality indices, they might feel offended by their position, is one more thing to consider.

Instead of or in addition to a payment, respondents should get some feedback, ideally something tailored especially for them, which is at the moment still hard to realize, because of the lack of appriopriate network software.

2.1.5. Dangerous information

Just imagine network studies on HIV, sexual activity, social connections, illegal activities, etc., and you can instantly create examples of worrying situations. Some people do not want to know their HIV serostatus, and a disclosure of test results can have adverse psychological, social, financial, or legal consequences.

What would *you* do, if (in your networks) you discover an HIV-negative couple, the woman is pregnant, and all of the sexual contacts of the husband (that pretends to be faithful) are HIV positive? Talk or not talk? To whom?

Being considered to be a terrorist, is another example of today. Network studies can probably identify subgroups (of e.g. telephone users) that get into clandestine behaviour, but sharing this information with a killing government is another question.

These are extreme examples, but in general, non-anonymous studies pose the question what to do with any information that might have (negative or positive) consequences for the individual. Even if the data itself was already in public (and not just collected from consenting individuals), the network analysis makes visible that which is not apparent to the "naked eye", and the person might never have expected that her answer would be used for this kind of analysis.

2.1.6. Your powerful tools and your decisions

With network analysis methods, you as a researcher have yet another powerful tool in your hands, and thus need to use it responsibly. I would like to close this short chapter with a quote from Kadushin [24], which sums it all up in a few lines:

> "But there are conditions upon which I will not compromise: the data are always under my direct control, must be collected under guidelines that I describe, must reside on my computers,

> as do the names associated with the data. Confidentiality is always guaranteed. The data are never the property of the firm for whom I am a consultant. Names are never associated with network graphs or with network indices and are never revealed to either management or employees. Rather, general patterns are described and used to suggest they way things currently flow and how matters might be changed. "Things" as flows depend on the purpose of the investigation—communication, prestige, authority, and even friendship. If the organization cannot meet these conditions, then they must look elsewhere for someone to carry out their investigation. Typically, these conditions cannot be met by classified or military research and so I do not do this kind of work."
>
> Kadushin, Social Networks 27 (2005) 139-153

2.2. Glassy Privacy

Those were rather concrete and study-related questions. Also connected with ethics of network research are questions of more philosophical or more political colour. We enter a stage in human evolution, in which not only an abundance of facts about humanity is measured and stored as data on computers, but also new and powerful methods are developed that allow to explore this massive dataset with by all means concrete results. This conflicts with the whole construction of privacy[2]. We explore some of the related thoughts in a more essayist manner:

2.2.1. Privacy and the state

When the modern state of nowadays was excogitated, originally a protection of privacy of the single citizen was included in the constitution to protect the single individual from some adverse effects of the herding - back then there was a clear split between "public" and "private". But recently, these guarantees have been removed bit by bit (for example, in Germany the constitutional "Briefgeheimnis" = privacy-of-correspondence was virtually removed in 2001 - researchers should now also consider this as part of their ethics questions, whenever study data is sent by unencrypted email!). There *are* laws of data protection, but many of them exist only in theory or idealism. The state is exclusively powerful (some states even still kill their citizens) and thus the protection of the individual *against* the state *is* especially important, but the data stored in *companies* all around this globalized world will even outperform the nations' data by several orders of magnitude. The ineffectiveness of laws for data protection is obvious.

Data protection politicians or jurisdiction might suggest that in order to protect privacy, we should refrain from developing or applying new methods, or from measuring the world around us; the necessary self-restriction of mankind has many times proven to be rather weak, though. And network methods are mostly simple applications of basic arithmetics so they can be developed even without much education or expensive machinery.

Data-wise, the single person will very soon be a trace-able collection of RFIDs (radio frequency identification), because products are planned to have and are already manufactured with such RFID-chips. A certain *combination* of products (including e.g. clothes) might be *unique to your person* - and thus the above questions of lost anonymity in network studies applies to our very existence in cities. Whoever reads the RFID chips (and they contain no security system but will

[2] For a nicely done introduction what is already possible, have a look at this Flash presentation. A Bachelor Thesis done in Ulm, 2006 (in German language): `http://www.spiegel.de/flash/0,5532,15385,00.html`

readily answer any question posed by any RFID detector) can then *identify you* - and e.g. trace your way through a surveilled area.

2.2.2. Possible futures without privacy

Extrapolating the development of methods and the data abundance into the future, it looks as if *privacy might completely disappear*, or perhaps becomes something that you can sometimes afford to pay for[3]?

For the rest of this chapter please imagine that this is actually possible. I suggest a *Gedanken-experiment*: *Imagine your world without any privacy.* Even if you cannot think of privacy to completely disappear from your world, this extreme standpoint might shed some light onto central questions.

One possible future: "legalized outlaws".

Not every individual will readily accept the glassiness of his own existence; on the other hand, participation in modern society automatically leaves data traces all over the place (supermarket bills, taxes, traffic lights, leaving and entering a room, house, quartier, city, state, ... etc.). So what will all the refusing people do? One might imagine non-surveilled areas left to "savages" like in Brave New World (Huxley 1932), who do not profit as much from modern society, but on the other hand are also not obliged to follow all the rules - and from this new viewpoint, are also not obliged to constantly leave about all their data. In converse, can you imagine a non-privacy world with a inevitable obligation to be part of it? Or are these questions only worrying to those who still experienced a strong private/public division in the past, and *all* younger people will easily be used to glassiness?

One other possible future: "fuzzier rules"

At the moment we live in a euphemistic, rule-based society that presents itself and the typical member in an idealistic way, and bound to strict rules - exceptions only exist if they are not detected. But what if the "true reality" bubbles up by massive measurement and data processing, and the empirics show a strong deviation from the ideal: Does the society then force all of its citizens to obey, or does it soften its rules accordingly?

One of the most human examples are "faithful relationships"; the majority in a developed country will probably pretend to be faithful (reality 1), but an estimated third acts differently (reality 2). If you are cool about *this* question, think of your own "perversions", for a second: In some of your behaviour, where do you deviate from the official ideal of your society, or from the first standard deviation of the population? And which aspects of that behaviour do you consider to happen only in your privacy?

In the old world *with* (guaranteed and technical) privacy, these competing realities could both exist at the same time (one open to the public, one hidden), but in a glassy society you cannot conceal anything for a really long time. So if that described development into a future without privacy (or "glassy privacy"), shall be a smooth transition, people should get more realistic and fault-tolerant towards each other, and the hard rules of society might become softer and fuzzier, alone because the alternative of massive behaviour change ("straighten the dishonest") sounds rather fascist.

---

[3] most toilets are not yet on video surveillance

2.2.3. The beginning already lies behind us

We would like to close these thoughts with a rather excentrical example, to show that deviation from the average of society is not automatically something inherently unethical; and if such a second reality is revealed, it may have simply bad results in a world of strict rules: Recently [22], an "illegal" private primary school in Northern Germany was detected that operated for 14 years, with more than 200 pupils. Now that school has been closed. Whatever the reason were for the teachers, parents and the pupils, to create and run an own school - a strict society does not seem to allow for that. And when data and data processing leads to detection of such "wrong" behaviour, diversity decreases.

This chapter contains first thoughts to open up this field, please contact us, if you are interested in discussing them deeper. In this case, science fiction might actually be a helpful vehicle to elaborate a bit on the *future* of our study object: The human existence. In the end, something like a honest TIC (truely informed consent) form will never be handed to us by any institution, but privacy will nevertheless gradually disappear. Probably the only thing we can realistically do is elucidate ourselves about the new situation.

3. PAPER 1: THE NETWORK OF EU-FUNDED COLLABORATIVE R&D PROJECTS

3.1. Overview

About 6% of research and development in Europe is funded by the EU. The EU commission publishes the EU-funded Research & Development projects in the CORDIS database on the WWW [12]. The Systems Research department of the Austrian Research Center (Vienna) downloaded the database and standardized the organisation names, cleaned up the data, broke meta-organisations into smaller parts, etc. - until a sufficient status of node identity was reached.

We were asked to help with modelling, so with the applied *set model* we created networks: Whenever the set of organisations of two projects are overlapping, so at least one organisation takes part in both projects, we created a link between these two projects, so we got the *projects network*. The dual brother of this unimodal projection from a bimodal network is to create the *organisations network*, by overlapping projects sets.

The degree distributions looked like any other of the network physics community (In the end, the exponents of the degree distribution tail turned out to be between 2.0 and 3.7). So we were "in the game", these CORDIS networks seemed to be part of that complex world that unfolds before our eyes in these years. A fruitful ongoing cooperation between economists and physicists began.

This important and realistic dataset became my first learning object for representing graphs in datastructures and algorithms. I programmed the Molloy Reed model (the easiest possible random pairing model with preservation of scale-freeness) in a bipartite version to "recreate Europe in the computer", with *only* the size distribution of projects and organisations kept from the empirical world. We found the synthetic networks to be highly similiar to the empirical networks in many respects, and soon understood that there is a need for better observables that look finer into the differences.

Recent developments [3] suggest new observables (*portraits of networks*), now the modelling can continue, e.g. we recently already saw that neither the additive nor the multiplicative Molloy Reed model is a good model for the EU networks.

In this first paper we studied several *global* features of the whole network, in the future, also the *mesocopic* scale, the clustering (partioning, dividing), has to be studied.

The Network of EU-Funded Collaborative R&D Projects
M. J. Barber, A. Krueger, T. Krueger, T. Roediger-Schluga
Phys. Rev. E 73, 036132 (2006) arXiv:physics/0509119v2

Abstract: We describe collaboration networks consisting of research projects funded by the European Union and the organizations involved in those projects. The networks are of substantial size and complexity, but are important to understand due to the significant impact they could have on research policies and national economies in the EU. In empirical determinations of the network properties, we observe characteristics similar to other collaboration networks, including scale-free degree distributions, small diameter, and high clustering. We present some plausible models for the formation and structure of networks with the observed properties.

You find paper 1 at the end of this framework text.

4. PAPER 2: CORRUPTION AS A GENERALIZED EPIDEMIC PROCESS (GEP)

4.1. Overview

With the MolloyReed (MR) model of paper 1, we saw that structural processes can generate networks with *scale-free degree* distribution, and of *small-diameter* type - however, simple (unimodal) MR cannot create a realistically high number of triangles, so the MR-networks are *not* automatically *highly-clustered* (With projections from bipartite graphs, though, we get a high clustering in the unimodal projections). Thus we can create semi-realistic networks, what is next?

What we now wanted to see was *processes* living on the nodes and edges of the network, with percolation phenomena no longer confined to Euclidean space. The far future hope is a network generating process that itself lives on the network, so with interaction of different network emergence levels; but first we need to understand the processural aspect, so in this paper, we kept the networks constant. A macroscopic jump over a critical point into a another domain can usually only be noticed after or during it is happening, so also if we want to improve our intuition for critical phenomena before they arise, we should study the microscopic processes themselves which lead to or hinder that macroscopic behaviour - these processes might even interact and resonate.

With the largest degree of freedom being the network structures itself, the *process* in our final model has a handful of free variables: Including Δ our phase space is 5-dimensional, and thereby so huge that we could only probe for single points in the phase diagram. It is however a difficult challenge to model social processes with such reductionist methods, in all our discussions about improving the model we always had to make *choices* for what we feel to be the best extension of the already running machine (in this respect it prepared for the work of paper 3, which was full of such decisions). Many ideas have never been implemented, others tried out and left as uninteresting, and the core that now represents our corruption-GEP-process is what we suggest to be a simple model for a complex social trait called corruption, using as few as possible parameters, and still allowing for a very complex and colourful behaviour.

4.1.1. The phase space dimensions of "corruption"

Let a network of nodes $x \in V$ be statically connected. Now we add a function ω, that returns a temporary state for each node x:

$$\omega(x,t) \in \{0,1\} \qquad \text{node } x \text{ is (not) corrupt at time } t \tag{2}$$

$$\Omega(x,t) = \sum_{y \sim x} \omega(y,t) \qquad \text{number of infected neighbours of node } x \tag{3}$$

$$\Pr_{\alpha,\epsilon,\Delta}(\omega_{t+1} = 1 | \omega_t = 0) = \left\{ \begin{array}{c} 0 \text{ for } \Omega = 0 \\ \epsilon \text{ for } 1 \leq \Omega < \Delta \\ \alpha \text{ for } \Omega \geq \Delta \end{array} \right\} \text{ with } \alpha \gg \epsilon \tag{4}$$

The new ingredient is the threshold function (4) for the infection itself, with Δ being the absolute number of neighbours that need to be infected to corrupt me - once my network N_1-neighbourhood corruption sum Ω counts at least Δ corrupt individuals, I become corrupt myself with a probability α. Beforehand, for less than Δ corrupt neighbours, I only have a very low probability ϵ to get infected by my neighbours. The idea behind that was that we all have a strong conditioning against corruption (which usually means breaking-the-rules of the society I live in), but if too many around me are already corrupt, I switch my inner state, and follow the criminals by imitation. This is a **local** process, see figure 10 for a visualization.

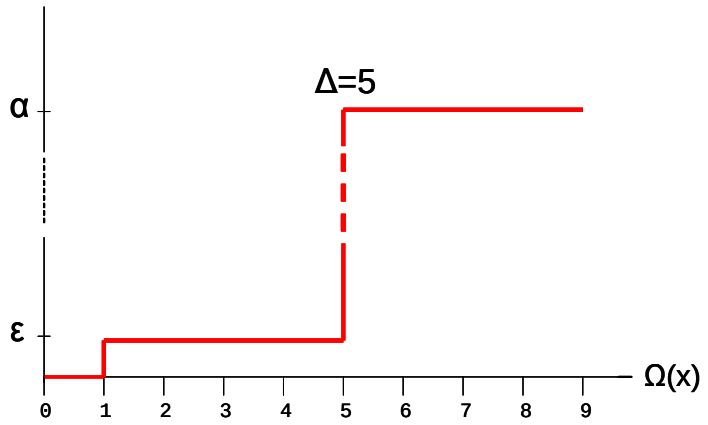


FIG. 10: A visualization of the local infection threshold probabilities α and ϵ ($\alpha >> \epsilon$). If the (absolute) number of corrupt neighbours $\Omega(x)$ of node x reaches $\Delta = 5$, the node x suddenly becomes much more susceptible for corruption.

The **global** "mean-field" terms use the probabilities β and γ, for infection and desinfection; because additional to the local infection by my neighbours, I can also get infected by mass media, atmosphere, the overall feeling how corrupt my society is - and that feeling is taken to be proportional to the prevalence of corruption ($|V|$ =number of nodes):

$$b_t = \frac{1}{|V|}\sum_{y\in V}\omega(y,t) = \text{prevalence of corruption} \tag{5}$$

$$\Pr_\beta(\omega_{t+1} = 1|\omega_t = 0) = \beta(b_t)(1-(1-b_t)) = \beta(b_t)^2 \tag{6}$$

$$\Pr_\gamma(\omega_{t+1} = 0|\omega_t = 1) = \gamma(1-b_t) \tag{7}$$

In (6), you see the proportionality (b_t) to the total prevalence, but you also need to see that the "fear" that has to be overcome is proportional to the number of still uninfected nodes $(1-b_t)$ (because the more of them, the higher the chance to get caught), and if I have to overcome my fear, that is $(1-(1-b_t))$. The combined effect makes the mean-field infection probability $\Pr_\beta$ proportional to the *square* of the prevalence.

Discovery and **Desinfection** can be visualized by taking the corrupt individual out of the system, and replacing him by an uninfected node. All his former connections to others are transfered to the new node, so the network structure itself stays constant. The desinfection term in our simulation is (7), so curing from corruption does not happen locally at all, but by some global institution, and is proportional to the uninfected part of society $(1-b_t)$, in other words: If there are no noncorrupt people left, nothing is done against corruption anymore.

The following table gives an overview over all 6 parameters, and some ideas how the terms can be interpreted:

Δ	Absolute threshold for local infection	for $\Omega < \Delta$, I believe that corruption is bad
α	Probability for local infection if $\Omega \geq \Delta$	Influenceability-by-others, -Decisiveness
ϵ	Low probability for local infection if $\Omega < \Delta$	Naiveity, criminal inclination
β	Probability of infection by global prevalence	"Random infection", belief in mass media
γ	Probability of des-infection by global prevalence	Random resistance / recovering / cleaning
b_{crit}	critical initial infection ratio	at which the whole society gets infected

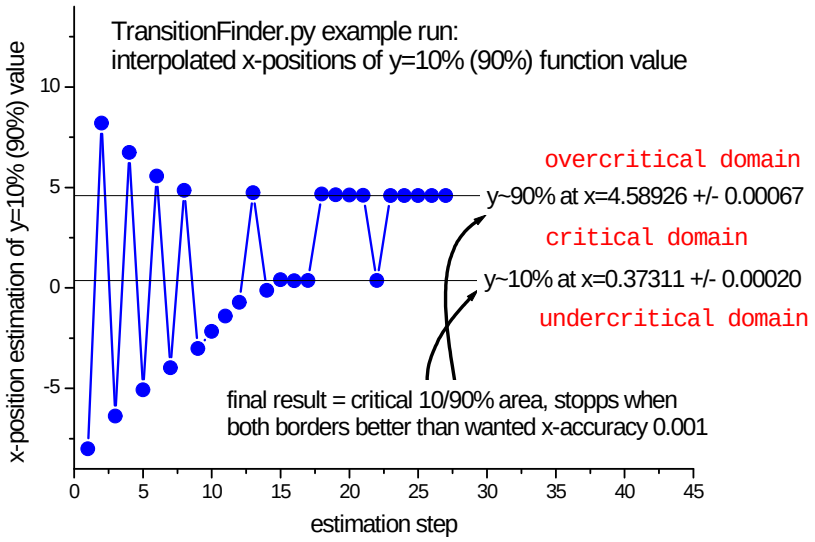


FIG. 11: My method for numerically finding the critical point avoids to simulate inside the critical area in which a single simulation has to be repeated many times due to the exploding variance.

Unlike classical epidemics, in our generalized epidemic model with threshold infection we actually observe a critical initial infection ratio b_{crit}, below which the society gets cleaned, and above which the society gets completely corrupt. This non-vanishing initial infection ratio complies with the unconcious "herd behaviour" that we often observe in society; once a certain group size believes in something (the "herd"), they influence all the others quickly.

(A provocative connection to the 1920ies in Germany could be drawn: The Nazis needed to persuade only a rather small group of size b_{crit} to believe in fascism and act accordingly, then afterwards the German society could not stop the complete infection anymore. So arguing in this model, the "masses" were as innocent and helpless as they always said after that terrible war.)

4.1.2. A new algorithm for estimating critical points

In statistical physics, we often study *critical* processes. At the critical point with its infinite correlation length, there is a dramatic increase in calculation time due to the huge variance of the results. In simple words, an almost-critical society needs a very very long simulation run to find out if it is over- or under-critical, much longer than a society far away from that critical point.

The experiencies of programming my diploma thesis (critical points of d-dim percolation, [26]) now became useful. I thought a lot about that computational problem, and came up with a good method to avoid these lengthy simulation runs as often as possible: The *position* of the critical point on the x-axis is usually more interesting than actual y-axis-*value*, so my 10/90 method tries to interpolate both the 10% and the 90% y-axis values by a linear function, and all the simulation runs take place *outside* the critical area (in which the variance explodes).
So we avoid the computationally expensive critical area by ideally never even touching it. The drawback is a slower convergence of the position estimation (see figure 11), the advantage is the much faster convergence of the single runs, because they are only simulated in the off-critical areas. This strategy was programmed as an abstract Python library, that can locate critical parameter

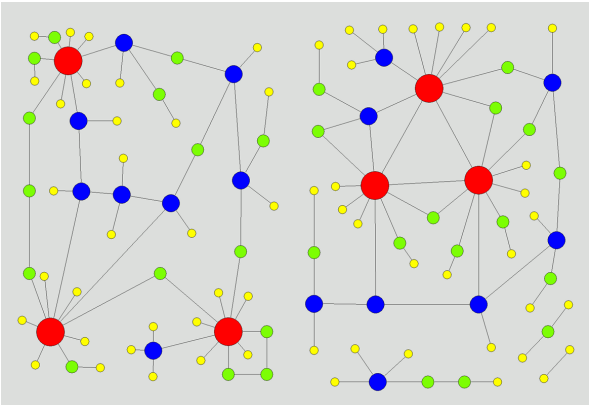


FIG. 12: Identical degree distribution, but (left) additive degree correlation, (right) multiplicative degree correlation. PLUS EVENTUELL averageRemnDegree plots dieser uniMag Netzwerke?

points of *any* function, once a <-relation exists. If you are interested in that library, please send me a message.

4.1.3. Degree correlations make a difference

We studied the described process on a variation of static networks: Erdösz-Renyi-Random-Graphs (ER-RGs), triangle-modified-ER-RGs (with a minimum number of triangles created), empirical EU-networks (of paper 1, see the chapter before), set graphs, and MolloyReed–(MR)-generated scale-free networks with two different degree correlations. The normal MR-algorithm creates multiplicative degree-degree-correlation, so high-degree nodes tend to get linked with other high-degree nodes. This resembles a stratified, dictatorial, centrical, hierarchical society - more like France (with the one center Paris, and an elite that went to school together) than like the federal Germany with its Bundesland-centers. In a variation of the normal MR-algorithm we chose a constant outdegree for all nodes, which results in a additive degree-correlation; now the hubs are more often not directly connected with each other, so a more polycentric or democratic society is modelled. Please see the two example societies in figure 12, their degree *distribution* is identical, but the degree-degree *correlation* is very different.

Figure 5 in paper 2 shows the difference of the two societies. With identical degree distribution, the multiplicative degree-correlation society is easier to infect by our corruption model.

4.1.4. Programming details, and plans

In order to avoid a path dependency of the update functions ("first node A is changed, then node B" - or "first node B is changed, then node A" would make a huge difference), we chose a "synchroneous update" by always working on the last copy of all the states $\omega(x,t)$. Only when all the updates for all the nodes have been done, the states are actualized. It is a little bit as if everyone decides on the basis of the knowledge of yesterday, and while everyone is at sleep overnight, the knowledge is updated synchroneously for everyone.

While developing this program, we had many more ideas, some of which we tried out (e.g. the relative threshold δ) some of which we put aside for later analysis (e.g. heterogeneous agents). To mention only some of the possible variations of the program:

- A *relative* threshold $\delta \leq \frac{\Omega(x)}{\deg(x)}$ instead of the absolute threshold $\Delta \leq \Omega(x)$, which however does not really make sense for e.g. scale-free networks, because hubs (with hundreds of neighbours, $deg(x) >> 1$) are almost immune against corruption, then. And as the relative threshold will enable small-degree nodes to get *very* easily infected, and they in turn will then infect the whole system, this *relative* threshold thereby overestimates the corruption infection. We went for the absolute threshold.

- At the moment all individuals (nodes) obey to identical process parameters $\alpha\beta\gamma\epsilon\Delta$, so the nodes are only distinguished on a network level (connectivity, degree). A more realistic model could work with individual parameters for each node, e.g. varying $\Delta = \Delta(x)$: For an weak and easy-to-influence person x' a lower $\Delta(x')$ threshold; and for a moral person x'' a high $\Delta(x'')$, because x'' is so blind for what is going on around him, that he only becomes corrupt if almost the whole neighbourhood is corrupt.

- We already thought about implementing a cleaning troup into the network, as nodes. At the moment, des-infection is only done by some global mean-field term. Local cleaning (I might get caught by some moralistic network neighbour) however, is not observed in any study about corruption, it just doesn't happen.

- A very interesting but still hard-to-grasp layer of reality would include "corrupt organizations" that create themselves in the network if a certain level of corruption prevalence is reached. Emergence of structure that becomes static over time.

- Most important at the moment:
  Change our model from a corruption model to a knowledge representation and transfer model. As mentioned for paper 1 above, Uni Bielefeld is an actor in a EU project about research networks in Europe. The idea now is to create a GEP (generalized epidemic process) that is suitable to represent the kind of information that is living in and travelling on the EU research networks.

As a concluding remark, I would like to thank Tyll Krueger especially for the phase in which I iteratively programmed this model, we were in constant exchange about the best possible models. It was very interesting and fruitful to combine the theoretical and the inductive viewpoints, and especially the moments when the computer programm surprised us were really instructive. The close contact of theoretical modelling and numerical programming made possible the creation of this versatile new computer program.

Corruption as a generalized epidemic process (GEP)
Ph.Blanchard, A. Krueger, T. Krueger, P. Martin
arXiv:physics/0505031

Abstract: We study corruption as a generalized epidemic process on the graph of social relationships. The main difference to classical epidemic processes is the strong nonlinear dependence of the transmission probability on the local density of corruption and the mean field influence of the overall corruption in the society. Network clustering and the degree-degree correlation play an essential role in corruption dynamics. We discuss phase transitions, the influence of the graph structure and the implications for epidemic control. Structural and dynamical arguments are given why strongly hierarchically organized societies like systems with dictatorial tendency are more vulnerable to corruption than democracies. A similar type of modelling can be applied to other social

contagion spreading processes like opinion formation, doping usage, social disorders or innovation dynamics.

You will find paper 2 at the end of this framework text.

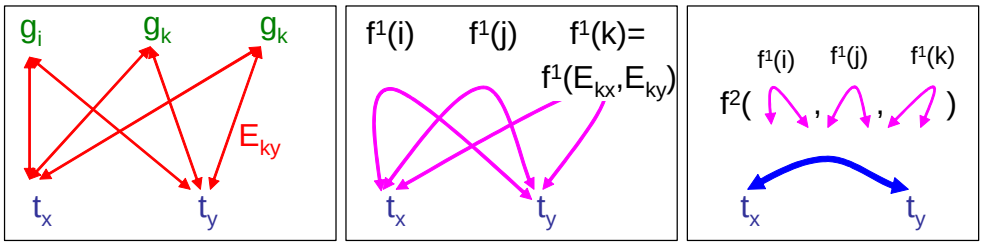


FIG. 13: The weighted projection scheme for a weighted bipartite graph with nodes g_i and t_x and weights E_{ix}. First a 2-step path over a to-be-eliminated node is contracted to one path by $f^1 = f^{\rightarrow\rightarrow}$, then all of those paths added up by $f^2 = f^{\rightrightarrows}$.

5. PAPER 3: CAMBO

In the course of working on gene-tumor data, the necessity for the *clustering* of networks arouse, so we developed an own approach towards this important perspective on the mesoscopic scale of networks. "Clustering" of all the nodes into subgroups is possible if they can be put into groups *with higher connectivity within than between the groups.*

5.1. Genes and Tumors

Why have we taken up the study of genes? The starting idea was that we could draw an analogy between genes/tumors and organisations/projects of paper 1. Imagine the genes of an individual to take the place of an organisation in Europe, and a tumor to be like a research project: Many genes work together to create a tumor, several tumors might need the influence of the same gene. Now, analogeous to the situation in paper 1, we can look at only the genes by a gene-gene projection, and look at only the tumors by a tumor-tumor projection.

One complication newly introduced for this examination, though, was that we did not want to simplify the given data as much as for the first experiments on the European Research Networks; in the case of the gene data we wanted to keep the *weights* on the links, so we had to extend the concept of projection to a weighted projection scheme.

5.1.1. Projection of Weighted Bipartite Networks

In some cases like this one it is straightforward how to project the data. Generally speaking, one always needs at least a rudimentary knowledge what kind of data is represented by the numbers in the bipartite network, to be able to decide how to combine such numbers with themselves.

The dataset used is the Alizadeh *et. al.* [1]-lymphoma-tumor-dataset of microarray gene log-expression-levels with almost complete 96×4026 tumor×gene information; each of the almost 400 000 numbers gives the logarithm of a microarray measurement of the expression level of that gene on that tumor.

Let us now first look at the *tumor-tumor-projection*, so we aggregate all information given by all the 4026 genes into only one number for each of the tumor-tumor-relatedness in terms of their gene-similarity:

Let g_i and g_j be two genes, and t_x and t_y two tumors, and let their log-expression-levels be E_{ix}, E_{iy}, E_{jx} and E_{jy}. The first operation is to contract the 2 paths $t_x \overset{E_{ix}}{\longleftrightarrow} g_i \overset{E_{iy}}{\longleftrightarrow} t_y$ and

$t_x \overset{E_{jx}}{\longleftrightarrow} g_j \overset{E_{jy}}{\longleftrightarrow} t_y$ to joint weights $t_x \overset{f^{\rightarrow\rightarrow}(i)}{\longleftrightarrow} t_y$ and $t_x \overset{f^{\rightarrow\rightarrow}(j)}{\longleftrightarrow} t_y$. The second operation is to combine several of such joint weights into one single weight $t_x \overset{f^{\rightrightarrows}}{\longleftrightarrow} t_y$ between t_x and t_y. After long discussions we decided to choose for both 2 functions the *sum*:

$$f^{\rightarrow\rightarrow}(i) = E_{ix} + E_{iy} \tag{8}$$

$$f^{\rightrightarrows} = \sum_i f^{\rightarrow\rightarrow}(i) \tag{9}$$

One interpretation why we chose this is: The E_{ix} contain *log*-expression levels, so contracting two consecutive paths (8) by addition is really *multiplying* their expression levels, comparable to the combination of two independent stochastic variables. For the path combination (9) of the 4026 tumor-gene-tumor paths we also chose the simple sum to reflect the fact that all genes connect the 2 tumors in the same, additive way.

In other situations, $f^{\rightarrow\rightarrow}$ and $f^{\rightrightarrows}$ must be chosen differently. A nice example is the inner structure of research projects described by *project partners* t_x and *work packages* g_i that they work in. Imagine that the E_{ix} and E_{iy} give the *number of hours* that people of projects t_x and t_y spend in a work package g_i, and that we want to do the partners-partners projection, to see how much a given work package structure brings the partners into contact. In this case, obviously *not the sum* of the two participation hours, but rather the number of *common hours* in one work package is an appropriate proxy for the possible contact, so we would choose the *minimum* of the two durations

$$f^{\rightarrow\rightarrow}(i) = \min(E_{ix}, E_{iy}) \tag{10}$$

for the path contraction. And for the path combination $f^{\rightrightarrows}$ again the sum (9), to simply add up all the effects of the different workpackages.

Please see figure 13 for a sketch of this concept of weighted projection. An interesting side effect of this construction is the natural incorporation of *missing data* - if some E_{ix} have not be measured, they are simply not used for the path contraction.

5.1.2. Network-Of-Clusters generated from Any-Data (NOCAD)

We work on the projection onto genes to identify those genes that are collectively switched on or off for certain tumors; then -with clustering methods- they can be separated into **clusters of similiar genes**. That is not the complete genetical process information yet, because the interaction of genes is not analyzed, what we see is only a static snapshot. Probably, genetically encoded algorithms are string-like gene sequences, that (like a Turing machine [[41]]) control the protein factory between a start and stop marker. Our method can only identify and distinguish the probable building blocks of such algorithms.

But if that information can be extracted, the next step will probably allow e.g. to draw medically interesting conclusions: Which gene groups cause which tumors? During the ongoing research we will hopefully be able to identify the *gene groups* that are highly correlated with certain *tumor groups*. Please see figure 14 for the planned research project.

We postulate that the same scheme can actually even be applied to *any* data table E_{ix}, as long as we can agree on any weighted projection scheme appropriate for the type of data that is kept in the table E_{ix}. Even freely formed database tables (with strings, dates, numbers, etc.) may probably be incorporated, as long as we can formulate a tuple metrics that gives the distance of two tuples in the table.

The CAMBO clustering algorithm was found while thinking about step 3 in that network-of-clusters from any-data scheme (NOCAD) in figure 14.

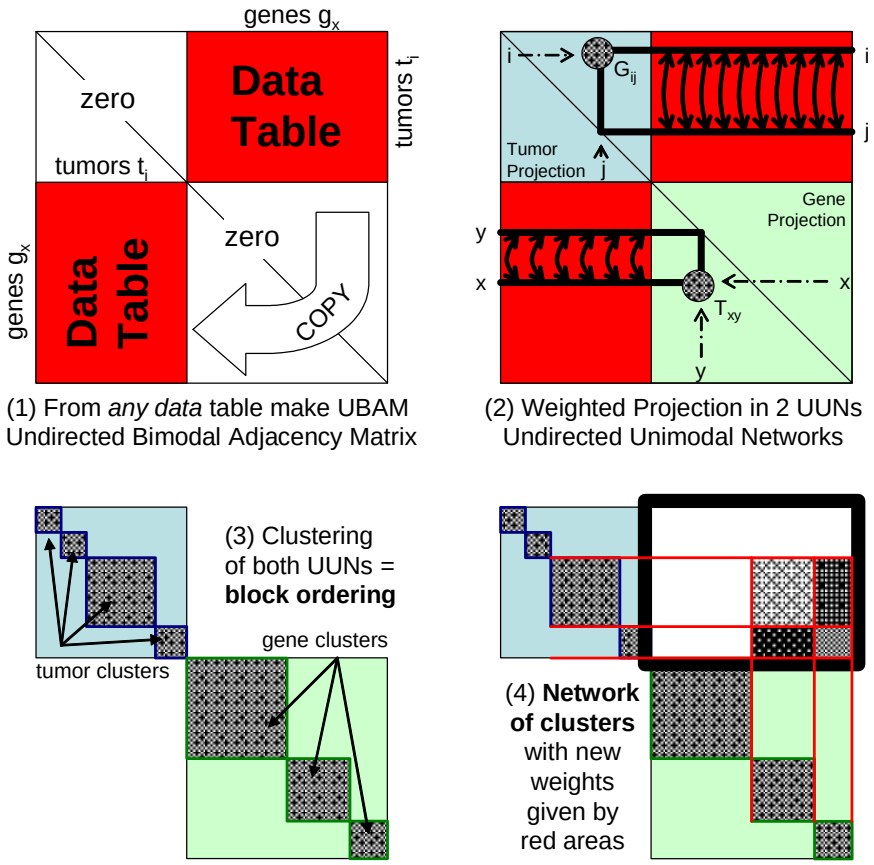


FIG. 14: The scheme how to create networks of clusters from *any* data table.

5.2. CAMBO - Clustering by Adjacency Matrix Block Ordering

There are already many good clustering algorithms but the physics community is still developing and has not converged towards THE best clustering algo, so why not think about an own approach? The idea of directly sorting the adjacency matrix sounded clever, so we spent the time. The algorithm itself is explained in the paper, so here we want to mention only some minor aspects.

CAMBO can be seen as a prototype for algorithms that cluster with parameter "dials" that represent important network features like edge weights, triangles, N_1- and N_2-structural equivalence, etc. (Please suggest others!) - we focussed on the most straightforward adjacency matrix operations like row distances, square of the matrix, etc..

As there are usually many good clusterings, the exhaustive search in parameter space usually gives several answers that are all quite good. To be able to choose one "best" clustering, there are not many accepted ways, one is to calculate the Newman Modularity (see paper 3 for details). That is a one dimensional number, so you can imagine that it will always only capture one viewpoint onto "best" clusterings of the networks. If theoreticians will develop other modularity measures, they can be put into the algorithm instead of the Newman Modularity - until then we have to put up with this, because for an algorithm like CAMBO we *need* a criterion to compare two clusterings on an order scale.

5.2.1. Metric vs. parametrized heuristics

In order to sort the lines of the matrix into block order, we calculate mutual pseudo-distances for all line-pairs, please see paper 3 for details. This pseudo-distance is no real distance, though:

Let M be any set. A function $d : M \times M \rightarrow \mathbb{R}$ is called a **metric** if

1) $d(x, y) \geq 0$ (non-negativity)

2) $d(x, y) = 0$ iff $x = y$ (identity)

3) $d(x, y) = d(y, x)$ (symmetry)

4) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

A metric is also called *distance function*, or simply *distance*.

In our case we have symmetric adjacency matrices as the networks are undirected, so the line difference heuristics obeys symmetry (rule 3), but it is not always non-negative due to the subtraction, and thus two lines can have a "line difference" of zero and still be different, and the triangle equation does not hold.

Still we are able to iteratively sort the matrix into a block order by always choosing the next-best line in relation to the already chosen lines, with "next-best" meaning that line differences pseudo-distance.

5.3. Trees as a simple example

An instructive question was how the CAMBO algorithm will cluster rooted regular *trees*. We tried with regular rooted 3-trees (so per generation each node gives birth to 3 more nodes). You see the resulting clusterings as colours in figures 15 and 16, and their modularities in figure 17, with some causing (τ, β, γ)-parameters mentioned on the x-axis.

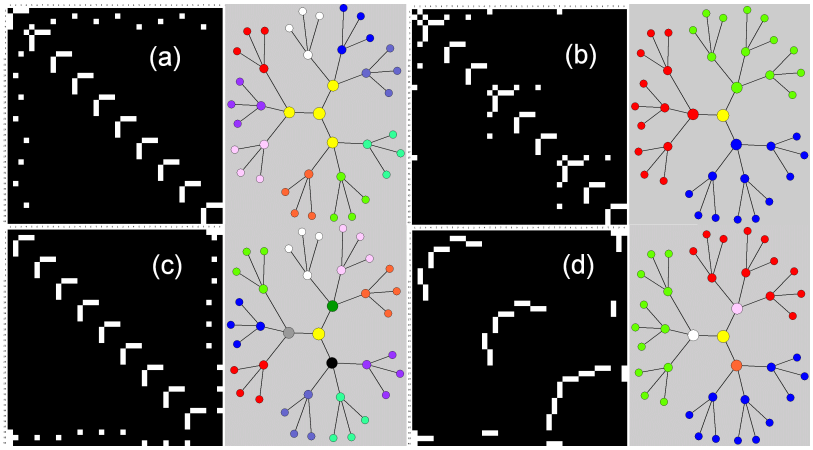


FIG. 15: Clusterings of a rooted 3-tree with 3 generations. Different colours mean different clusters, and to the left is always shown the adjacency matrix.

One interesting feature of the CAMBO algorithm is that it can group together nodes only due to their neighbourhood similiarity, even if there is no direct connection between these nodes, e.g. example (d) in figure 15 is such a case - all red, green and blue nodes are counted into each one cluster even without being directly connected.

The overall *best* clustering (a), though, is the one that puts the four nodes of the root 3-star into one cluster, and all subsequent subtrees into one cluster each. From there we get to the second

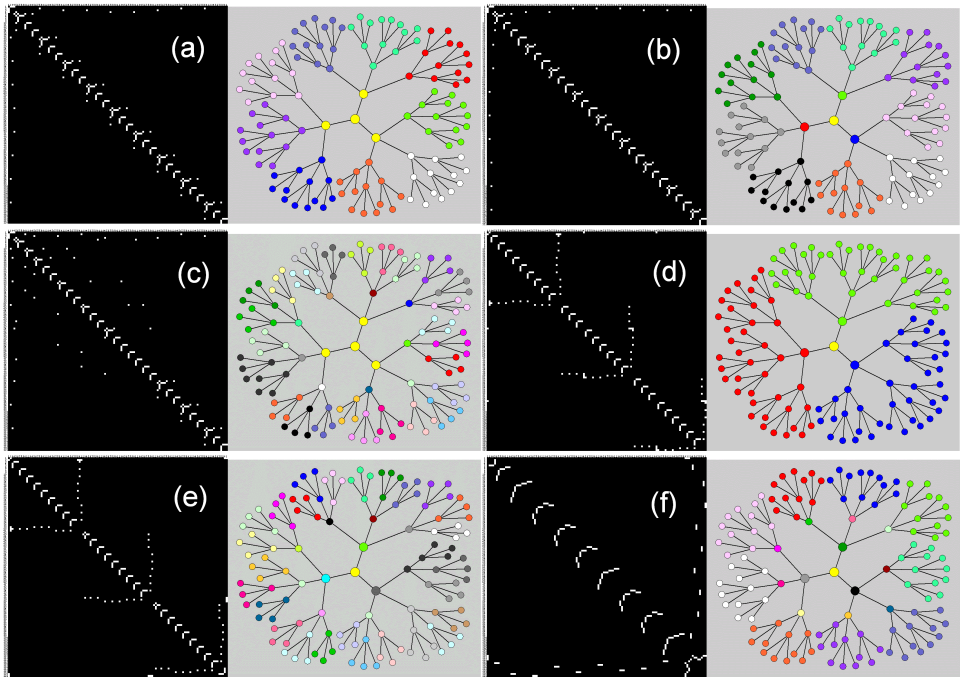


FIG. 16: Clusterings of a rooted 3-tree with 4 generations. Different colours mean different clusters, and to the left is always shown the adjacency matrix.

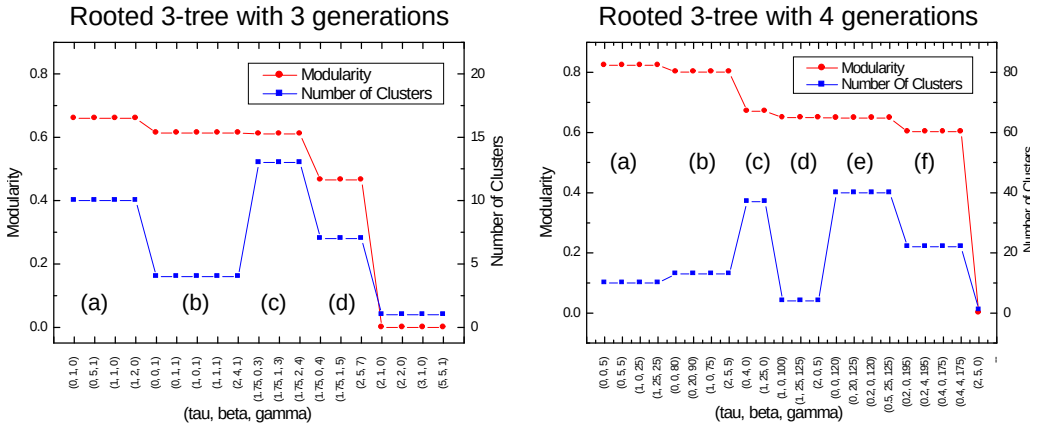


FIG. 17: Modularity and number of clusters of the 4 (6) best clusterings for the trees in figures 15 and 16 found by the CAMBO algorithm. The clusterings are shown in decreasing Newman modularity. On the x-axis, some (τ, β, γ)-parameter points are given for each such clustering.

best clustering (b) with only the root in the first yellow cluster, if we increase the γ-contribution of the second-next-neighbours $(\tau, \beta, \gamma) = (0, 0, 1)$ or if we also increase the τ-contribution of the N_1-neighbourhood similarity $(\tau, \beta, \gamma) = (1, 0, 1)$. Both effects "pull away" the subtrees from the root, into own clusters.

Clustering by Adjacency Matrix Block Ordering (CAMBO)
A. Krueger, 2007

Abstract: Clustering results are often visualized as block-structured adjacency matrices. When the nodes are clustered and sorted by their cluster order, the adjacency matrix shows *blocks* of more-strongly connected subspaces along the matrix diagonal. The inspiring idea of our new algorithm was: Why not **directly** sort the nodes into such a block-structure?
We inductively developed a deterministic algorithm that uses a parametrized heuristic of mutual 'distances' of all nodes, reorders them by smallest distances in a linear chain, cuts between clusters at the highest distance jumps, and takes the one clustering with the best modularity as the end result. The three parameters influence the mixing of the direct connection weight A_{ij}, the two-step connections $(A^2)_{ij}$, the N_1-neighbourhood similarity, and the N_2-neighbourhood similarity. A proof-of-concept-implementation suitable for small networks is described. The algorithmic time complexity is $O(N^3)$ due to the matrix multiplication, we give a discussion of possible enhancements to the algorithm. The fruitfulness of this approach is shown through application to several networks: the Zachary Karate Club, where an unknown high-modularity 3-clustering could be found by our method; a set of 96 tumors that are clustered by their gene-similarity; and clustered topics of 27000 EU-funded R&D projects.

You will find paper 3 at the end of this framework text.

6. OUTLOOK

In this thesis, three different aspects of the current network hype in physics were examined: Real-world data and modelling (EU), processes on networks (GEP), and clustering (CAMBO).

For the future, they will converge in several ways: By the application of the GEP process model for the modelling of knowledge transfer in EU networks. And by applying the bipartite clustering scheme for any data (NOCAD) to the EU networks and to the genes and tumors.

If the NOCAD research plan really succeeds with step 4 (the *networks of clusters,* which then again can be clustered themselves!), then with this "renormalization scheme" we will be able to create a "data clustering machine" DCM that will cluster any given 2-dimensional data tables, so we will be able to say:

> Give us any data, and with network methods
> we will always extract some inner order.

Acknowledgement 1 *This work has been supported in part by the European FP6-NEST-Adventure Programme, contract number 028875. Andreas Krueger would like to thank: The Volkswagen Foundation, and the Systems Research unit of the Austrian Research Center, for financing this PhD; Tyll Krueger and Philippe Blanchard for our always-inspiring working atmosphere; Madeleine Sirugue-Collin, Pierre Chiappetta for fruitful discussions about genes and tumors, Michael Barber by far not only, but especially for persuading me to use Python, Thomas Roediger-Schluga for pursuing the idea of a project between social and natural sciences, and Hanne for being the good soul of our department.*

[1] Alizadeh et. al.: *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*, Nature **403**: 503-511 (2000), http://llmpp.nih.gov/lymphoma

[2] Bagrow, Bollt, da F. Costa, **2006**, *Network Structure Revealed by Short Cycles*, cond-mat/0612502

[3] Bagrow, Bollt, Skufca, ben-Avraham, **2007**, *Portraits of Complex Networks*, cond-mat/0703470

[4] A.L.Barabasi, R.Albert, *Emergence of Scaling in Random Networks*, Science 286, 509 (**1999**), arXiv:cond-mat/9910332

[5] A.L.Barabasi, R.Albert, R.Albert, *The diameter of the world-wide web*, **1999**, Nature (London) 401, 130-131; arXiv:cond-mat/9907038

[6] R.Albert , A.-L. Barabasi : *Statistical Mechanics of Complex Networks*, Reviews of Modern Physics, 74, 47 (**2002**), arXiv:cond-mat/0106096

[7] Barabasi, *Linked: How Everything Is Connected to Everything Else and What It Means*, 2003, ISBN 978-0452284395, http://www.nd.edu/~networks/Linked

[8] Borgatti, S.P. and Molina, J-L. **2003**. *Ethical and strategic issues in organizational network analysis*. Journal of Applied Behavioral Science. 39(3): 337-350. http://www.analytictech.com/borgatti/papers/ethics.pdf

[9] Stephen P. Borgatti and Jose-Luis Molina, *Toward ethical guidelines for network research in organizations*, Social Networks, Volume 27, Issue 2, Ethical Dilemmas in Social Network Research, May **2005**, Pages 107-117.

[10] Ronald L. Breiger, *Introduction to special issue: ethical dilemmas in social network research*, Social Networks, Volume 27, Issue 2, Ethical Dilemmas in Social Network Research, May **2005**, Pages 89-93. http://www.u.arizona.edu/~breiger/

[11] *CAMBO*-Website will be linked from http://www.AndreasKrueger.de/networks

[12] CORDIS *Community Research & Development Information Service, European Communities*, **1990-2007**, http://cordis.europa.eu/search

[13] Robin Cowan, MERIT, **2004**, *Network models of innovation and knowledge diffusion*, http://www.merit.unu.edu/publications/rmpdf/2004/rm2004-016.pdf

[14] *CPT* Centre de Physique Theoretique, Marseille Luminy, France, http://www.cpt.univ-mrs.fr

[15] L. Danon, J. Duch, A. Diaz-Guilera, A. Arenas: *Comparing community structure identification*. Journal of Statistical Mechanics: Theory and Experiment, **2005** doi:10.1088/1742-5468/2005/09/P09008

[16] de Moura A.P., Lai Y.C., Motter A.E., *Signatures of smallworld and scale-free properties in large computer programs*, Physical Review E 68, 017102 July **2003**.

[17] S.N. Dorogovtsev, J.F.F. Mendes, *The shortest path to complex networks*, **2004,** arXiv:cond-mat/0404593

[18] *EconoPhysics* conferences at http://www.ge.infm.it/~ecph/events/

[19] Girvan M and Newman M E J, 2002, *Community structure in social and biological networks*, Proc. Natl. Acad. Sci., 99, 7821.

[20] P. M. Gleiser: *How to become a superhero*, J. Stat. Mech. (2007) P09020, arXiv:0708.2410

[21] For an excellent portal visit the WikipediaPage *GraphTheory* http://en.wikipedia.org/wiki/Graph_theory

[22] *Illegale Schule in Bremer Villa blieb 14 Jahre unentdeckt*, Die Welt, 13.10.2007, http://www.welt.de/welt_print/article1261911/ Illegale_Schule_in_Bremer_Villa_blieb_14_Jahre_unentdeckt.html

[23] E. Ising, Zeitschrift f. Physik 31, 253 (**1925**)

[24] Charles Kadushin, *Who benefits from network analysis: ethics of social network research*, Social Networks, Volume 27, Issue 2, Ethical Dilemmas in Social Network Research, May **2005**, Pages 139-153. http://www.sciencedirect.com/science/article/B6VD1-4FH0W77-1/2/f4f6cc8f4842373f1665496fef96fb4e

[25] *Das Ising-Modell - gestern und heute*, Sigismund Kobe, TU Dresden http://www.physik.tu-dresden.de/itp/members/kobe/isingphbl

[26] Andreas Krueger, *Dimensionality in Continuum Percolation Thresholds*, **2002,**Physics Diploma Thesis Uni Bielefeld, http://www.andreaskrueger.de/thesis

[27] M. Barber, A. Krueger, T. Krueger and T. Roediger-Schluga: *The Network of EU-Funded Collaborative R&D Projects,* Phys. Rev. E 73, 036132 (**2006**), arXiv: physics/0509119

[28] Andreas Krueger, Clustering by Adjacency Matrix Block Ordering (CAMBO)

[29] T.Krueger, A.Krueger, Corruption as a generalized epidemic process (GEP)

[30] Tyll Krueger, personal communication, 24.10.2007

[31] S.Milgram, (1967). *The small world problem.* Psychology Today, 2, 60-67.
[32] *NEMO - Network Models, Governance and R&D collaboration networks*, EU-project contract#028875, http://www.nemo-net.eu
[33] Newman, M. E. J., *Assortative Mixing in Networks*, PhysRevLett.89.208701, **2002**
[34] M.˜E.˜J. Newman, M. Girvan, *Finding and evaluating community structure in networks*, Physical Review E, 69, 026113 (**2004**), cond-mat/0308217
[35] Clauset, Newman, Moore, *Finding community structure in very large networks*, Phys. Rev. E 70, 066111 (**2004**), arxiv:condmat/0408187
[36] M.˜E.˜J. Newman, **2006**, *Finding community structure in networks using the eigenvectors of matrices*, Physical Review E, 74, 036104
[37] V.Batagelj, A.Mrvar, *Pajek - Program for Large Network Analysis*, http://vlado.fmf.uni-lj.si/pub/networks/pajek
[38] http://www.python.org
[39] Jorg Reichardt and Stefan Bornholdt, **2004**, *Detecting Fuzzy Community Structures in Complex Networks with a Potts Model*, Phys. Rev. Lett. 93, 218701
[40] *SocioPhysics*, Abstract Book, ZIF Uni Bielefeld, **6.6.2002**, Frank Schweitzer, Klaus G. Troitzsch, http://www.uni-bielefeld.de/ZIF/AG/2002
[41] *Turing machine* (Wikipedia article) http://en.wikipedia.org/wiki/Turing_machine
[42] Stijn van Dongen, *Graph Clustering by Flow Simulation.* PhD thesis, University of Utrecht, May **2000**, http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm
[43] Stijn van Dongen, *MCL = Markov Cluster Algorithm*, http://micans.org/mcl
[44] D.J.Watts, S.H.Strogatz, *Collective dynamics of 'small-world' networks*, Nature 393, 440 (**1998**)
[45] Douglas R. White (editor), *Networks And Complexity*, Complexity Special Issue **2002,** vl. 7 no. 6, http://cse.ucdavis.edu/˜cmg/netdyn/SpecialIssue.html
[46] Woodhouse DE, Potterat JJ, Rothenberg RB, Darrow WW, Klovdahl AS, Muth SQ. *Ethical and legal issues in social networks research: the real and the ideal*, in Needle RH, Genser SG, Trotter II RT (eds): Social Networks, Drug Abuse and HIV Transmission. National Institute on Drug Abuse Monograph No. 151 (NIH Publication No. 953889); **1995** :131-143.

The Network of EU-Funded Collaborative R&D Projects

M. J. Barber
Centro de Ciências Matemáticas, Universidade da Madeira, Funchal, Portugal

A. Krueger and T. Krueger
Universität Bielefeld, Bielefeld, Germany

T. Roediger-Schluga
ARC systems research, Vienna, Austria

(Dated: June 6, 2007)

We describe collaboration networks consisting of research projects funded by the European Union and the organizations involved in those projects. The networks are of substantial size and complexity, but are important to understand due to the significant impact they could have on research policies and national economies in the EU. In empirical determinations of the network properties, we observe characteristics similar to other collaboration networks, including scale-free degree distributions, small diameter, and high clustering. We present some plausible models for the formation and structure of networks with the observed properties.

I. INTRODUCTION

Real world network analysis has become a major issue of research in the last years. Most prominent are perhaps the investigations of the structure of the World Wide Web, the network of internet routers, and certain social networks like citation networks. On the theoretical side, one tries to understand the mechanisms of formation of such networks and to derive statistical properties of the networks from the generating rules. On the rigorous mathematical side, there are only a few results for specific models, indicating the difficulty of a purely mathematical approach (for a survey of recent results in this direction, see [7]). Thus, the main approach is to use some mean field assumption to get relevant information about the corresponding graphs. Although it is not clear where the limits of this approach lie, in many cases the results match well with numerical simulations and empirical data.

In this article, we study a particular collaboration network. Its vertices are research projects funded by the European Union and the organizations involved in those projects. In total, the data base contains over 20000 projects and 35000 participating organizations. The network shows all the main characteristics known from other complex network structures, such as scale-free degree distribution, small diameter, high clustering, and inhomogeneous vertex correlations.

Besides the general interest in studying a new, real-world network of large size and high complexity, the study could have a significant economic impact. Improving collaboration between actors involved in innovation processes is a key objective of current science, technology, and innovation policy in industrialized countries. However, very little is known about what kind of network structures emerge from such initiatives. Moreover, it is quite likely that network structure affects network functions such as knowledge creation, knowledge diffusion, and the collaboration of particular types of actors. Presumably, this is determined by both endogenous formation mechanisms and exogenous framework conditions. In order to progress in our understanding, it is therefore essential to have sound statistics on the structure of networks we observe and to develop plausible models of how these are formed and evolve over time.

The model networks we use to compare with the empirical data are random intersection graphs, a natural framework for describing projections of bipartite graphs. Discrete intersection graphs similar to the ones we use were first discussed in [8]. We extend and refine the construction from [8] to be more applicable to real world graphs.

Perhaps the most important finding from our model approach is the strong determination of the real network structure by the degree distribution. That is, most statistical properties we measure in the EU research project networks are the ones observed in a typical realization of a uniform weighted random graph model with given (bipartite) degree distribution as in the EU networks. Since this distribution is characterized by two exponents—one for each partition—we have essentially only four parameters (size, edge number, and exponents) which are needed to describe the entire network. This is a tremendous reduction of complexity indicating that only a few basic formation rules are driving the network evolution.

In section II, we describe the preparation of the data on the EU research programs. We present empirical determination of the network properties in section III, followed by an explanation of these properties using a random intersection graph model in section IV. Finally, in section V, we summarize the key results and consider implications of the network properties on EU research programs.

II. THE DATA SET

In this work, we study research collaboration networks that have emerged in the European Union's first four successive four-year Framework Programs (FPs) on Research and Technological Development. Since their inception in 1984, six FPs have been launched, on four of which we have comprehensive data. FPs are organized in priority areas, which include information and communication technologies (ICTs), energy, industrial technologies, life sciences, environment, transportation, and a number of additional activities. In line with economic structural change, the main thematic focus of the FPs has shifted somewhat over time from energy and industrial technologies to the application of ICTs and life sciences. The majority of funding activities are aimed at stimulating research partnerships between firms, universities, research organizations, governmental actors, NGOs, lobby groups, etc.. Since FP4, the scope of activities has been expanded to also cover training, networking, demonstration, and preparatory activities (for details, see reference [1]). In order to keep our data set compatible over the different FPs, we have excluded the latter set of projects from FP4 and only focus on collaborative research projects (see table I).

In order to receive funding, projects in FP1 to FP4 had to comprise at least two organizations from at least two member states. We have retrieved data on these projects from the publicly available CORDIS (Community Research and Development Information Service) projects database [10]. This database contains information on all funded projects as well as a reasonably complete listing of all participating organizations.

The raw data on participating organizations is rather inconsistent. Apart from incoherent spelling in up to four languages per country, organizations are labelled inhomogeneously. Entries may range from large corporate groupings, such as Siemens, or large public research organizations like the Spanish CSIC to individual departments or labs and are listed as valid at the time the respective project was carried out. Among heterogeneous organizations, only a subset contains information on the unit actually participating or on geographical location (address, city, region and/or country). Information on older entries and the substructure of firms tends to be less complete.

Because of these difficulties, any automatic standardization method akin to the one utilized by Newman [9] is inappropriate to this kind of data. Rather, the raw data has to be cleaned and completed manually, which is an ongoing project at ARC systems research. The objective of this work is to produce a data set useful for policy advice by identifying homogeneous, economically meaningful organizational entities. To this end, organizational boundaries are defined by legal control and entries are assigned to the respective organizations. Resulting heterogeneous organizations, such as universities, large research centres, or conglomerate firms are broken down into subentities that operate in fairly coherent areas of activity, such as faculties, institutes, divisions or subsidiaries. These can be identified for a large number of entries, based on the available contact information of participants, and are comparable across organizations.

The case of the French Centre National de la Recherche Scientifique (CNRS), the most active participant in the EU FPs may serve as an illustration. First, 785 separate entries were summarized under a unique organizational label. Next, these 785 entries were broken down into the eight areas of research activity in which CNRS is currently organized. Based on available information on participating units and geographical location, 732 of the 785 entries could be assigned to one of these subentities. For the remaining 53 entries, the nonspecific label CNRS was used.

Comparable success rates were achieved for other large public research organizations and universities. Due to scarcer information, firms could not be broken down at a comparable rate. Moreover, due to resource constraints, standardization work has focused on the major players in the FPs. Organizations participating in fewer than a total of 30 projects in FP1–4 have not been broken down yet. Due to these limitations in processing the data, we cannot rule out the possibility of a bias in analysing our data. However, we have run all the reported analyses with the undivided organizations and have obtained qualitatively similar results, apart from different extreme values, e.g., maximum degree.

Table I displays information on the present data set, which contains information on a total of 27,758 projects, carried out over the period 1984 to 2004. It shows that the total budget as well as number of funded projects has increased dramatically from FP1 to FP4. Moreover, it provides a rough measure on the completeness of the available data. For a sizeable number of projects, the CORDIS project database lists information only on the project co-ordinator. This is due to the age of the data and inhomogeneous disclosure policies of different units at the European Commission. Comparing the number of projects containing information on more than one participant with the total number of projects funded in each FP shows that the data is fairly complete as of FP2.

The fact that FP1 was the first program launched and that the available data are rather incomplete make it exceptional in many respects. We therefore focus our analyses on FP2–4 and only give graph characteristic values for FP1 to indicate the difference to the networks created by the subsequent FPs.

III. THE NETWORK STRUCTURE

In this section, we present the basic properties of the network structure for projects and organizations in the first four EU Framework Programs. We consider both graphs as intersection graphs, each being the dual of the other, which, for our purposes, is generally more convenient than the usual bipartite-graph point of view. Recall that an intersection graph is given by an enumerated collection of sets—the vertices of the intersection graph—with elements from a given fixed base-set and edges defined via the intersection property (edge $\triangleq$ nonempty intersection of two sets). The sets need not be distinct.

We denote by $\mathcal{P} = \{P_1; ... : P_M\}$ the family of projects and by $\mathcal{O} = \{O_1; ...; O_N\}$ the family of organizations. Projects are understood as labeled sets of organizations and organizations as labeled sets of projects. The corresponding intersection-graphs are denoted by G_P and G_O and we will sometimes use the terms P-graph and O-graph for them. The size $|x|$ of a vertex x from G_P or G_O is the cardinality of the set corresponding to the vertex; in the picture of bipartite graphs, the size is just the degree of the vertex. In tables II and III, we give some basic parameters measured on the P- and O-graphs from the four Framework Programs. Since the degree distribution for P-graphs is a superposition of two power-law distributions (one for small degree values and one for large values), we give the corresponding values for the exponents parenthetically.

As expected, FP1–4 are of small world type: high clustering coefficient and small diameter of the giant component. There is a slight increase in the clustering coefficient of the O-graphs from FP1 to FP4, indicating a stronger integration amongst groups of collaborating organizations. This is also reflected in the mean project size which increases from 2.4 to 6.2. There is an interesting jump in the P-graph mean degree values and the mean triangle numbers between FP1 and 2 and between FP2 and 3. The maximal degree of the O-graphs are very high in comparison with the mean degree, which is a consequence of the power law degree structure. For the P-graphs, the gap between mean and maximal degree is less pronounced.

More information is contained in the statistical properties of the relevant distributions. The numerical data strongly indicate that the size distributions follow power laws. Also, the O-graph degree distribution is of power-law type, while the project-graph degree distribution is a superposition of two scale free distributions, one dominating the distribution for small degree values (up to 100) and one relevant for the large degree values. We discuss these properties at greater length in the following sections.

A. Size distributions

The size distributions are the basic distributions for the EU-networks since, as will be shown in section section IV B, a typical sample from the random graph space with fixed size distributions like in FP 2-4 will have very similar statistical properties to FP 2-4. This strongly suggests that there is essentially no additional correlation in the data once the size distribution is known. Both the O-graph and P-graph size distributions show clear asymptotic power law distributions for FP1–4 (figs. 1 and 2). In terms of the corresponding bipartite graph, these are just the degree distributions of the project and organization partitions. While the O-graph size distribution is of power law type over the whole size range, the P-graph size distribution deviates strongly from the power-law for small size values. In section IV, we give a possible explanation for the appearance of the power law distribution for size.

The numerical values for the exponents of the organization size distributions from FP2–4 are slightly below 2, but constant within the error tolerance. This indicates that the distribution of organizations able to carry out a particular number of projects has not changed in the three Framework Programs. A complementary interpretation of this finding is that the underlying research activities, which we know to have changed over time, have not altered the mix of organizations participating in a particular number of projects in each Framework Program. It is further worth noting that the values of the O-graph exponents are close to the critical value 2, hence the size expectation could diverge for large graphs (whether the value is really below 2 or not is still unclear due to the error tolerance).

The picture is similar for the P-graphs, although there are some differences in the initial behavior (that is, for small project sizes) and in the exponent value. The local minima at size 2 is decreasing from FP2–4. This points to the existence of an optimal project size within the regime of the EU FPs. Moreover, the rise in the average project size indicates that increases in the available funding from FP2 to FP4 not only lead to more projects, but also slightly larger projects. This is consistent with recommendations from evaluation studies and the stated attempts of the EU commission to reduce its administrative burden. As a whole, the size distribution for the P-graphs matches in the asymptotic regime very well to a power law with exponent around -3, hence indicating that the mechanisms for coagulation of organizations into a project did not greatly change from FP2–4.

B. The degree distribution

Since the degree distribution in the projection graphs is just the distribution of the size of the 2-neighborhood $N_2(x) := \#\{y : d^{bi}(x,y) = 2\}$, it is not surprising that this quantity is closely connected to the size distribution. In the absence of other special correlations, it can be shown (see section IV) that the degree distribution is determined by the size distribution in a rather simple way. Namely, for the case when both size distributions are scale-free with exponents, say α (O-size) and β (P-size), the P-graph degree distribution is a superposition of two power-law distributions with exponents $\alpha - 1$ (and cutoff given by the maximal O-size value) and β. The same holds vice versa for the O-graph.

In figs. 3 and 4, we show the degree-distribution for the P- and O-graphs in a log-log plot. . While the organization graphs for FP2–4 show a clear power law, the picture for the project graphs is more complicated. As previously mentioned, the P-graph degree distribution shows two different power laws, one for the initial segment up to degree 150 and another one for large degrees. Nevertheless, there is still a widely scattered heavy tail in the degree distribution. The deviation from a power law in the P-graphs indicates a kind of anticorrelation: large projects above a size of 15 are mainly formed by organizations of small size. A possible explanation is that large projects have a time- and resource-demanding intrinsic network structure, making it more unlikely that a participating organization has other projects (of course, with the exception of hub-like organizations such as CNRS with *a priori* unlimited capacity).

C. Clustering, correlation and edge multiplicity

By their construction process, intersection graphs have a naturally high clustering coefficient. This is easily seen, since an organization which participates in, say, k projects generates a complete subgraph of order k in the P-graph amongst these projects. If the probability for an organization to be in more than one project is asymptotically bound away from zero, it follows that the P-graph (and similarly for the O-graph through an analogous argument) has a nonvanishing clustering coefficient. In the present study, we focus on the triangle number $\triangle(x) := \#\{\text{triangles containing } x; x \in (\mathcal{P} \text{ or } \mathcal{O})\}$ as a measure of local clustering. We define the degree-conditional mean triangle number as $\triangle_k := \mathbb{E}\{\triangle(x) \mid d(x) = k\}$. As seen in figs. 5 and 6, we have $\triangle_k \sim k$ for both graph types.

There is a good explanation for this type of behavior in the framework of intersection graphs (see section IV). As noted above, high clustering in intersection graphs is not necessarily an indication of local correlations between vertices. This is already seen in the case of an Erdös-Renyi random bipartite graph where an edge between any project and organization is drawn i.i.d. with probability p. If $\mathcal{P}$ and $\mathcal{O}$ are of equal cardinality N and $p = \frac{c}{N}$, the expected bipartite degree equals c. For large N a typical realization of the random graph looks locally like a tree with branching number $c - 1$. However, for the projection graphs, we obtain an positive clustering coefficient that is independent of N, since most projects and organizations cause complete graphs of order c and a typical vertex is therefore a member of $\sim c$ cliques of order c.

A better indication for the presence of correlations is given by the so-called multiplicity of edges. For a link between two organizations or projects it is sufficient to have just one project or organization, respectively, in common, but of course there could be more. Given an edge $x \sim y$, we define $m(x,y) := |x \cap y| - 1$ and call it the multiplicity of the edge. As will be discussed in the next section, random intersection graphs without local search rules can nevertheless admit a high edge multiplicity. In fig. 7 and 8, the multiplicity distribution is shown for P- and O-graphs of FP2–4. There is an almost perfect power-law behavior with exponent 4.3. Note that positive multiplicity in the projection graphs translates in the bipartite graph picture into the presence of cycles of length four. The presence of exceptionally high multiplicity in the P-graphs may be caused by memory effects due to prior collaborative experience. Also, a greater edge multiplicity may result from the fact that organizations are active in a wider set of complementary activities. In this case, intra-organizational spillovers may also be of importance as search for potential partners may be influenced by the collaboration behavior of other actors within an organization. Such effects should be detectable from a fine structure analysis of the time evolution of the corresponding graphs.

D. Diameter and mean path length

There is essentially no difference in the diameter value of the largest component in the four Framework Program networks. A classical random graph of the same size and the same edge number would have a diameter about $\log_{\bar{d}} N$. The mean path length is about a third of the diameter and and shows a slightly higher variation between the different framework programs. It is well known that the expected path length in random graphs with a scale free degree distribution and exponent less than 3 is essentially independent of the graph size (the diameter of the largest component still increases in N but only as $\log\log N$). The same holds for random intersection graphs with power law

size and degree distributions. Since the the O-graphs seem to fall into that class, the almost constant diameter and path length is not surprising. Although the P-graphs do not show an asymptotic power law structure for the degree, there is a strong increase in the edge density from FP2 to FP4, keeping the diameter of the largest component almost fixed.

IV. A RANDOM INTERSECTION GRAPH MODEL

Intersection graphs are a natural framework for networks derived from a membership relation, such as citation networks, actors networks, or networks reflecting any other kind of cooperation. As previously mentioned, intersection graphs by construction have a high clustering coefficient. As explained below, the clique distribution of a random intersection graph is almost given by the size distribution of the dual graph.

A. Random intersection graphs with given size distribution

One of the simplest random intersection models is constructed in the following way. Knowing the size of a set to be constructed, we generate a random subset from a finite base set $X = \{a_1, a_2, ..., a_N\}$ of N elements, such that each set element is drawn i.i.d. uniformly from X. These subsets constitute the vertices of a random graph. Edges are defined via the set intersection property, namely we have an edge between i and j (denoted by $i \sim j$) if and only if the associated subsets A_i and A_j have nonempty intersection (to compare with earlier sections, A stands here for either projects sets P or organization sets O). The size (cardinality) of the subsets is either itself a random variable drawn i.i.d. from a probability distribution $\varphi(k)$ or given by a list $\{D_k := \#\{A_i : |A_i|\} = k\}$ (where for each i a conditional random choice is made to which size class it belongs). For the latter case, we define again $\varphi(k) := \frac{D_k}{M}$ where M is the total number of sets to be formed.

Since we want to compare the model with the EU- cooperation network we are mainly interested in the situation when φ is an asymptotic power law distribution

$$\varphi(k) = \frac{1}{k^{\alpha+o(1)}}; \alpha > 2 \quad . \tag{1}$$

This assumption is also reasonable for many other applications where vertices are formed from a base set of elements. To obtain an interesting limiting random graph space, we further assume that the number of chosen subsets is $C_1 \cdot N$ where C_1 is neither too large nor too small (for FP2–4 we have about twice as many organization as projects hence hence C_1 is either 2 or 0.5).

A basic quantity for the analysis of intersection graphs is the conditional edge probability given the size of two subsets:

$$\begin{aligned} P_{k,l}(N) &:= \Pr\{i \sim j \mid |A_i| = k \text{ and } |A_j| = l\} && (2)\\ &= \Pr\{A_i \cap A_j \neq \emptyset \mid |A_i| = k \text{ and } |A_j| = l\} && (3)\\ &= 1 - \frac{\binom{N-k}{l}}{\binom{N}{l}} && (4)\\ &= 1 - \frac{(N-k)!\,(N-l)!}{N!\,(N-k-l)!} && (5)\\ &= 1 - \frac{(N-k)\,(N-k-1)\cdot ...\cdot (N-k-l+1)}{N\,(N-1)\,(N-2)\cdot ...\cdot (N-l+1)} \quad . && (6) \end{aligned}$$

Using the condition $lk \ll N$, we obtain

$$\begin{aligned} P_{k,l}(N) &= 1 - \frac{\left(1-\frac{k}{N}\right)\left(1-\frac{k+1}{N}\right)...\left(1-\frac{k+l-1}{N}\right)}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)...\left(1-\frac{l-1}{N}\right)} && (7)\\ &= 1 - \frac{1 - \frac{lk+\frac{1}{2}(l-1)(l-2)}{N} + o\left(\frac{1}{N}\right)}{1 - \frac{\frac{1}{2}(l-1)(l-2)}{N} + o\left(\frac{1}{N}\right)} && (8)\\ &= \frac{lk}{N} + o\left(\frac{1}{N}\right) \quad . && (9) \end{aligned}$$

With this result, we can easily calculate the conditional degree distribution for a vertex of given size. First, we estimate the conditional subdegree distribution with respect to a given group of vertices of size m. Here, the subdegree $d_m(i)$ of a vertex i is defined as the number of edges i has with vertices of size m. Clearly $d(i) = \sum_m d_m(i)$. We have

$$\psi_l(k,m) := \Pr\{d_m(i) = k \mid |A_i| = l\} \tag{10}$$

$$= \sum_G \Pr\{\sharp\{j \mid |A_j| = m\} = G\} \binom{G}{k} \left(\frac{ml}{N} + o\left(\frac{1}{N}\right)\right)^k \left(1 - \frac{ml}{N} + o\left(\frac{1}{N}\right)\right)^{G-k} . \tag{11}$$

The probability that a randomly chosen vertex j has size m equals, by assumption, $\frac{C_2}{m^{\alpha+o(1)}}$ with normalization constant C_2 ($1 = \sum_m \frac{C_2}{m^{\alpha+o(1)}}$). We therefore obtain

$$\psi_l(k,m) = \lim_{N\to\infty} \binom{C_1 N \cdot \frac{C_2}{m^\alpha}}{k} \left(\frac{ml}{N} + o\left(\frac{1}{N}\right)\right)^k \left(1 - \frac{ml}{N} + o\left(\frac{1}{N}\right)\right)^{C_1 N \cdot \frac{C_2}{m^\alpha} - k} , \tag{12}$$

which converges to a Poisson distribution

$$\psi_l(k,m) = \frac{c(m)^k}{k!} e^{-c(m)} \tag{13}$$

with $c(m) = m^{1-\alpha} l C_1 C_2$. Since the distribution $\psi_l(k)$ of the degree of vertices i with $|A_i| = l$ is the convolution of the Poisson distributions $\psi_l(k,m)$, we obtain again a Poisson distribution for $\psi_l(k)$:

$$\psi_l(k) = \frac{c_l^k}{k!} e^{-c_l} \tag{14}$$

with $c_l = \sum_m c(m) = l \cdot C_3$, where $C_3 = \sum_m m^{1-\alpha} C_1 C_2$ is a well defined constant since $\alpha > 2$. It remains to estimate the total degree distribution $\psi(k)$. In [2], conditions were given describing when a superposition of Poisson distributions results in a scale-free distribution. Specifically, we get the following asymptotic estimate:

$$\psi(k) = \sum_m \varphi(m) \frac{(mC_3)^k}{k!} e^{-mC_3} \tag{15}$$

$$= \sum_m \frac{1}{m^{\alpha+o(1)}} \cdot \frac{(mC_3)^k}{k!} e^{-mC_3} . \tag{16}$$

The main contribution to $\psi(k)$ comes from a rather small interval of m-values, called $I_{ess}(k)$. This interval has the property that for $m \in I_{ess}(k)$, the expectation $\mathbb{E}(d(i) \mid |A_i| = m)$ is of order k. The exponential decay of the Poisson distribution guarantees that the remaining parts of the sum become arbitrarily small for large k. It is important that the constant c_l has a linear $l-$dependence since an $l-$proportionality with exponent larger than one would force the degree distribution to have gaps due to a lack of overlap of the individual Poisson distributions. We therefore obtain for the degree distribution a power law with the same exponent α as in the size distribution.

Although the intersection model gives a power-law degree distribution when the size distribution is already of power-law type, we will not obtain a power-law distribution for the size on the dual graph unless additional assumptions are made on the set formation rules. It is easy to see that the size distribution on the dual graph is asymptotically Poisson. Since $\Pr\{|x| = k\} \sim \binom{M}{k} \left(\frac{\mathbb{E}(|A|)}{N}\right)^k \left(1 - \frac{\mathbb{E}(|A|)}{N}\right)^{M-k}$ and $\mathbb{E}(|A|)$ converges as well as $\frac{M}{N}$ for $M, N \to \infty$, we obtain in the limit a Poisson distribution. Nevertheless, the degree distribution on the dual graph still admits a scale-free part induced by the scale-free size distribution of the intersection graph. We will not discuss many of the details, but instead provide a simple estimation for the lower bound on the number of elements a_i with $d(a_i) = k$. Namely, the number of elements a_i which are members of sets A_j with $|A_j| = k$ is for large k and $M, N >> k$ about $\frac{k \cdot M \cdot const}{k^\alpha} = \frac{N \cdot const}{k^{\alpha-1}}$. Since $d(a_i) \geq k$ for $a_i \in A_j$ with $|A_j| = k$, we obtain $\frac{const}{k^{\alpha-2}}$ as a lower bound on the density of elements a_i with degree greater than or equal to k (note that we assumed $\alpha > 2$). This estimate holds of course only up to the maximal size value k, which is in the range of the power law distribution for the set sizes $|A_i|$. For larger k-values there is a rapid exponential decay.

The last argument clarifies also the situation when one wants to impose conditions on the size distribution and the dual size distribution. Without going into the details of the rather involved analysis, we simply state that the

resulting degree distribution is given by a superposition of the size distibution and the dual size distribution (the last one enters with an exponent reduced by one). This explains essentially the picture for the degree distribution for the P-graph.

Finally we want to discuss the mean triangle (conditioned on the degree) - degree dependence which shows a clear linear behavior in the empirical data. We argue that this is again a consequence of the power law distribution for the size. First observe that a size k element $a_i \in A_j$ induces a $k-1$ complete subgraph on the neighborhood vertices of A_j. Furthermore, each maximal $k-$clique in which A_j is a member generates $(k-1)(k-2)/2$ triangles for A_j. Since the size distribution of the elements a_i is Poisson with expectation of, say, c and the degree of A_j is proportional to the size $|A_j|$, we obtain for the conditional expected number of triangles $\triangle_k$ given the degree k:

$$\triangle_k := \mathbb{E}\left(\#\text{triangles containing } A \mid d(A) = k\right) \sim \frac{c^2}{2} const \cdot k \quad . \tag{17}$$

In deriving eq. (17), we used the facts that with high probability the size of the intersection between two sets A_i and A_j has cardinality 1 (conditioned on the two sets having a nonempty intersection) and that the Poisson distribution has an exponentially decaying tail.

B. A Molloy-Reed version of random intersection graphs and a Bernoulli type model

We sketch the construction of random intersection graphs with given size distribution φ and size distribution ψ on the dual. The two distributions are not independent but have to fulfill the condition $\sum_i [\varphi(i) - \psi(i)] = 0$. There are further restrictions on the maximal size in order to get a reasonable random graph model. Note that the problem is equivalent to the construction of a random bipartite graph given the degree sequence on the two partitions.

Assign first to each set A and each element a from the base set a random size value according to the given distributions φ and ψ. Let D_k be the resulting set of elements a_i with size k. Replace each element from D_k by k virtual elements $a_{i,l}, l = 1, 2, \ldots, k$ and form a new base set X' with all the virtual elements. The set formation process for the sets $\{A_i\}$ is now the same as in the previous section except that each chosen virtual element $a_{i,l}$ will be removed from X' when it was selected first into a set. After the sets are constructed we identify the virtual elements back into the original ones and define the corresponding set graph in the usual way.

By construction the resulting size distribution on the dual graph will be given by ψ as long as the probability of choosing two virtual elements $a_{i,l}$ and $a_{i,m}$ (corresponding to the same element a_i) is sufficiently small. To ensure this one has to impose restrictions on the maximal size values. It is not difficult to show that the correlation between the size of A and the size of an element a is multiplicative. In case of a linear relation between the number of sets N and the number of elements M we have

$$\Pr\{a \in A \mid |A| = k \wedge |a| = l\} \sim \frac{const}{N} k \cdot l \quad . \tag{18}$$

To see this observe that

$$\Pr\{a \in A \mid |A| = k \wedge |a| = l\} = 1 - \Pr\left\{ \begin{array}{c} \text{among the } k \text{ choices to generate } A \\ \text{is no virtual } a- \text{ element} \end{array} \right\} \tag{19}$$

$$= 1 - \frac{M^* - l}{M^*} \cdot \frac{M^* - 1 - l}{M^* - 1} \cdot \ldots \cdot \frac{M^* - k - l + 1}{M^* - k + 1} \tag{20}$$

with M^* being the number of virtual elements. The last formula has the same structure as the expression for the pairing probability in the previous section hence we get, for $lk \ll M^*$ and bounded first moments of the ψ-distribution, the claimed multiplicative correlation. We note that there is also a variant of the Molloy-Reed construction which produces an additive size-size correlation such that $\Pr\{a \in A \mid |A| = k \wedge |a| = l\} \sim \frac{const}{N}(k + l)$ holds (see [5] for details of the algorithm).

We next present a simulation-based comparison of the multiplicative and additive Molley-Reed model with the FP4 network. The input size distributions for the Molloy-Reed simulations are the same as in FP4. For completeness we also include the simulation results based on the simple random intersection graph model defined in the previous section. To make clear which size distribution is given in that case we use the notation P-model (O-model) for the intersection graph with fixed P (O) size distribution and denote by PO-model the corresponding Molloy-Reed graphs since both size distributions are fixed therein. Figs. 9 and 10 show the degree distribution for the O- and P-graphs. There is a very good agreement over the whole range of degree values between the real FP4 network projections and typical samples of the multiplicative Molloy-Reed model. This is quite remarkable since a considerable bias from the almost independence of the Molloy-Reed model should be visible in the degree distributions. The fact that there is no

deviation between the degree distributions indicates that the majority of project-organization alignments is essentially a random process. Furthermore, the additive model reproduces the FP4 P-graph degree distribution only well for large degree values indicating that the correlation is indeed multiplicative.

Two quantities measuring local correlations are the triangle-degree dependence and the distribution of edge multiplicity introduced earlier. Fig. 11 compares the triangle-degree correlation for the O-graph. Although the overall picture is similar (linear dependence up to medium degree) there is a clear tendency for higher triangle numbers in FP4 for large degree values. Again the multiplicative version matches better with the data then the additive model. The edge multiplicity—again for the O-graphs—is shown in fig. 12. The real graph has a considerably smaller value in the exponent and extends to almost twice as large a maximal multiplicity value. Nevertheless, both Molloy-Reed models show a sharp scale-free distribution for the multiplicity. This is quite surprising, since, naively, one would expect the probability for positive edge multiplicity to go to zero as N becomes large. In summary, one has a strong agreement between the real data and the multiplicative Molloy-Reed model (the comparison results for FP2 and FP3 are almost identical to the situation with FP4 and have therefore not been depicted here). Only in the fine structure of clustering characteristics are some differences observed.

Finally, we briefly outline why, under certain circumstances, almost independent models like the Molloy-Reed one can have a scale-free edge-multiplicity distribution. To keep the discussion as transparent as possible, we study the question in a pure bipartite Bernoulli model, which can be thought of as a kind of predecessor to the Cameo-model discussed below.

To each vertex from the O- and P- partitions (with cardinality N and M), we assign a power-law distributed, positive integer parameter $\mu(P)$ and $\nu(O)$ with exponents α and β. That is we partition the P- and O-vertices into sets $D_\mu := \#\{P \mid \mu(P) = \mu\}$ and $G_\nu := \#\{O \mid \nu(O) = \nu\}$ such that $|D_\mu| = \frac{C_P M}{\mu^\alpha}$ and $|G_\nu| = \frac{C_O N}{\nu^\beta}$ where C_P and C_O are normalization constants. We further assume $N = C_{op} \cdot M$ and put

$$\Pr\{P \sim O\} := \frac{c}{N}\mu(P)\,\nu(O) \quad . \tag{21}$$

It is easy to see that the expected degree, conditioned on the μ or ν value, is proportional to μ or ν, respectively, and therefore the (bipartite) degree distribution on each partition has the same exponent as μ or ν. Note that the maximal μ and ν values are given by $\mu_{\max} \sim M^{\frac{1}{\alpha}}$ and $\nu_{\max} \sim N^{\frac{1}{\beta}}$.

Since the edge multiplicity in the projection graph corresponds to the number of paths of length 2 in the bipartite graph, we define $E_k^{(P2)} := \mathbb{E}\#\{\ (P,P') :$ there are exactly k paths of length 2 between P and $P'\}$ and $E^{(P2)} := \sum kE_k^{(P)}$. For fixed P and P' with parameters μ and μ' the expected number of paths of lenght 2 between the two vertices is given by

$$\sum_\nu \frac{c^2}{N^2}\mu\mu'\nu^2\,|G_\nu| \tag{22}$$

and therefore the expected total number of 2 paths in the P-partition is

$$E^{(P2)} = \sum_{\mu,\mu'} |D_\mu|\,|D_{\mu'}| \sum_\nu \frac{c^2}{N^2}\mu\mu'\nu^2\,|G_\nu| \tag{23}$$

$$= \sum_{\mu,\mu'}\sum_\nu \frac{C_O C_P^2 M}{C_{op}\,(\mu\mu')^{\alpha-1}\,\nu^{\beta-2}} \quad . \tag{24}$$

On the other hand, we have for the probability of an edge between P and P' in the P-projection graph the estimate

$$\Pr\{P \sim P'\} = 1 - \prod_\nu \left(1 - \frac{c^2}{N^2}\mu\mu'\nu^2\right)^{|G_\nu|} \tag{25}$$

$$\simeq 1 - \exp\left(-\sum_\nu \frac{C_O c^2 \mu\mu'}{C_{op} M \nu^{\beta-2}}\right) \tag{26}$$

and hence for the expected total number of edges E

$$E \simeq \sum_{\mu,\mu'} \frac{C_P^2 M^2}{(\mu\mu')^\alpha}\left(1 - \exp\left(-\sum_\nu \frac{C_O c^2 \mu\mu'}{C_{op} M \nu^{\beta-2}}\right)\right) \quad . \tag{27}$$

Several cases are now possible. For $\beta > 3$ and $\alpha > 2$, it is easy to see that $\lim_{N\to\infty} \frac{E^{(P2)}}{E} = 1$ and higher edge multiplicities have essentially zero probability.

The situation is different if either condition is violated, since in this case $E^{(P2)} - E$ diverges and can become of the same order as E. For instance, we obtain for $\beta < 3, \alpha < 2$

$$E^{(P2)} - E \simeq \sum_{\mu,\mu'}^{\mu_{\max}} \frac{C_P^2 M^2}{(\mu\mu')^\alpha} \sum_{k\geq 2} \frac{(-1)^k}{k!} \left[\sum_{\nu}^{\nu_{\max}} \frac{C_O c^2 \mu\mu'}{C_{op} M \nu^{\beta-2}} \right]^k \tag{28}$$

$$\simeq \sum_{\mu,\mu'}^{\mu_{\max}} \frac{const \cdot M^2}{(\mu\mu')^\alpha} \sum_{k\geq 2} \frac{(-1)^k}{k!} \left[const \cdot \mu\mu' M^{\frac{3}{\beta}-2} \right]^k \tag{29}$$

$$\simeq \sum_{k\geq 2} const \cdot \frac{(-1)^k}{k!} M^{\frac{2}{\alpha} + k\left(\frac{3}{\beta}+\frac{2}{\alpha}-2\right)} \tag{30}$$

From the last formula, we see that the expected edge multiplicity $\frac{E^{(P2)}}{E} - 1$ can become positive for proper choices of α and β. We show that $\frac{E}{E^{(p2)}} < 1$ under the above assumptions. Since

$$E^{(P2)} = \sum_{\mu,\mu'} \sum_{\nu} \frac{C_O C_P^2 M}{C_{op} \left(\mu\mu'\right)^{\alpha-1} \nu^{\beta-2}} \tag{31}$$

$$\simeq const \cdot M^{\frac{1}{\alpha}2(2-\alpha)+1+\frac{1}{\beta}(3-\beta)} \tag{32}$$

$$= const \cdot M^{\frac{4}{\alpha}+\frac{3}{\beta}-2} \tag{33}$$

and

$$E \simeq \sum_{k\geq 1} const \cdot \frac{(-1)^{k+1}}{k!} M^{\frac{2}{\alpha} + k\left(\frac{3}{\beta}+\frac{2}{\alpha}-2\right)} \quad , \tag{34}$$

one gets

$$\frac{E}{E^{(P2)}} \simeq 1 - \sum_{k\geq 2} const \cdot \frac{(-1)^k}{k!} M^{\frac{2(k-1)}{\alpha}+\frac{3(k-1)}{\beta}-2k} \tag{35}$$

$$\simeq 1 - const \cdot M^{-\frac{2}{\alpha}-\frac{3}{\beta}} \left(M^{\frac{2}{\alpha}+\frac{3}{\beta}} - 1 + o\left(1\right) \right) \tag{36}$$

$$= 1 - const + o\left(1\right) \quad . \tag{37}$$

Since the involved constant is positive we get the desired result. A more carefully analysis, which will be part of a forthcoming paper, shows that one also obtains a power law for the edge multiplicity, as observed in the simulations.

C. Random intersection graphs and the "Cameo" principle

In this section, we give a possible explanation for the appearance of power laws in the size distribution. In most models of complex networks with power-law like degree distributions, one assumes a kind of preferential attachment rule as in the Albert and Barabasi model. This makes little sense in our framework. Instead we propose a rule called the "Cameo Principle" first formulated in [2].

Before giving an interpretation and motivation we briefly describe the formal setting. Assign to each project a positive φ distributed random variable (r.v.) ω and to each organization a positive $\psi-$ distributed r.v. μ (note that, in contrast to section IV B, φ and ψ are not the size distributions). We assume φ and ψ to be supported on $(1,\infty)$ and monotone decaying as ω and μ tend to infinity. On the bipartite graph an edge between an organization O and a project P is then formed with probability

$$p_{o,p} := \frac{c_0}{\psi^\alpha\left(P\right)} \cdot \frac{1}{\sum\limits_P \psi^{-\alpha}\left(P\right)} + \frac{c_1}{\varphi^\beta\left(O\right)} \cdot \frac{1}{\sum\limits_O \varphi^{-\beta}\left(O\right)} \quad , \tag{38}$$

where c_0 and c_1 are positive constants, $\alpha, \beta \in (0,1)$, and all edges are drawn independently of one another. We are interested in the properties of the corresponding random P and O-graphs for typical realizations of the ω and μ

variable. The word typical is here understood in the sense of the ergodic theorem, namely we assume $\frac{1}{N}\sum_O \varphi^{-\beta}(O) \sim \int \varphi^{1-\beta} d\varphi =: C_0^{-1}$ and $\frac{1}{M}\sum_P \psi^{-\alpha}(P) \sim \int \psi^{1-\alpha} d\psi =: C_1^{-1}$, where N and M are the cardinalities of the O- and P-partitions and α and β are such that the integral is bounded. The above formula reduces then to

$$p_{o,p} := \frac{c_0 \cdot C_0}{M\psi^{\alpha}(P)} + \frac{c_1 \cdot C_1}{N\varphi^{\beta}(O)} \quad . \tag{39}$$

The expected conditional size of a vertex is then given by

$$\mathbb{E}(|P| \mid \psi(P)) = \frac{Nc_0 \cdot C_0}{C_1 M \cdot \psi^{\alpha}(P)} + c_1 \tag{40}$$

and

$$\mathbb{E}(|O| \mid \varphi(O)) = \frac{Mc_1 \cdot C_1}{C_0 N \cdot \varphi^{\beta}(O)} + c_0 \quad . \tag{41}$$

The interpretation behind the special form of edge probability in eq. (38) is the following. The ω and μ values describe a kind of attractivity property inherent to projects and organizations. Thinking in terms of a virtual project formation process the final set of organizations belonging to a project P can either join the project actively—in which case the μ value of P is important—or the organization more passively enters the project on the request of organizations already involved—in which case the attractivity ω of the the corresponding organization is important. The attractivity of an organization could, for instance, be related to its reputation, financial strength, or quality of earlier projects in which the organization was involved. Extrapolating from human behavior, it is not directly the ω or μ value which enters the pairing probability, but rather the relative frequency of the ω or μ values: the rarer a property, the more attractive it becomes. This is in essence the content of the "Cameo" principle.

The parameters α and β can be seen as a kind of affinity to following the above rule; for $\alpha, \beta \to 0$ the rule is switched off and we recover a classical Erdös-Renyi intersection graph. In general the values of α and β are themselves quenched random variables with their own—usually unknown—distribution. As shown in [4], only the maximal α and β values matter for the resulting degree distribution of the graphs. We therefore restrict ourself in the following to constant values.

Since the conditional expectation of the size values (eqs. (40) and (41)) are proportional to $\varphi^{-\beta}$ and $\psi^{-\alpha}$, we have to estimate their induced distribution. It can be shown [3] that $z := \varphi^{-\beta}(\omega)$ is asymptotically distributed with density $z^{-\left(1+\frac{1}{\alpha}+o(1)\right)}$ when $\varphi(\omega)$ decays monotone and faster than any power law to zero as $\omega \to \infty$. When $\varphi(\omega)$ is itself a power-law distribution with exponent γ, the resulting distribution for z will be $z^{-\left(1+\frac{1}{\alpha}-\frac{1}{\alpha\gamma}+o(1)\right)}$. Therefore, the induced distribution is always a power law and independent of the details of φ. Applying this result to our model, we obtain immediately a power law distribution for the size distribution on the P- and O-graphs with exponents depending essentially only on α and β. It is not difficult to see that, due to the edge independence in the model definition, the resulting degree distributions are again of power-law type. The Cameo *Ansatz* hence generates in a natural way a bipartite graph, where both projections admit two of the main features of the FP-networks. Furthermore, we obtain a linear dependence of the mean triangle number $\triangle_k$ on the degree, as in section IV A.

None of the models discussed in section IV can reproduce scale-free distribution of the edge multiplicity with the same low exponent as observed in each of the FP networks. It will be interesting to see whether the inclusion of memory effects like the "My friends are your friends" principle [6] will change the picture.

V. CONCLUSIONS

In this work, we have described research collaboration networks determined from research projects funded by the European Union. The networks are large in terms of size, complexity, and economic impact. We observed numerous characteristics known from other complex networks, including scale-free degree distribution, small diameter, and high clustering. Using a random intersection-graph model, we were able to reproduce many properties of the actual networks. The empirical and theoretical investigations together shed light on the properties of these complex networks, in particular that the EU-funded R&D networks match well with typical realizations of random graph models characterized by just four parameters: the size, edge number, exponent of project-projection degree distribution, and exponent of organization-projection degree distribution.

In terms of real-world interpretation, the present analysis yields three major insights. First, based on the fact that the size distribution of projects did not change significantly between the Framework Programs, any possible changes

in project formation rules—which we do not know at this stage—did not affect the aggregate structure of the resulting research networks. Second, the fact that integration between collaborating organizations has increased over time, as measured by the average clustering coefficient, indicates that Europe has already been moving towards a more closely integrated European Research Area in the earlier Framework Programs. Finally, the fact that a sizeable number of organizations collaborate more than once in each Frame Program shows that there appears to be a kind of robust backbone structure in place, which may constitute the core of the European Research Area.

In terms of application, the present results suggest a number of extensions. First, it is essential to learn more about the properties of the vertices in our networks. To what extent can they be characterized and classified? What kind of structural patterns emerge if we add this information? Second, we need to know more about the micro-structure of the networks. In which areas are the networks highly clustered and where does this clustering come from? What kind of subgroups can be identified? Third, we need to learn more about where the observed distribution of edge multiplicity comes from. Finally, it would be desirable to explicitly include edge weights into the analysis. Presumable, actors who collaborate more frequently are more proximate to each other than actors who collaborate only once. This may significantly impact the structural features we are able to observe, as well as the conclusions we might draw concerning the link between network structure and function.

Acknowledgments

We would like to acknowledge support from the Portuguese Fundação para a Ciência e a Tecnologia (Bolsa de Investigação SFRH/BPD/9417/2002 and FEDER/POCTI-SFA-1-219), ARC systems research (W4570000294-3), and from the VW Stiftung (I/80496). We thank Ph. Blanchard and L. Streit for useful discussions and commentary. Portions of this work were done at the Vienna Thematic Institute for Complexity and Innovation, EXYSTENCE Network of Excellence: IST-2001-32802.

---

[1] K. Barker and H. Cameron: *European Union science and technology policy, RJV collaboration and competition policy,* in Y. Caloghirou, N.S. Vonortas and S. Ioannides (eds.), *European Collaboration in Research and Development*, Edwar Elgar: Cheltenham, UK and Northampton, MA, US (2004)

[2] Ph. Blanchard, T. Krueger: *The "Cameo principle" and the origin of Scale-free graphs in social networks,* Journal of Statistical Physics, **114**, 5-6 (2004), arXiv: cond-mat/0302611

[3] Ph. Blanchard, T. Krueger: *Networks of the extreme: a search for the exceptional,* to appear in "Extreme Events in Nature and Society", The Frontier Collections, Springer (2005)

[4] Ph. Blanchard, S. Fortunato, T. Krueger: *Importance of extremists for the structure of social networks*, arXiv:cond-mat/0407434 (2004), Phys.Rev.E, **71**, (2005)

[5] Ph. Blanchard, A. Krüger, T. Krueger, P.Martin: *The Epidemics of Corruption* ,submitted to Phys. Rev. E, (2005), arXiv:physics/0505031

[6] Ph. Blanchard, T. Krueger, A. Ruschhaup: *Small world graphs by iterated local edge formation*, Phys. Rev. E, **71** (2005), arXiv: cond-mat/0304563 (2003)

[7] B. Bollobas, J. Riordan: *Mathematical results on scale-free random graphs*, Handbook of graphs and networks, (2003)

[8] M. Karonski, E.R. Scheinerman, K.B. Singer-Cohen: *On random intersection graphs: the subgraph problem,* Combinatorics, Probability and Computing, **8,** (1999)

[9] Newman, M.E.J. : *Scientific collaboration networks: I. Network construction and fundamental results*, Physical Review E, **64** (016131), (2001)

[10] CORDIS (2004): *Projects Database—Advanced and Professional Database Search*, available from http://dbs.cordis.lu/cordis-cgi/ EI?CALLER=EIPROF_EN_PROJ&MODE=N&LANGUAGE=EN&DATABASE=PROJ.

[11] CORDIS(2002): *Proposal evaluation,* available from http://www.cordis.lu/fp5/managment/eval/hp_evaluation.htm

Figures

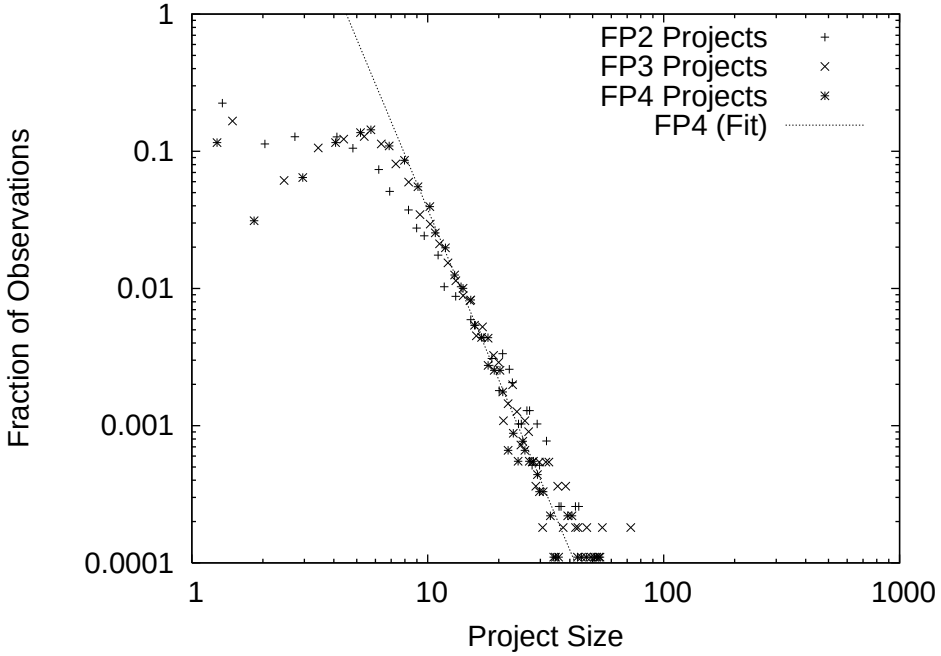


FIG. 1: Distribution of project sizes.

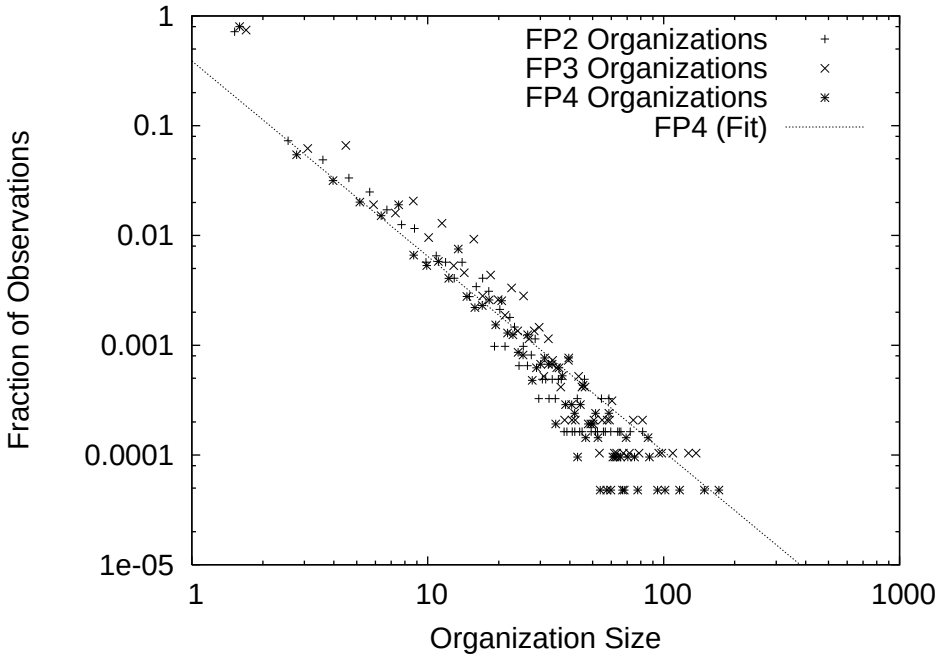


FIG. 2: Distribution of organization sizes.

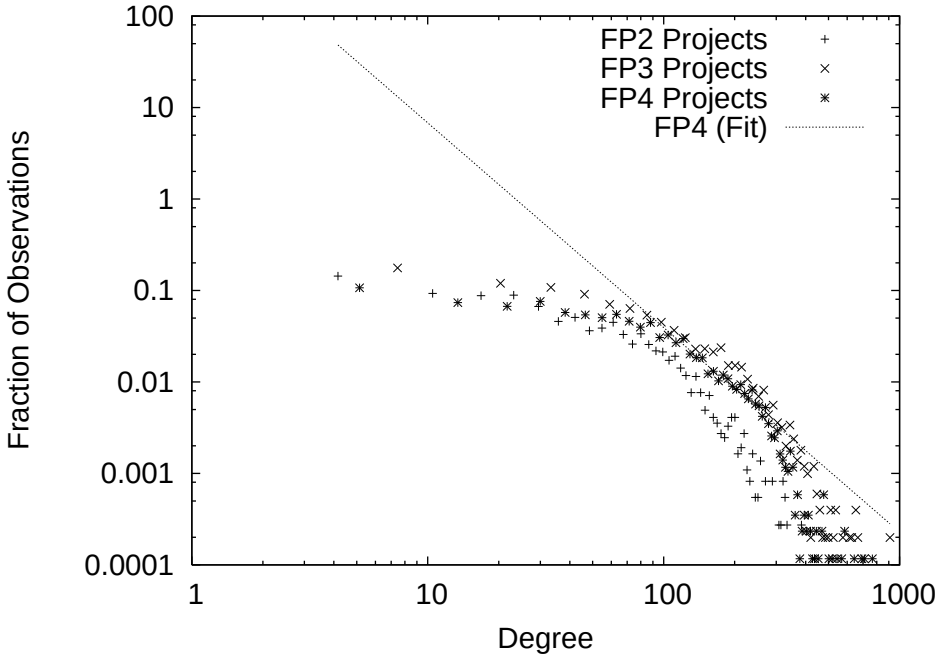


FIG. 3: Degree distribution of projects projection.

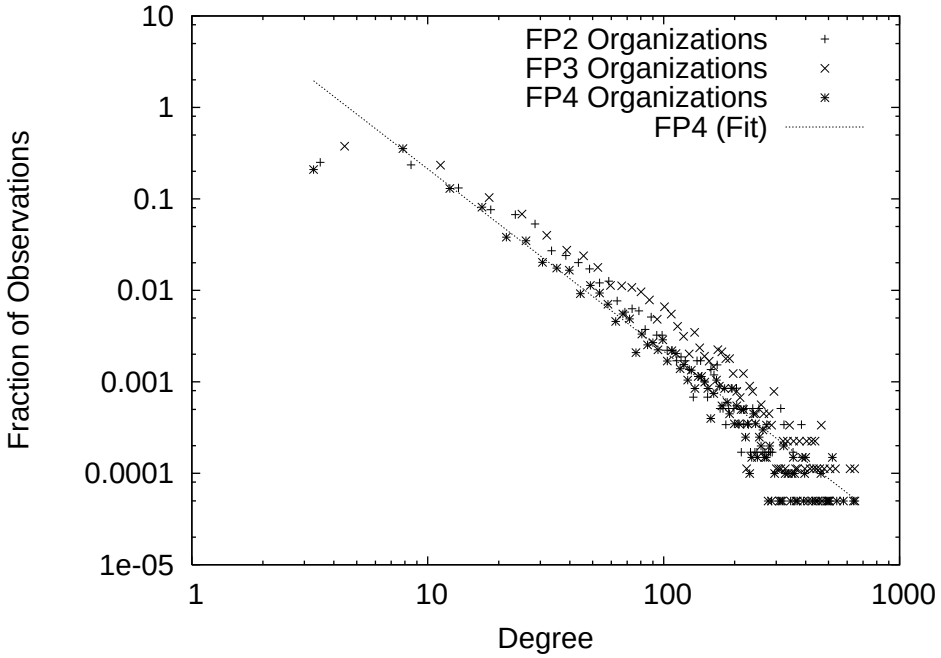


FIG. 4: Degree distribution of organizations projection.

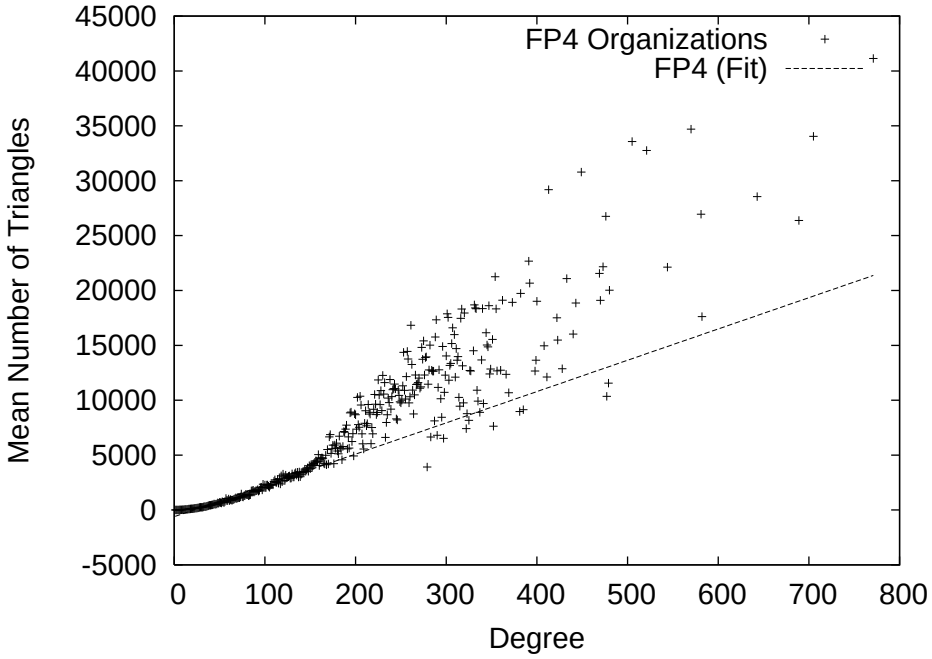


FIG. 5: Relation between degree and number of triangles in the projects projection.

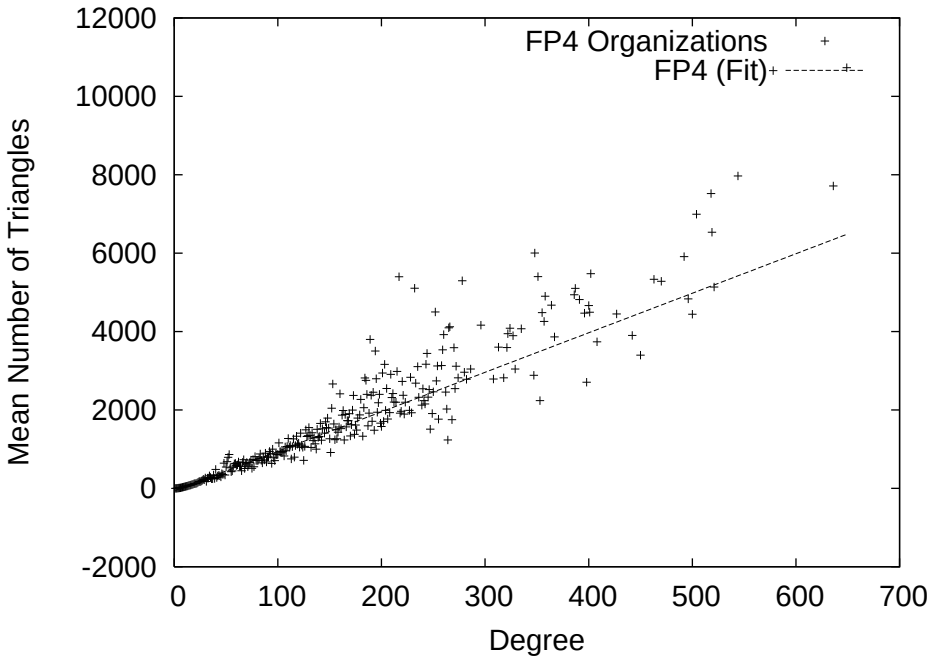


FIG. 6: Relation between degree and number of triangles in the organizations projection.

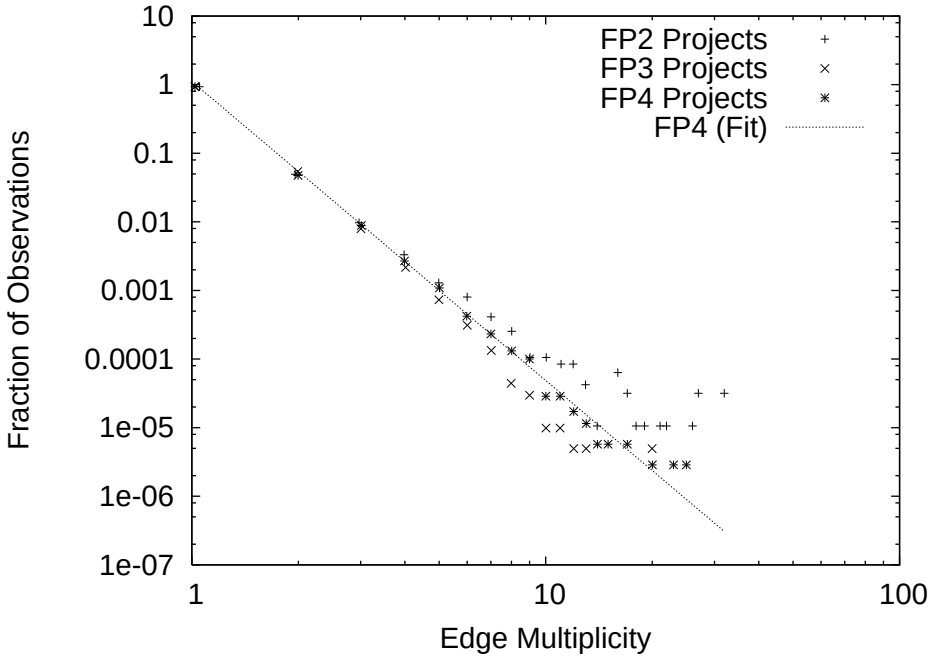


FIG. 7: Distribution of edge multiplicities in the projects projection.

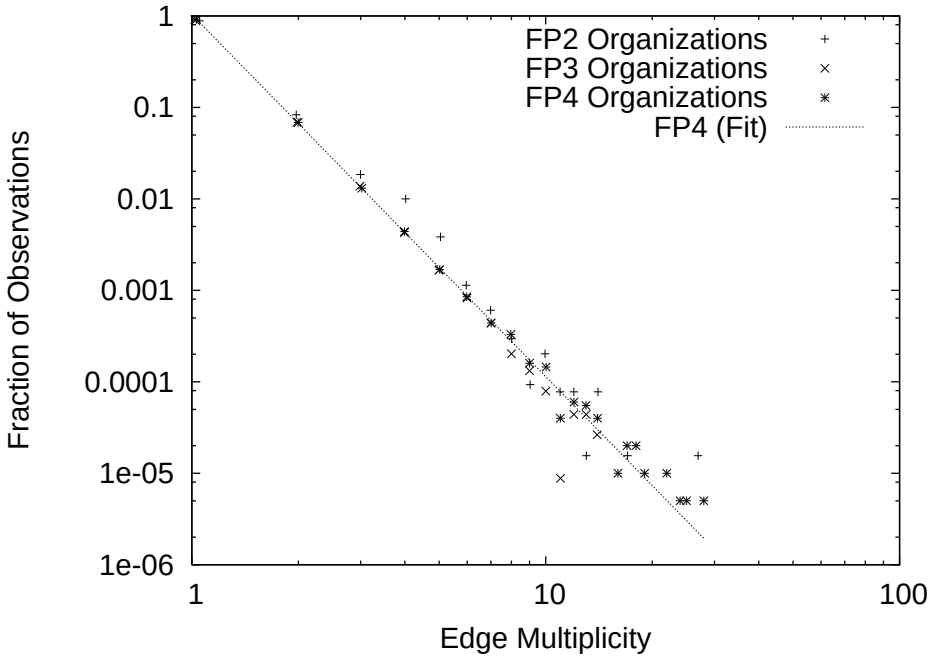


FIG. 8: Distribution of edge multiplicities in the projects projection.

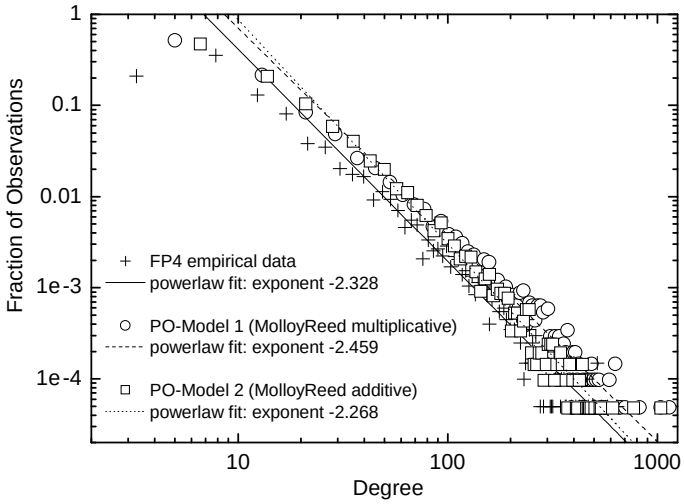


FIG. 9: Degree distribution for the O-graphs.

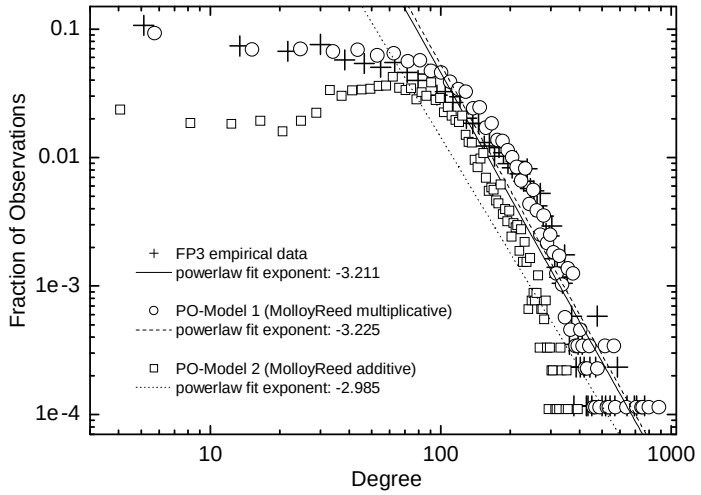


FIG. 10: Degree distribution for the P-graphs.

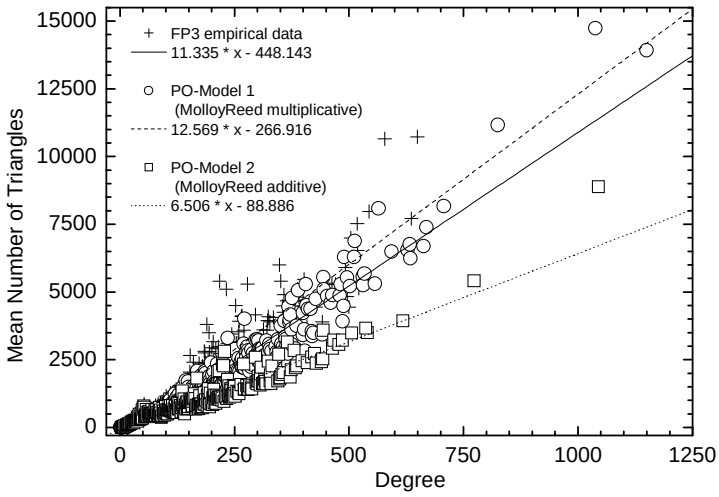


FIG. 11: Triangle-degree correlation for the O-graphs.

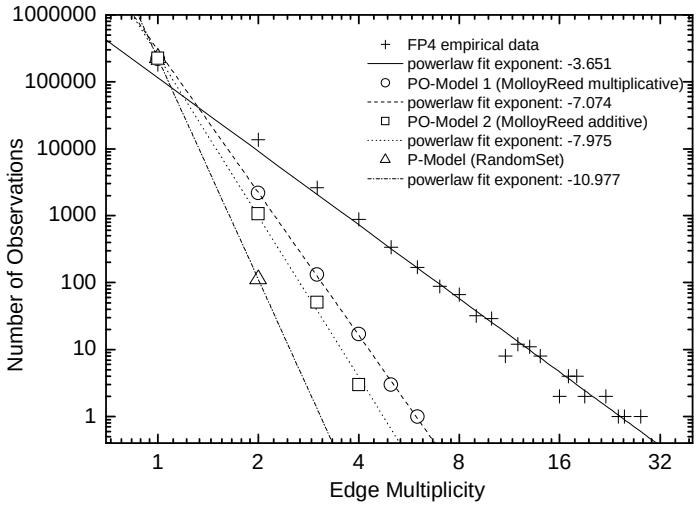


FIG. 12: Edge multiplicity for the O-graphs.

Tables

Framework Program	**budget**[a]	**# P**	**million EUR/P**	**#(P >1)**[b]	**# O**	**million EUR/O**
FP1 (1984–1988)	3.8	3283	1.15	1696	2500	1.52
FP2 (1987–1991)	5.4	3885	1.39	3013	6135	0.88
FP3 (1990–1994)	6.65	5294	1.25	4611	9615	0.69
FP4[c] (1994–1998)	13.3	15061 (9087)	0.88	11374 (8039)	20873	0.64

[a] billion ECU/EUR
[b] projects with more then one participating organization
[c] R&D projects listed in parentheses. The number excludes all projects devoted to preparatory, demonstration, and training activities.

TABLE I: FP1–4 total budget and number of funded projects. The smaller average funding per project and org in FP4 is an artefact as it involves a large number of scholarships and the like, which are smaller than research projects (however, we cannot isolate the bias created).

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.

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.

Corruption as a generalized epidemic process

Ph.Blanchard,* A.Krüger,† and T.Krueger‡

University of Bielefeld, Faculty of Physics and BiBoS

P.Martin§

FU-Berlin, Department of Law

Abstract

We study corruption as a generalized epidemic process on the graph of social relationships. The main difference to classical epidemic processes is the strong nonlinear dependence of the transmission probability on the local density of corruption and the mean field influence of the overall corruption in the society. Network clustering and the degree-degree correlation play an essential role in corruption dynamics. We discuss phase transitions, the influence of the graph structure and the implications for epidemic control. Structural and dynamical arguments are given why strongly hierarchically organized societies like systems with dictatorial tendency are more vulnerable to corruption than democracies. A similar type of modelling can be applied to other social contagion spreading processes like opinion formation, doping usage, social disorders or innovation dynamics.

*Electronic address: blanchard@physik.uni-bielefeld.de

†Electronic address: networks@andreaskrueger.de

‡Electronic address: tkrueger@physik.uni-bielefeld.de

§Electronic address: peter.martin@schulz-berlin.de

I. INTRODUCTION

Corruption seems to be an unavoidable part of human social interaction, prevalent in every society at any time since the very beginning of human history till today. Common sense associates corruption mainly with a deviation from fair play interaction in the development of social relations. Clearly what is meant be fair play depends on the cultural context of a given population/society. This vague description of corruption is in the spirit of sociology and psychology and differs from the narrower corruption concepts usually considered in economics or political sciences. There, corruption is mainly seen as a misuse of public power to gain profit in a more or less illegal way. For the model approach developed in this article we will use the notion of corruption in the more general, first sense. More precisely our intention is to describe changes in mind ranging from damming of corruption as a criminal act to accepting corruption as an attractive option. Therefore in this paper we do not introduce a group of state representatives or officials since we assume that the essential changes in mind which allow corrupt acts happen long before an individual is in the position to act corruptly. Empirical investigations about motives and "typology" of corrupt actors (see [3] for results from case studies in Germany) have shown that the majority of individuals involved in corruption affairs are highly educated, well positioned with respect to social status and do not think to have done something wrong, indicating the importance of mind changes prior to corrupt acts.

In sharp contrast to the high prevalence of corruption in many countries and the rather large literature on political, social and economical aspects of corruption there is only a small number of attempts to model the dynamics of corruption in a mathematically quantified way. The modelling approach in these few attempts essentially follows two paths. The first is in the sense of microeconomics and incorporates game theoretic aspects (for a recent model in this direction see the book by Steinrücken [20] and the references therein) or rules for maximizing a certain economically based profit functional ([19][11]). Then a set of differential equations for the evolution of the mean corruption is derived and a stability analysis done on that basis. In these models one usually makes rather detailed assumptions about the underlying organization structure on which the individuals interact. The second line of approach is more in the sense of cellular automata (CA) models with rather simple state variables and local interaction dynamics. For example in the article by Wirl [22]

a simple 1-dimensional deterministic cellular lattice automata model is used to describe the propagation of corruption. Nevertheless, as is well known in CA-modelling, the global dynamical picture can be highly complex and nontrivial.

Up to now all these attempts did not take into account the complex network of social relationships as the underlying structure for the spread of corruption. In this article we will present a model for the spread of corruption on complex networks in the spirit of epidemiology. The model describes aspects of the evolution of corruption in a virtual population and incorporates some basic universal features of corruption. The local interaction dynamics of the model is similar to probabilistic cellular automata but "lives" not on a lattice type graph like most of the CA-models but on complex networks.

Considering corruption as a nonstandard epidemic process relies on the plausible assumption that corruption rarely emerges out of nothing but is usually related to some already corrupt environment which may "infect" susceptibles. Of course the spontaneous decision of somebody to act corruptly is possible and can easy be handled in the model as an external weak source of infection. One of the very special features in corruption propagation which differs from what is used in describing classical epidemic processes is the threshold like dependence of the local transition probabilities. By this we mean that a noncorrupt individual gets infected with high probability if the number of corrupt individuals in the group of his direct social contacts (encoded as the set of neighbors in a "friendship" or acquaintance graph) exceeds a certain threshold number. Otherwise if the number of corrupt individuals in somebodies social neighborhood is below that threshold value there is only a small probability to get corrupt via such "local" interactions. The second main difference to classical epidemic processes is the mean field dependence of the corruption process. By this we mean that an individual can get corrupt just because there is a high prevalence (or believed prevalence) in the society even when there is no corruption in the local neighborhood. There is another interesting mean field term entering the game, namely the society strikes back to corruption with an efficiency proportional to the fraction of the noncorrupt people. Both mean field terms are nonlinear and together with the local propagation mechanisms they give rise to a rather complex dynamical picture.

There is a notorious problem in finding good empirical data which would allow to estimate the real prevalence of corruption. Probably the greatest effort over the last years to measure the degree of corruption in various countries was made by "Transparency International"

(TA), a non profit group of individuals and organizations which are highly concerned by the lack of sound data. Since 1995 they publish a yearly corruption report and a so called Corruption Perception Index (CPI) [21].

It is not our aim to explain the values of the CPI or other corruption data sets, since this would require a semirealistic modelling of the social and economical structure of individual countries which is completely illusionary at the present stage of research. Rather we want to demonstrate which scenarios are dynamically possible and whether there are interesting phase transitions.

The paper is structured as follows. In section 2 we present the description of the model in a fairly general way- that means the underlying graph and the parameter values are not yet specified. In section 3 we start the numerical investigation of phase transitions and their dependence on the parameters for the corruption process. In section 5 we investigate phase transitions for the mean field infection process and it's interaction with the local processes. Section 6 is devoted to numerical results about the time evolution of the corruption process, showing the diversity of the model. Finally in section 7 we discuss some conclusions for the prevention of corruption and give an outlook to further work.

II. CORRUPTION AS A GENERALIZED EPIDEMIC PROCESS

In this section we first describe the basic setting for our model structure. Refinements and more detailed aspects will be discussed later on. Due to the common view, corruption is first of all a property of the relations between individuals irrespective which definition of corruption one uses. Since an act of corruption requires that at least one of the participants in a corrupt relation has a mental state which tolerates or even assigns a positive value to (his personal view of) corruption we will focus mainly on the spread of this mental state change (from not accepting to accepting corrupt acts as an option for one's own activities). Therefore to discuss corruption as an epidemic process in the afore mentioned sense it is useful to assign a corruption property to the individuals themselves. In the simplest case we just have a time dependent $0-1$ state variable $\omega(x,t)$ assigned to each individual, encoding whether the vertex is corrupt (1) or not (0) at time t. Of course more refined scales for the degree of corruption are possible like additional states for wether an individual is actively corrupt or which type and strength of corruption is considered. Since we are in this

paper mostly interested in phase transitions and the principal structures of the dynamical evolution one can reduce many of these refinements to the present case (at least in the sense of obtaining upper and lower bounds for phase transitions). The underlying structure on which corruption spreads is a given finite graph G with fixed vertex set $V = \{1, ..., n\}$. We consider in this article only stationary graphs with no changes in time on the underlying graph structure. The dynamics is specified by conditional transition probabilities $(p_{ij}(x))$ which depend only on the states on $B_1(x) = \{y : d(x,y) \leq 1\}$ and a meanfield term reflecting the influence of the total prevalence of corruption in the society. Here $d(.,.)$ is the usual graph metric on G and $d(x)$ is the degree of x. We define $b_t := \frac{1}{N} \sum_{y \in V} \omega(y,t)$ as the density of corruption at time t. The standing assumptions on $(p_{ij}(x))$ are the following:

$$p_{01}(x) = \Pr\{\omega(x,t+1) = 1 \mid \omega(x,t) = 0\} = \min\left(1, f_x\left(\sum_{y \sim x} \omega(y,t)\right) + \beta(x) \cdot b_t^2\right) \tag{1}$$

$$p_{10}(x) = \Pr\{\omega(x,t+1) = 0 \mid \omega(x,t) = 1\} = \gamma(x) \cdot [1 - b_t] \tag{2}$$

with $\beta(x) \geq 0$ and $\gamma(x) \in [0,1]$

. In other words the probability to become corrupt depends only on the local prevalence of corruption among the neighbors and the mean corruption in the society and individuals who became corrupt can cure from corruption with a rate proportional to the density of the noncorrupt individuals in the society. The reason behind the quadratic dependence of the mean field term on the density $(\beta(x) \cdot b_t^2)$ is the following. First, corruption becomes attractive as more individuals are corrupt and in the simplest case this attractiveness is proportional to the density of corrupt individuals b_t. Second, a person has to overcome some fear to get uncovered and if fear is proportional to $1 - b_t$ we get a term $const \cdot (1 - (1 - b_t))$ for the probability of neglecting the fear. By assuming further that attraction and fear are independent we obtain the quadratic dependence on the density in formula 1.

In classical i.i.d. epidemics one would have the following functional dependence for the local part of the conditional probabilities: $f(k) = 1 - (1 - \varepsilon)^k \simeq \varepsilon k$ (for ε sufficiently small). For the corruption process the function f_x is more like in voter models, that is below a critical value $\Delta(x)$ of the number of corrupt individuals in $B_1(x)$ the value of f_x is close to zero and above $\Delta(x)$ it is a number $\alpha(x)$ much larger then zero. We will show in section IV that due to this property local clustering can force the epidemics to spread whereas in classical

process name	*characteristic*	*typical value*
α - process	the local transmission process for # of corrupt neighbors $\geq \Delta$	$\alpha >> \varepsilon, \beta, \gamma$
β - process	the mean field transmission process due to the total prevalence or perception of corruption	$\varepsilon < \beta < \gamma$
γ - process	the corruption recover/elimination process due to the fight of the society against corruption	$\beta \leq \gamma < \alpha$
ε - process	the classical local epidemic process for # of corrupt neighbors $< \Delta$	$\varepsilon << \alpha, \beta, \gamma$

TABLE I: the different processes for the corruption dynamics

epidemic processes high clustering slows down the spread of an infection due to reinfection of the already infected (more precisely in corruption processes the gain in spread due to high clustering is dominating the slow-down due to reinfection). In this paper we only study the case where f is a vertex independent, fixed threshold infection function. That means there is a $\Delta > 1$ such that $f(i) = \varepsilon$ for $0 < i < \Delta$ and $f(i) = \alpha \gg \varepsilon$ for $i \geq \Delta$ where ε and α are a priory chosen parameters. Note the difference to the classical voter-type infection function where f_x depends on the degree $d(x)$ of a vertex x (typically one has $f_x(i) = \varepsilon$ for $i < \left\lfloor \frac{1}{2} d(x) \right\rfloor$ and $f(i,x) = \alpha$ for $i \geq \left\lfloor \frac{1}{2} d(x) \right\rfloor$). In general a threshold proportional to the degree seems not appropriate for corruption modelling since this would imply that hubs (high degree vertices) are more immune to infection than low degree vertices contrary to real life experience.

To distinguish between the different ways in which an individual can become corrupt we will speak about the $\alpha, \beta, \varepsilon$ or $\gamma-$ process. For convenience of the reader we give in tabular 1 a summary of the different processes.

Note that in contrast to standard voter models we do not have the possibility of a locally induced backflip from the corrupt state to the noncorrupt. Generalizations of classical epidemic dynamics to processes with a local threshold have recently also been studied in the context of models of contagion (see [10] and references therein) but not yet been mixed with global mean field processes and not specifically adapted to the corruption topic.

We think that our model captures several essential parts of the real corruption transmis-

sion process, namely the local threshold dependence and the mean field influence. Although the "Ansatz" we have chosen can as well be applied to the spread of other social "infection" phenomena e.g. political opinion formation, the concrete assumptions on the transmission functional and the range of the parameters is specific for corruption.

III. PHASE TRANSITIONS IN THE CORRUPTION PROCESS- NUMERICAL RESULTS

In this section we want to look at some threshold properties of the corruption process as a function of either process parameters like Δ or graph properties like edge density, local clustering or scale - freeness.

One of the remarkable differences between a classical epidemic process and a process based on local threshold dynamics is the dependence on the initial number of "infected" vertices in the latter case. In classical epidemics an epidemic process is either overcritical (reproduction number $R_0 > 1$) and a single initial infected vertex infects with positive probability a positive fraction of the whole population, or the process is below criticality ($R_0 < 1$) and all infected will die out respectively become healthy in relatively short time. In corruption epidemics both parts - the mean field process as well the local $\alpha-$ process - can have phase transitions with respect to the initial number of corrupt vertices. That means, there is critical initial density of corrupt vertices b_0^c such that for initial densities below b_0^c the number of infected stays as it is or goes down to zero. Above b_0^c the entire population becomes corrupt with high probability. In Fig. 1 we show the dependence of b_0^c on the edge density $\frac{M}{N}$ on a classical random graph space $\mathcal{G}(N, M)$ with N vertices and M edges. The initial infected vertices are always randomly chosen from the vertex set. There is a clear linear relation between edge density and b_0^c. The jump to $b_0^c \approx 0$ above an edge density of 4.25 is due to the effect that the infection rate due to the classical epidemic $\varepsilon-$ process dominates already the $\gamma-$ process and is therefore overcritical.

As already mentioned in section II one expects that the presence of clustering (that is the number of triangles is at least proportional to the number of vertices) decreases the critical density b_0^c since the α- process can propagate locally more easy. In Fig. 2 the effect of the increase of the triangle number is clearly to see. Here we used a modification of the $\mathcal{G}(N, M)$ random graph space where randomly triangles are added in such a way that the

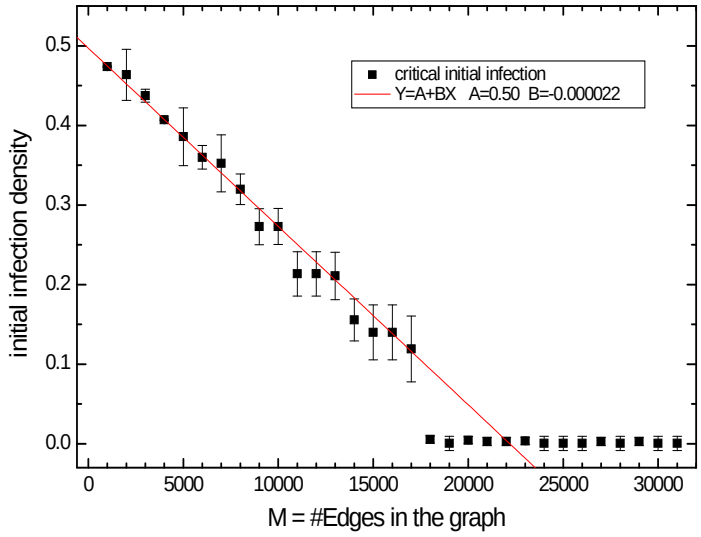


FIG. 1: numerical estimation of the critical initial density b_0^c as a function of the number of edges M for the following parameter values: $\Delta = 5; \alpha = 0.35; \beta = 0.08; \gamma = 0.04; \varepsilon = 0.005; N = 4000$ .Vertical segments are errorbars over 20 runs.

total number of edges remains constant. The threshold value Δ was chosen to be 2 since for higher values of Δ one should add higher order complete subgraphs to see a compatible effect.

The next figure (Fig. 3) shows the dependence of the critical density on Δ for $\mathcal{G}\,(N = 1500, M = 5000)$ The two curves represent the threshold values for an end-prevalence of 10 respectively 90 percent. Since the mean degree in this simulation is about 6.5 one has a vanishing contribution of the α- process above $\Delta = 7$. The critical threshold b_0^c stays than essentially at a value given by the mean field process (for analytic estimations of the critical densities for the pure $\beta -$ process see section V).

To get an impression of the contribution of the different kind of processes (local α and ε, global β and γ - process) to the end-prevalence we give in Fig. 4 the accumulated number of state changes caused by each of the subprocesses till saturation. For small values of Δ ($\Delta \leq 5$) the α- process dominates all others.

Finally we want to look at the dependence of the critical initial density from the exponent in graphs with a power law degree distribution. In Fig.5 we give numerical results for the

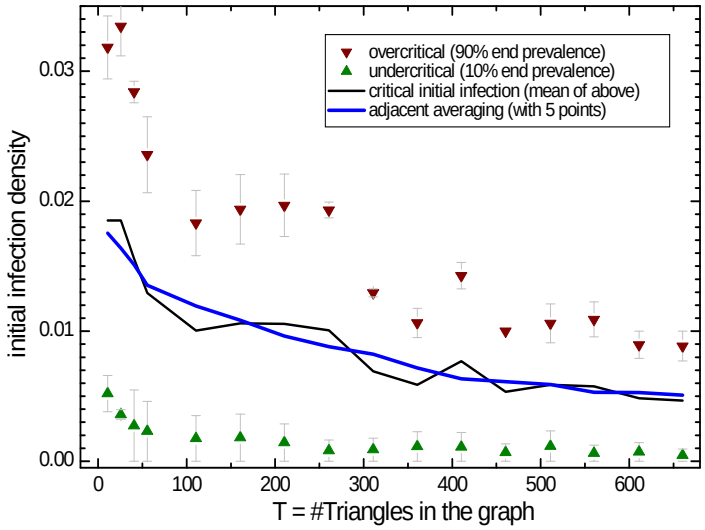


FIG. 2: Critical density b_0^c versus triangle density for the parameter values: $\Delta = 2; \alpha = 0.3; \beta = 0.08; \gamma = 0.04; \varepsilon = 0.005; N = 1000; M = 2000$.

relation between the critical density b_0^c and the exponent λ while keeping the edge density fixed. There is a clear phase transition around $\lambda \sim 2.3$ for $\Delta = 5$.

The explanation of this observation is closely related to a structural phase transition in scale free random graphs at $\lambda = 3$ - namely that for most vertices x an asymptotically positive fraction of all vertices has bounded distance to x. To link this property with the α - process one has to look more closely on the degree-degree correlation in scale-free graphs. Depending on the choice of the model one can have very different correlations like:

$$\Pr\{x \sim y \mid d(x) = k \wedge d(y) = k'\} \simeq const \cdot \frac{k + k'}{N} \text{ or} \tag{3}$$

$$\Pr\{x \sim y \mid d(x) = k \wedge d(y) = k'\} \simeq const \cdot \frac{k \cdot k'}{N} \tag{4}$$

. Formula 3 holds for instance for the Cameo - model ([5]) whereas formula 4 is valid for scale-free graphs generated via the Molloy&Reed algorithm (the later one represents the random graph space containing all graphs with a given degree distribution equipped with the uniform measure and was used for the simulations of the multiplicative case in Fig.5). Evolutionary graphs like in the Albert&Barabasi model have usually more complicated correlations. To

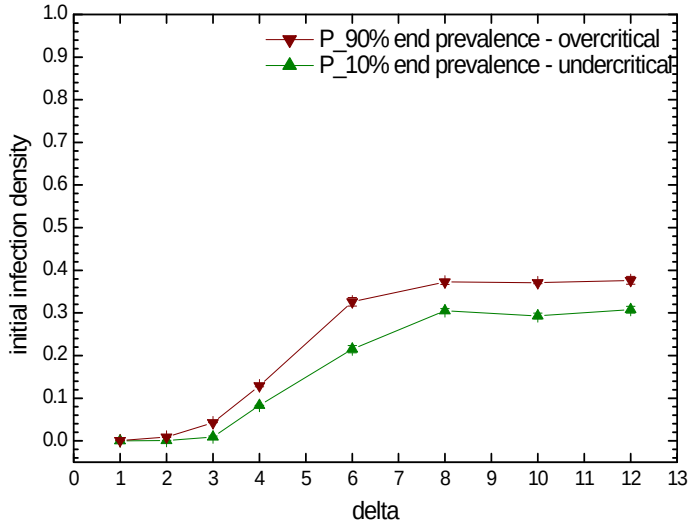


FIG. 3: Lower and upper bounds for the critical density b_0^c as a function of Δ for the following parameter values: $N = 1500; M = 5000; \alpha = 0.35; \beta = 0.08; \gamma = 0.04; \varepsilon = 0.005$

compare with the multiplicative case we have in Fig.5 chosen the same parameters and degree-distribution for the additive case. There is a clear increase of b_0^c to observe but, although unlikely, it remains open wether there is a vanishing threshold in the limit $N \to \infty$. For intermediate couplings we still expect $b_0^c(N) \to 0$ as N diverges for $\lambda < \lambda_c \in (2, 3)$ where λ_c depends on the concrete model. It is remarkable that low λ and a tendency to multiplicative correlation is mainly expected to hold in societies with strong hierarchical structures of social dependencies e.g. dictatorships (see [9] for details), whereas democracies are characterized by less strong degree-correlation.

We close this section by presenting a numerical result showing the different contributions to the overall infection (end-prevalence) of the local and mean field processes as a function of the edge density in the random graph space $\mathcal{G}(N, M)$. Fig.6 gives the accumulated number of state changes (divided by N) caused by the α, β, γ and ε- process at initial density values slightly above the critical one.

.

Up to an edge density of 2 (corresponding to a mean degree of 4) the $\beta-$ process gives the major contribution to the end prevalence in the overcritical situation. Parallel to the

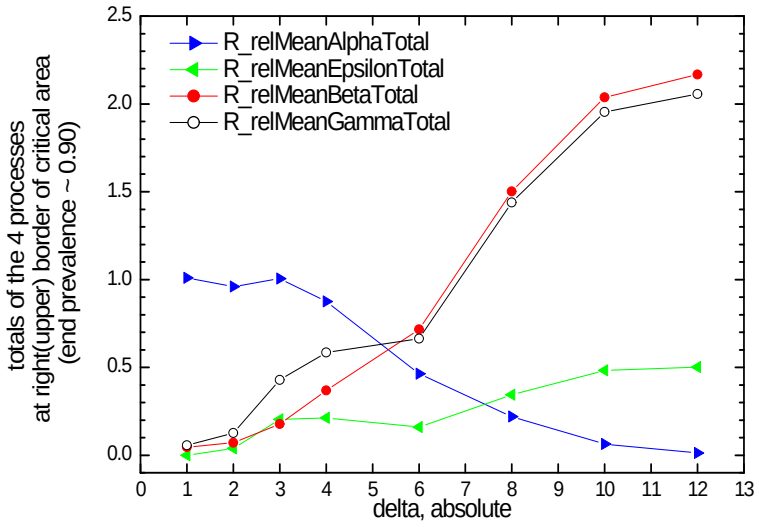


FIG. 4: Total number of state changes splitted according to the different subprocesses as a function of Δ for the same parameter values as in Fig.3.

increase in the edge density increases the contribution of the α- and ε- process (in the intermediate phase of density between 2 and 3 dominated by the α- process) till a sharp peak at edge density 4.5 where the $\varepsilon-$ process outperforms all the others (at the same time the critical initial corruption density b_0^c drops down and becomes almost zero). The peak is easy to understand since for the chosen parameters we have at an edge density of 4 an equality between the recover rate γ and the expected number of new corruptions caused by a single corrupt vertex via the $\varepsilon-$ process (which is $\mathbb{E}\left(d\left(x\right)\right)\cdot\varepsilon$). In terms of classical epidemic processes this corresponds to the case of reproduction number $R_0 = 1$. Above this value single initial corrupt vertex is already enough to cause in conjunction with the mean field process a total infection of the network.

IV. PHASE TRANSITIONS FOR THE $\alpha-$ PROCESS

In the previous chapter we have numerically studied the dependence of b_0^c on the network structure and the local threshold parameter Δ. In this chapter we will turn to more theoretical considerations about the $\alpha-$ process. First we will give a theoretical outline of the

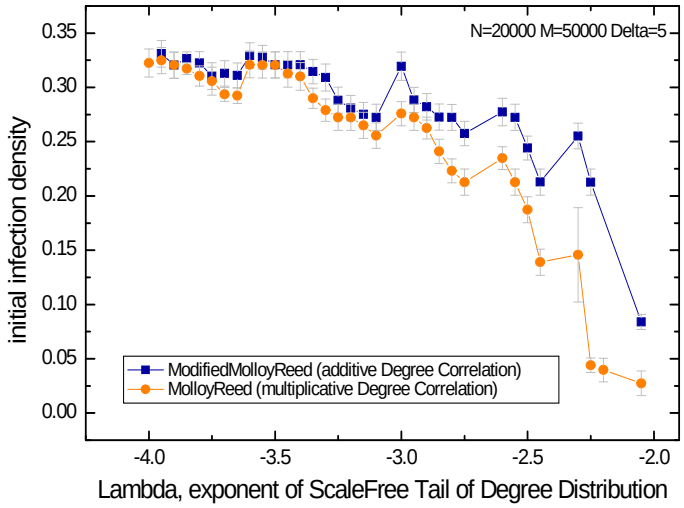


FIG. 5: b_0^c as a function of the exponet λ in an additive and a multiplicative scale-free degree distribution with parameters: $N = 20000; M = 50000; \Delta = 5; \alpha = 0.35; \beta = 0.08; \gamma = 0.04; \varepsilon = 0$

different threshold behavior in the case of additive and multiplicative degree correlations. Second we study the effect of clustering on the threshold value for several types of infinite tree $-$ like structures.

Since a rigorous mathematical analysis of the $\alpha-$ process is beyond the scope of this paper we just give a heuristic outline why in scale free graphs with a multiplicative degree - correlation (as in formula 4) the threshold density b_0^c tends to zero as $N \to \infty$ for exponents $\lambda < 3$ (note that for classical epidemic processes there is absence of an epidemic threshold in scale free graphs with exponent $\lambda < 3$ irrespective of the degree correlation). For fixed initial infection density $b_0 > N^{\frac{1}{\lambda}-\nu}$ and $\frac{1}{\lambda} > \nu > 0$ (note that the typical maximal degree is about $N^{\frac{1}{\lambda}}$) it is obvious that vertices x with $d(x) \geq k_0 >> \frac{\Delta}{b_0}$ get almost surely infected (as $N \to \infty$) via the α - process as soon as $\gamma < \alpha$. Let A_{k_0} be the set of such vertices. One the other side it follows from 4 that a vertex y with $d(y) = k < k_0$ is linked to the set A_{k_0} with

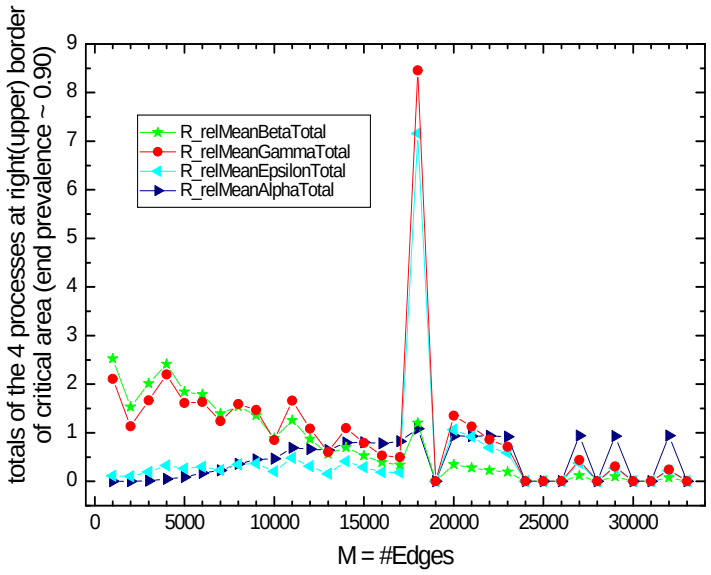


FIG. 6: The total number of state changes splitted according to the α, β, γ and ε- process for the following parameter values in a $\mathcal{G}(N, M)$ graph: $N = 4000; \alpha = 0.35; \beta = 0.08; \gamma = 0.04; \varepsilon = 0.005; \Delta = 5$.

probability

$$q_k \sim 1 - \prod_{k' \geq k_0}^{k_{\max} \sim N^{\frac{1}{\lambda}}} \left(1 - \frac{const \cdot k \cdot k'}{N}\right)^{const \cdot \frac{N}{(k')^\lambda}} \tag{5}$$

$$\sim 1 - e^{-const \cdot \frac{k}{N} \sum\limits_{k' \geq k_0} N \cdot \frac{k'}{(k')^\lambda}} \sim 1 - e^{-const \cdot k \frac{1}{k_0^{\lambda-2}}} \tag{6}$$

. Since q_k is close to 1 for $k > k_0^{\lambda-2}$ one has an almost sure multiple linkage of vertices y with $d(y) > k_0^{\lambda-2} < k_0$ to the set A_{k_0}. These vertices get now again infected via the α - process. By iterating this procedure one may arrive at a positive N- independent infection density $b_t >> b_0$ such that the β - process is overcritical and finally the whole vertex set becomes corrupt. The mechanism described requires N to be large and therefore we conjecture that the difference to the numerical results depicted in Fig.5 (phase transition at $\lambda < 2.3$ instead of 3) is due to finite size effects. In the case of $\Delta = 2$ the finite size effects are smaller and the phase transition is more close to 3. A similar kind of arguments shows, that the expected path-length is finite for $\lambda < 3$. Namely since the expected number S_l of vertices at distance l

from a vertex x with degree k_0 is approximately given by $(const)^l \cdot \sum_{k_1,\ldots,k_l}^{N^{\frac{1}{\lambda}}} \frac{k_o \cdot k_1 \cdot k_1 \cdot \ldots \cdot k_{l-1} \cdot k_{l-1} \cdot k_l}{(k_1)^\lambda \cdot \ldots \cdot (k_{l-1})^\lambda} \sim const \cdot k_0 \cdot N^{\frac{(l-1)(3-\lambda)}{\gamma}}$ for $\lambda < 3$ (note that this expression is only valid for l s.t. $\frac{(l-1)(3-\lambda)}{\gamma} \cdot \log k_0 < 1$). The essential diameter $diam_e$ (a large fraction if the whole vertex set is within a ball of diameter $diam_e$) is then given by the smallest l such that $\frac{(l-1)(3-\lambda)}{\gamma} > 1$ (for a more extensive discussion of the notion of essential diameter see [6]). For $\lambda = 2$ one obtains therefore $diam_e = 3$. For $\lambda \geq 3$ the essential diameter is no longer bounded but growths logarithmically in N. It is interesting that the jump in the critical density at 2.3 in Fig.5 coincides with a jump in diameter from 4 to 5. A small essential diameter can have fatal consequences for corruption epidemics since most vertices are closely linked to hubs and, as was outlined above, hubs are with high probability corrupt. A precise estimation of the dependence of b_0^c from N, M and λ requires a careful discussion of the involved constants. For scale-free graphs with additive degree correlation like Cameo-graphs one still has a bounded essential diameter for exponents less than 3. But the first argument about chains of almost sure linkages from high degree to low degree vertex sets can not be adopted. One expects therefore a higher value of the critical density b_0^c. This is also supported by the numerical results from the previous section.

As already mentioned one of the main differences between corruption epidemics and classical epidemics is the different effect of clustering on the epidemic threshold. In the classical situation any epidemics will be slowed down by the presence of local cycles due to the high probability of reinfection. In corruption epidemics local clustering may speed up the propagation of corruption due to the nonlinear dependence of the infection probability on the number of infected neighbors. In the following we will give two of examples where the strength of this effect can be analyzed and where the critical infection density can be explicitly computed. The first one is a regular infinite tree of degree 4 where of course no triangles are present (see Fig. 7).

. The second structure is a regular infinite graph of again of degree 4 with positive local cluster coefficient ($A(x) = 2$) and a global tree-like structure (see Fig.8).

. Infinite (or large) tree-like structures are important to understand since the local picture around a typical vertex in a reasonable random graph looks tree-like (a classical tree in the absence of clustering or a "fattened" tree in the case of local clustering). In both cases an exact computation of the critical infection density is possible. We give a short outline for the

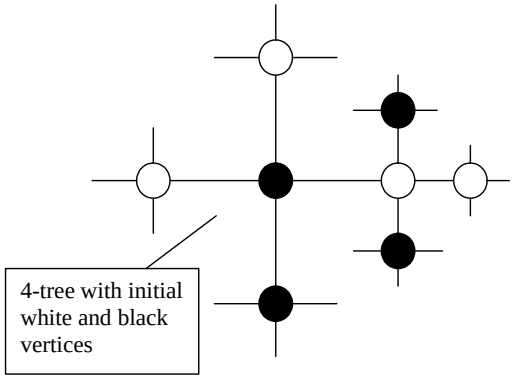


FIG. 7: Segment of a regular infinite tree of order 4

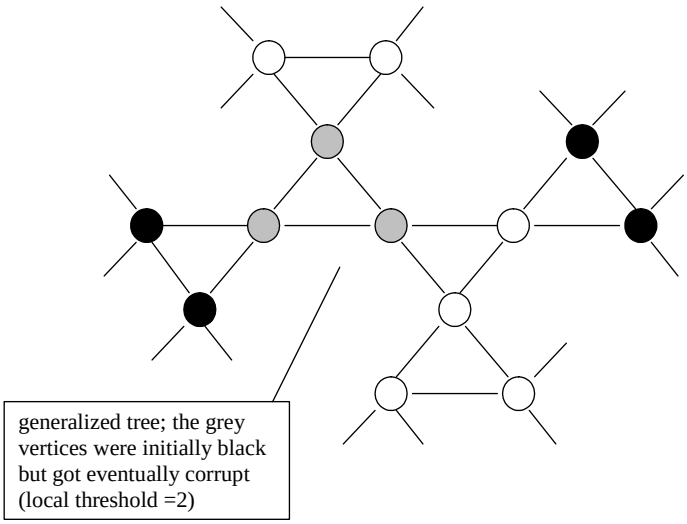


FIG. 8: Segment of an infinite generalized tree (degree 4 and branching number 3)

case of threshold value $\Delta = 2$ and $\alpha = 1$ (the case $\alpha < 1$ requires more lengthy computations but can be done in a similar fashion) and start with the case of the regular 4 tree. A random initial configuration is given by marking each vertex with probability p as noncorrupt (black) and with probability $1 - p$ as corrupt (white). We ask for the critical probability p_c such that for $p < p_c$ almost surely the entire tree becomes white (corrupt) and for $p > p_c$ there

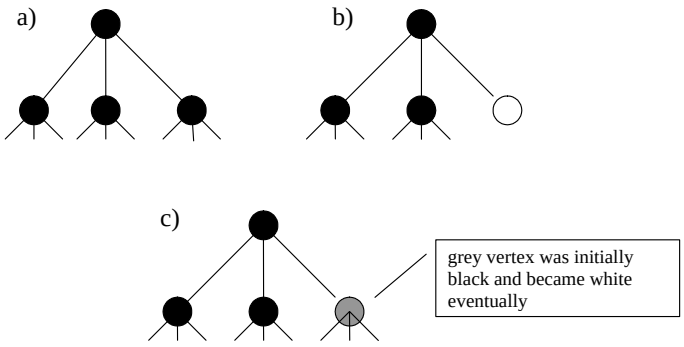


FIG. 9: Different configurations in the neighborhood of the root vertex.. Black denotes vertices in an immune cluster and grey an initial black vertex which became white.

remains an infinite cluster of noncorrupt (black) vertices with probability one. Note that no finite cluster of black vertices -that is a finite black subgraph surrounded by white vertices- can survive so there are either infinite black clusters or none. This property actually holds for all values of $\Delta \geq 2$ on trees with degree larger Δ. We call an invariant infinite black cluster immune. Since $\Delta = 2$ any vertex in an immune cluster must have at least three black neighbors from that cluster. Denote by $T_R(3)$ the rooted tree with outdegree 3 (fixing a root gives a canonical direction to the edges of the tree so it makes sense to speak about the outdegree of a vertex). Every vertex has degree 4 except the root which has degree 3. Let x be the $p-$ dependent probability that the root is contained in an immune cluster (as a subgraph of $T_R(3)$) conditioned that the root vertex is initially black. By arguments from the general theory of branching processes x equals the largest solution of the following recursion equation

$$x = \underbrace{p^3 x^3}_{a)} + \underbrace{3p^3 x^2 (1-x)}_{b)} + \underbrace{3p^2 (1-p) x^2}_{c)} \tag{7}$$

. Figure 9 displays the different situations which enter the above equation. The solutions are $\frac{1}{2p^3}\left(\frac{3}{2}p^2 \pm \frac{1}{2}\sqrt{-8p^3+9p^4}\right)$ and 0. Since $-8p^3+9p^4 \geq 0$ is needed to have a positive nonzero solution we get for the critical probability $p_c = \frac{8}{9} \simeq 0.888\,89$.

In a similar fashion one can derive a recursion equation for the generalized tree case. For that let $T_R(2,1)$ be the rooted generalized tree shown in Fig.10. To every vertex is attached an outgoing triangle, hence the degree of a vertex is 4 except the root which has degree 2.

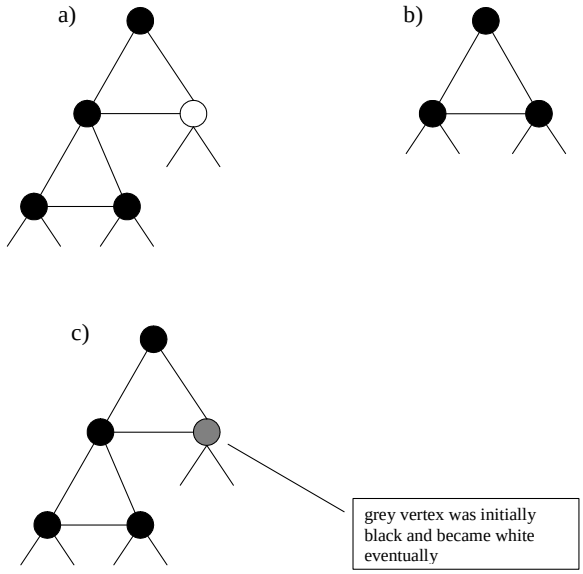


FIG. 10: The local picture around the root vertex in $T_R(2,1)$

To settle the question about p_c for the original generalized tree it is enough to analyze the corresponding problem for $T_R(2,1)$. Again let x be the probability that the root vertex is in an immune cluster conditioned that the root is initially black. One gets the following recursion equation

$$x = \underbrace{p^2x^2}_{b)} + \underbrace{2p^4x^2(1-x)}_{c)} + \underbrace{2p(1-p)p^2x^2}_{a)} \tag{8}$$

(see Fig.10). The solutions are $\frac{1}{2p^4}\left(\frac{1}{2}p^2 + p^3 \pm \frac{1}{2}\sqrt{-7p^4+4p^5+4p^6}\right)$ and 0. Again since $-7p^4+4p^5+4p^6 \geq 0$ is needed to get a positive nonzero solution we get for the critical probability $p_c = \sqrt{2} - \frac{1}{2} \simeq 0.914\,21$. That means the presence of clustering in this example lowers the critical initial density needed to infect the whole graph by almost a factor of $\frac{3}{4}$.

The study of the regular $4-$tree generalizes easily to the case of regular $n+1-$ trees $(n > 2)$. The recursion equation in this case is

$$x = p^n x^n + np^n x^{n-1}(1-x) + np^{n-1}(1-p)x^{n-1} \tag{9}$$

. A straightforward but lengthy computation gives for the critical probability

$$p_c = \frac{(n-1)^{2n-3}}{n^{n-1}(n-2)^{n-2}}; n > 2 \tag{10}$$

. In the special case of a 3− tree ($n = 2$) one obtains $p_c = \frac{1}{2}$. For completeness we give without proof the formula for the computation of the critical probability in case of a rooted random tree with arbitrary outdegree distribution. Let $g(z) = \sum_{i \geq 2} a_i z^i$ be the generating function for the outdegree; that is a_i is the probability that a random chosen vertex has outdegree i (and hence total degree $i + 1$). The critical probability p_c is given by the smallest p such that the equation

$$\frac{z}{p} = (1-z)\, g'(z) + g(z) \tag{11}$$

has a positive real solution.

. We want to close this section by an example where a single infected vertex can infect already a positive fraction of the whole vertex set. Again we chose $\Delta = 2$ but examples for larger Δ values are equally easy to construct in an analog fashion. The important new property of such graphs is the following: any two vertices can be linked by a chain of triangle where neighbor triangles always have a common edge. We start with a regular tree of degree 3. Replacing each vertex by a triangle and gluing the triangles along the former edges of the regular tree gives a regular graph of degree 4 where the triangle corners act now as the new vertices. In each neighbor pair of triangles (A, B) (that are the triangles which have a common vertex) we form an edge randomly between the set of vertices lying in $A \setminus B$ and $B \setminus A$ (see Fig. 1). . Once a triangle is infected the corruption jumps to all the three neighbor triangles due to the extra random edge present between each neighbor pairs of triangles . Hence we have a nonzero probability that the whole graph becomes infected. The graphs in the previous examples of this section do not have this property since neighboring triangles have only one common vertex. For threshold values $\Delta > 2$ one has to consider chains of $\Delta + 1$ cliques. We say that a graph is well k -linked if any pair of vertices can be linked by a chain of complete graphs of order k such that all neighboring $k-$ cliques have a $k-1$-clique in common. For well k- linked graphs the critical density b_c^0 is zero (a finite number of initially infected vertices can already infected a positive fraction of the vertex set) for $\alpha-$processes with $\Delta < k$ whereas for graphs which are not well linked one needs a positive critical density.

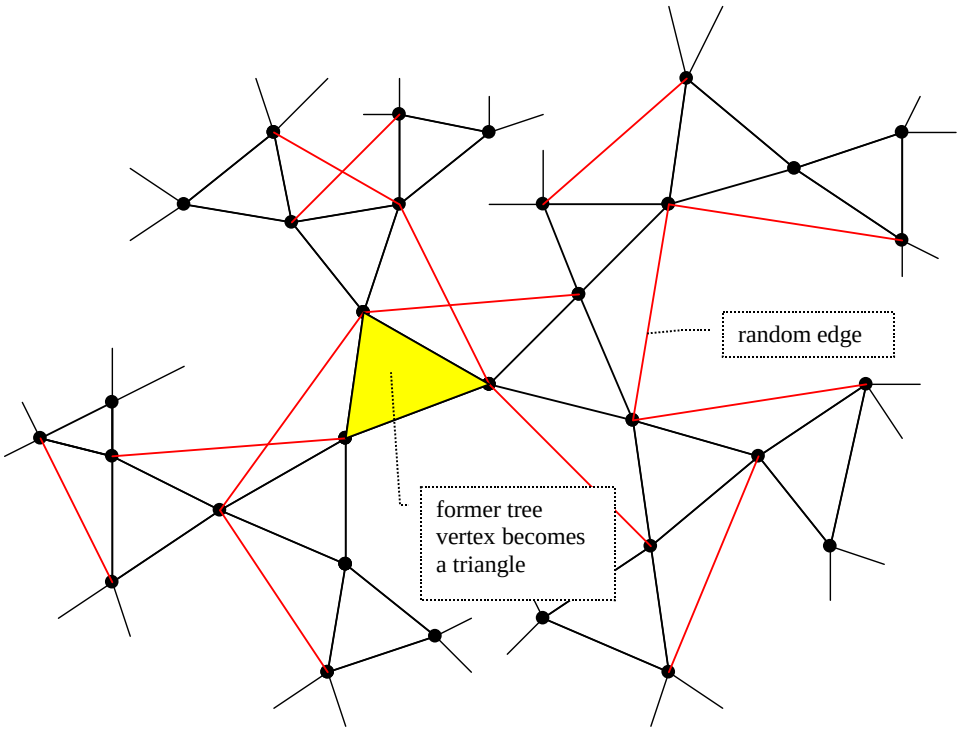


FIG. 11: A highly clustered network with underlying tree structure

The above study on trees or generalized trees is insofar important as in most random graph models used for complex networks one has as a tree or generalized tree as the typical local structure around a random chosen vertex. Furthermore the dependence of the corruption dynamics on graph properties like edge density or degree distribution is in large parts of the parameter space entirely caused by the $\alpha-$ process.

V. PHASE TRANSITIONS FOR THE MEAN FIELD $\beta-$ AND $\gamma-$ PROCESS

In this section we want to have a closer look at the mean field dependence of the corruption process. To gain some insight in the possible type of behavior we start with some simple assumptions which will be refined later on. Recall from section 2 that b_t is the density of corrupt people at time t. We assume that the affinity for an individual to change its behavior from noncorrupt to corrupt increases proportional to the corruption prevalence. Furthermore to become really corruptly minded an individual has to overcome some fear which we put proportional to $(1-b_t)$. Formally this reads as $\Pr\{\omega(x,t+1)=1 \mid \omega(x,t)=0\} = \beta b_t^2$ with $\beta \in [0,1]$. Corrupt individuals can recover due to state and police effects (uncovering, fear

etc.). Again it seems reasonable to assume that the probability to recover is proportional to $1-b_t$ since only the noncorrupt part of a society is willing to fight corruption. Therefore we have $\Pr\{\omega(x,t+1)=0 \mid \omega(x,t)=1\} = \gamma(1-b_t)$ with $\gamma \in [0,1]$. Combining both terms we get for large populations the following mean field dynamic:

$$\begin{aligned} b_{t+1} &= (1-b_t)\,\beta b_t^2 + b_t - b_t\gamma(1-b_t) \\ &= b_t(1-\gamma) + b_t^2(\gamma+\beta) - b_t^3\beta \end{aligned} \tag{12}$$

with the two obvious fixed points 0 and 1. For $\beta \neq 0$ there is a third intermediate fixed point $b^* := \frac{\gamma}{\beta}$. An interesting phenomena happens for parameter pairs (β,γ) s.t. $\gamma < \beta$ since under this conditions both fixed points at 0 and 1 are locally stable. Hence there are two basins of attraction- one for 0 and one for $1-$ with b^* as the boundary point. In other words, if the initial percentage of corruption is less b^* corruption stays under control whereas for an initial value larger b^* things run out of control and a corruption collapse takes place. Of course this mean field part of the model is still very simplistic but the qualitative statement seems to be quite stable with respect to modifications. For instance there are good reasons to believe that neither the mean field infection nor the mean field recover process are linear in b_t.

We want to end this section with a small modification of the mean field "Ansatz" where we include social weights. This is a natural and meanwhile very common approach in network dynamics and can easily be adopted to the corruption model. In the above argumentation on the attraction of becoming corrupt it is plausible to assume that corrupt individuals with high social influence have a stronger influence on the mean field probability to get corrupt than individuals with low social importance. A similar argument holds for the recover probability. As a simple measure for social strength we use the degree of the vertices since high degree vertices are more likely to play a dominant social role than low degree vertices. Formally we introduce the weighted density b_t^wat time t as

$$b_t^w := \frac{\sum\limits_k I_t^{(k)} d_k}{\sum\limits_k (d_k)^2} \tag{13}$$

where d_k is the number of vertices with degree k and $I_t^{(k)}$ the number of corrupt (state 1) vertices with degree k at time t. The mean field equation for group k is now given by

$$I_{t+1}^{(k)} = \beta (b_t^w)^2 \left(d_k - I_t^{(k)}\right) + I_t^{(k)} - \gamma(1-b_t^w) I_t^{(k)} \tag{14}$$

. Multiplying the last equation by $\frac{d_k}{\sum_k (d_k)^2}$ and summing over k gives

$$b_{t+1}^w = (1 - b_t^w)\,\beta\,(b_t^w)^2 + b_t^w - b_t^w \gamma\,(1 - b_t^w) \tag{15}$$

which is the same as equation (12). Therefore the introduction of social weights does not add anything new to the dynamical picture. There is of course a difference in the interpretation since a small real initial prevalence of corruption can give rise to a high initial value of b_0^w as soon as the corruption is concentrated at the high degree vertices. Here also a difference between scale free networks and classical random networks is seen since in the scale free case high degree vertices (hubs) are much more frequent than in the classical case.

VI. INTERACTION BETWEEN THE MEAN FIELD PROCESS AND THE LOCAL THRESHOLD DYNAMICS ($\beta + \gamma$ VERSUS α)

In this paragraph we will investigate some aspects of the interplay between the mean field process described in the previous section and the local, threshold dependent, corruption propagation. For $\alpha > \gamma$ there is a core infected component generated via the $\alpha-$ process. To gain some insight how such a core infected part of the population changes the mean field dynamics we will assume that a certain fraction, say a, of the population is permanently infected and resistant against the $\gamma-$deletion process. Denoting by $q_t = b_t - a$ the density in the noncore part of the population (the normalization here is still with respect to the total population size) we get the following mean field dynamics:

$$\begin{aligned} q_{t+1} &= (1 - a - q_t)\,\beta\,(q_t + a)^2 + q_t - q_t \gamma\,(1 - a - q_t) \\ &= a^2\beta - a^3\beta + q_t\left(2a\beta - \gamma + a\gamma - 3a^2\beta + 1\right) + \\ &+ q_t^2\,(\beta + \gamma - 3a\beta) - \beta q_t^3 \end{aligned} \tag{16}$$

. Since the state where all individuals are infected is stationary we get the following set of fixed points:

$$\left\{-a+1, \frac{1}{\beta}\left(\frac{1}{2}\gamma - a\beta - \frac{1}{2}\sqrt{-4a\beta\gamma + \gamma^2}\right), \frac{1}{\beta}\left(\frac{1}{2}\gamma - a\beta + \frac{1}{2}\sqrt{-4a\beta\gamma + \gamma^2}\right)\right\}$$

. For $-4a\beta\gamma + \gamma^2 < 0$ there are no real fixed points except $q^* = -a + 1$ which becomes globally stable under this condition. Since we have a polynomial of degree 3 we

get $\frac{1}{\beta}\left(\frac{1}{2}\gamma - a\beta + \frac{1}{2}\sqrt{-4a\beta\gamma + \gamma^2}\right) < 1$ as the condition for the fixed point at $1-a$ to be locally stable. Furthermore in this case also the fixed point at $\frac{1}{\beta}\left(\frac{1}{2}\gamma - a\beta - \frac{1}{2}\sqrt{-4a\beta\gamma + \gamma^2}\right)$ becomes a local attractor. This is for instance the case when a becomes very small and $\beta > \gamma$ - being back essentially in the situation of the previous section. In case when $\frac{1}{\beta}\left(\frac{1}{2}\gamma - a\beta + \frac{1}{2}\sqrt{-4a\beta\gamma + \gamma^2}\right) > 1$ it is easy to show that the fixed point at $\frac{1}{\beta}\left(\frac{1}{2}\gamma - a\beta - \frac{1}{2}\sqrt{-4a\beta\gamma + \gamma^2}\right)$ becomes a global attractor (to see this just note that the derivative at $q_t = 0$ is always positive for the relevant parameter intervals). The above considerations show that the possible dynamical evolution scenarios are the same for $a = 0$ and $a \neq 0$. But there is a very strong influence of a on the parameter regimes of β and γ for which one has a corruption collapse. Whereas in case $a = 0$ one is always in the basin of attraction of zero for b_0 sufficiently small and $\gamma \neq 0$ (in other words $b = 1$ is never a global attractor) one can now have the phenomenon that only the complete saturation with corruption is stable ($q = 1 - a$). As an example lets look at the case where $\beta = 2\gamma$. For $a = 0$ there is a fixed point at $b^* = 0.5$ and hence for an initial infection density $b_0 < 0.5$ the pure mean field dynamics converges to zero. In the case $a \neq 0$ one has for $a > 1/8$ only the stable fixed point $b^* = a + q^* = 1$. At $a = 1/8$ there is a phase transition since a new indifferent (slope 1) fixed point at $b^* = 1/4$ emerges. For $a < 1/8$ this fixed point bifurcates into two fixed points where the first one at $b^* = \frac{1}{4} - \frac{1}{4}\sqrt{1-8a}$ becomes locally stable with a basin of attraction given by $b_0 < \frac{1}{4} + \frac{1}{4}\sqrt{1-8a}$.

VII. TIME EVOLUTION

In the previous sections we looked at the dependence of the asymptotic prevalence on the parameter values and the graph structure. In this section we want to focus on the concrete time evolution of the corruption process for several given initial configurations on some medium size complex networks. Small graph sizes are interesting as they are typical for communities in highly social structured populations. As a simple to generate random graph space with high clustering and power law degree distribution we have chosen so-called intersection graphs. Intersection graphs can easily be defined as follows. First one forms random sets from a finite base set of N elements (random means in this context that the set elements are chosen uniform i.i.d. from the base set). These sets constitute the vertices of a random graph. Edges will be defined via the set intersection property, namely there is

graph characteristic	**FP2 (1987-1991)**	**FP3 (1990-1994)**
# vertices	4879	7710
# edges	57633	93852
mean degree	23.624	24.346
maximal degree	844	1014
# vertices with degree > 5	3865	6051
size of largest component	4775	7356
mean # triangles per vertex	256.89	418.09
exponent of degree distribution	2.1	2.4

TABLE II: Properties of the real networks FP2 and FP3

an edge between i and j if the associated sets A_i and A_j have nonempty intersection. The size (cardinality) $|A|$ of a set A is itself a random variable drawn i.i.d. from a pre-given probability distribution $\varphi(k)$. To get interesting graph spaces one furthermore requires $N < \sum |A_i| < const \cdot N$. For theoretical results about the structure of random intersection graphs see [8][15][12]. It is worth noting that intersection graphs have a high clustering by definition (if an element is contained in say k sets simultaneously this k sets form a complete subgraph). Most simulations were done for the case when φ is an asymptotic power law distribution with exponent 3 or when φ is singular (all sets have the same size). Random intersection graphs have a multiplicative degree correlation and therefore the critical threshold should be very low for exponents less than 3 be the arguments from section V. Above that value the form of the degree distribution has only little influence on the corruption propagation. Besides random intersection graphs generated according to some degree specifications we used also a collection of real collaboration graphs. These graphs come from a database about research and development projects funded by the European Community (FP2-3)[4]. It's vertices are organizations involved in European research projects. Two organizations are linked if they have a joint project (see table II for the main graph characteristics). In total the data base contains about 8000 projects and 13000 participating organizations. In essence the network shows all the main characteristics that are known from other complex network structures like scale free degree distribution (with exponent between 2 and 3), small diameter and high clustering and vertex correlation. The initial fraction of infected individuals was either

distributed at random over the vertex set or clumped together in a sufficiently large ball with a random chosen vertex as center.

In the following we want to give a small sample of simulations on the just mentioned graphs and try to discuss its main features. Fig.12 displays the prevalence of corruption on the real network FP2. The absolute threshold value $\Delta = 30$ is very high and does not allow for a big outbreak of corruption. But there is a metastable small community of individuals, highly linked and almost resistant to the $\gamma-$process. It took more then 800 complete updates till this structure broke down.

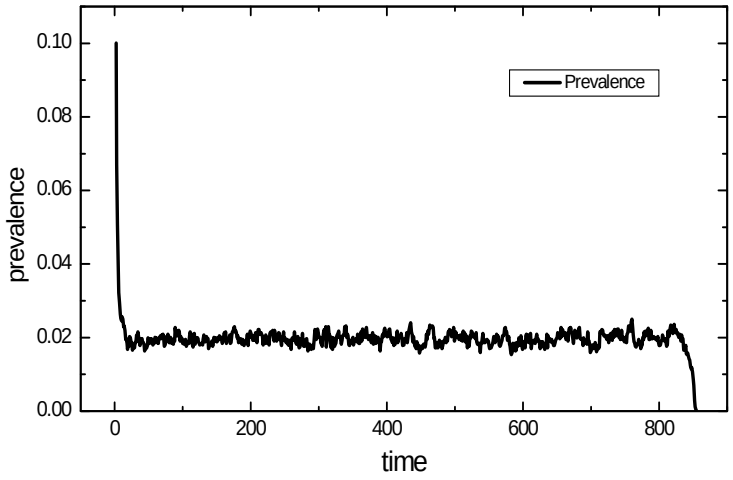


FIG. 12: Low, semistable prevalence in a real collaboration network (FP2), Δ=30 ε=0, α=0.99 β=0.09 γ=0.545 b_0=0.1 N=4879

The next figure 13 presents a similar situation on an almost twice as large real graph (FP3).

In contrast to the previous case we have a much smaller $\alpha-$ value and an only slightly reduced threshold Δ.

The network FP3 is extremely high clustered (mean degree = 48.6, mean triangle number = 418 and a total of 7710 nodes and 187704 edges) and stays metastable with a very small corruption cluster for about 200 updates till it jumps by a factor 10 to another metastable state. The sub-figure in Fig. 13 gives a more detailed view of the accumulated contributions

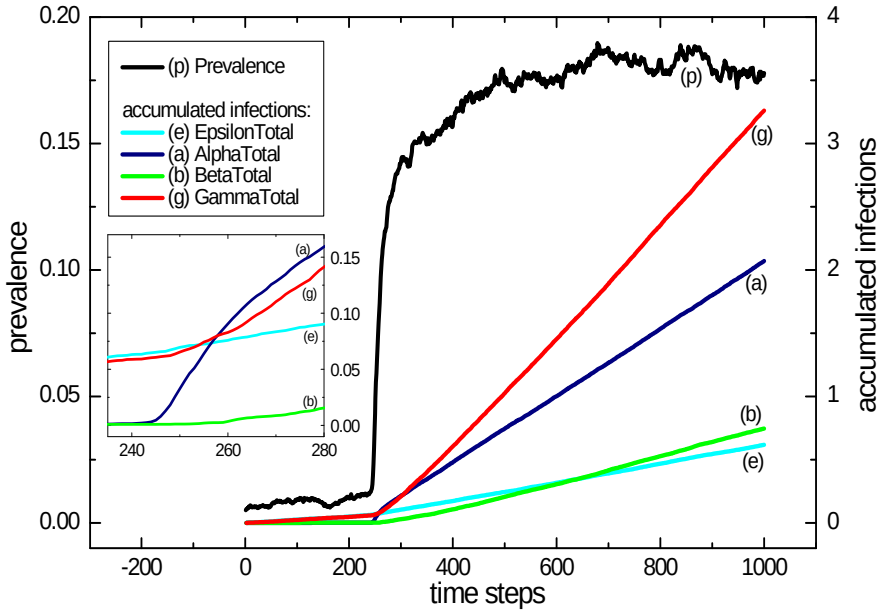


FIG. 13: switch from a very low prevalence semi stable state to a medium-low prevalence state (FP3), Δ=25, ε=0.001, α=0.2, β=0.04, γ=0.03, b_0=0.005, N=7710

by the different processes for a time interval around the jump in prevalence. In the initial phase the $\varepsilon-$process was dominating the $\beta-$process and vice versa in the second phase. The next pictures show a situation where after an initial phase of slow growth a corruption collapse happened. It seems that the absolute threshold value $\Delta = 20$ is well below the critical value where the system can still stabilize. It is surprising that the system semi-stabilizes after an initial rapid increase in the prevalence for a rather long time (Fig.14). In the subpicture of Fig.14 the accumulated infection processes for the initial phase are shown. Here the ε- process, although undercritical, causes a redistribution of infection till a clustered configuration is reached such that the α- process can start. Than the systems stays in almost complete balance till the $\beta-$ process (which is slow in all our examples) wins. Note the difference to Fig.**??**, where the $\beta-$ process never really contributes to the infection. Finally we show two simulations for a sample of a random set graph model with about 1000 vertices (Fig.16 and Fig.15). Although both prevalence curves look similar there is a clear difference in the process fine-structure (Fig.16 and Fig.15). In the first instance

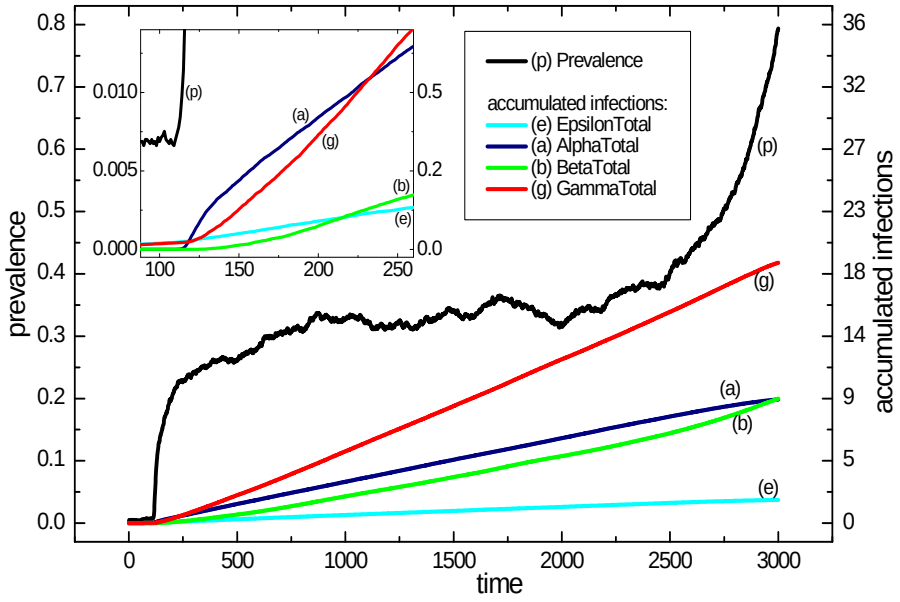


FIG. 14: Slow increase of prevalence till collapse (FP3),Δ=20 ε=0.001 α=0.2 β=0.04 γ=0.03 b_0=0.005

the $\varepsilon-$ and $\beta-$ process are causing the collapse whereas in the second case the $\alpha-$ process in conjunction with the $\varepsilon-$ process is the main booster.

The few examples of single simulation runs given in this section show already, that there are many different routes to obtain high prevalence in corruption typically interrupted by long phases of metastability. Similar to other complex systems with hidden phase transitions (e.g. the climate) there can be an unnoticed small accumulation of infection till a critical density- a point of no return- of corruption is reached from which on an almost complete saturation of the society (or a corresponding subsystem) by corruption becomes the normality.

VIII. EPIDEMIC CONTROL

One of the basic question in classical epidemics as well as in corruption dynamics is: what can be done to slow down the "infection" propagation or prevalence. Knowing the different

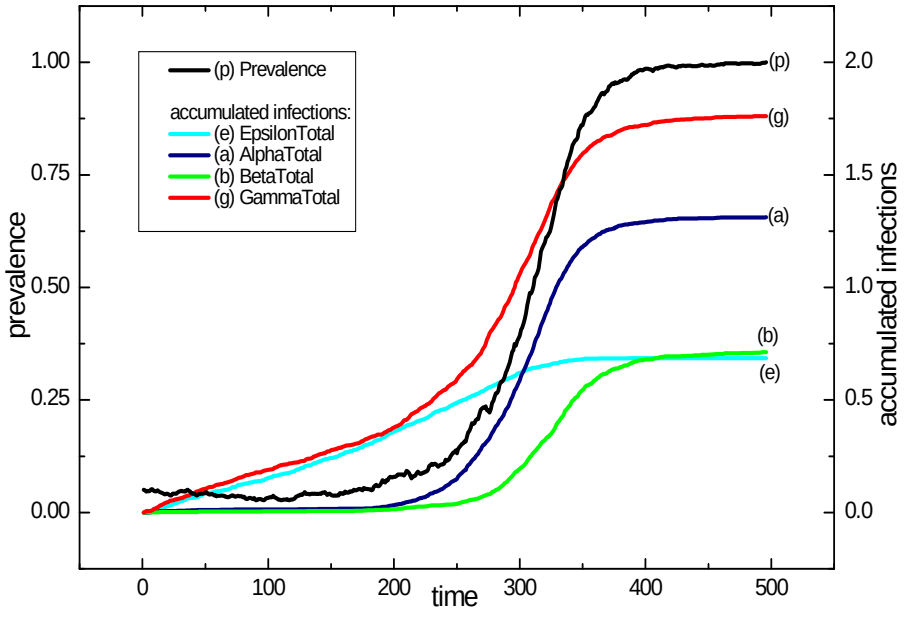


FIG. 15: Slow corruption collaps in an artificial net Δ=6 ε=0.005 α=0.1 β=0.06 γ=0.05 b_0=0.05 N=972

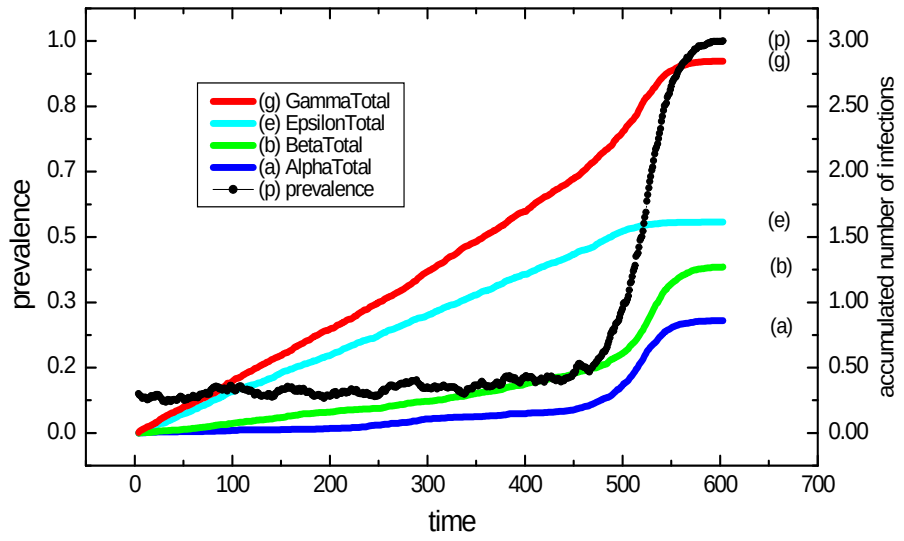


FIG. 16: Δ=8 ε=0.005 α=0.1 β=0.09 γ=0.045 b_0=0.1 N=952

phase transitions and their dependency on structure properties and social parameters is of great help in designing proper prevention scenarios. In the following we will try to relate some of the findings from our model to what is considered by practicians as useful in corruption reduction. First we would like to emphasize again that the present model deals in a rather abstract way with the propagation of mental willingness to be corrupt and not so much with realized corruption which always requires a specific environment and additional structural assumptions. Hence concerning corruption control, we only will be able to support certain prevention scenarios in the sense, that they go into the right direction and that there effect is strong or weak but without being able to make quantitative statements.

The model presented in this paper contains, besides structural parameters for the underlying network, 5 relevant parameters: α- characterizing the strength of the local threshold process, β- characterizing the strength of the mean field attraction on becoming corrupt, γ- the strength of the "society strikes back" term, ε- the strength of the classical epidemic process (assumed to be very small) and Δ- the height of the local threshold. Three of the parameters- α, β and ε- are positively correlated to the spread of corruption whereas 2 parameters- Δ and γ- are negatively correlated. As is well known from classical epidemic control for infectious diseases it is very hard if not impossible to change basic social parameters in a short time. This can only be achieved in a long running educational process. Therefore not much can be done in avoiding high clustering in certain relevant areas of the society in order to prevent the emergence of highly connected corruption nets.

As the name already indicates, Transparency International favours as an effective tool to decrease corruption especially the increase of transparency in all forms of administrative decision making as well as transparency in financial affairs of socially exposed persons, institutions and companies. The effect of an increase of transparency translates into our model as an increase of the value of Δ and a decrease of the values of β, α and ε. Strengthening of justice, police and similar instruments to fight and uncover corruption has again the effect of lowering β (via increase of fear) but may also increase the value of γ (uncovering rate). Since an increase of γ above the value of β and α would perhaps require a total police state, it is illusionary to overcome corruption just by means of law, justice and police. Besides necessary long term educational efforts in school and public to strengthen the moral resistance against corruption (increase of Δ and decrease of α) it seems a good strategy to make administrative and political decision hierarchies as independent and decentralized as

possible to avoid high clustering.

We would like to end these short remarks by a few comments on the role of hubs - the very high degree vertices typically present in scale free graphs - in corruption dynamics. While a priory not especially well suited to transmit corruption via the $\alpha-$ process due to the local tree like structure around the hubs (compared with low degree vertices) they nevertheless are more often exposed to corruption and have therefore a higher probability to get corrupt. If the hub density is sufficiently high (as is the case for scale-free degree distributions with exponent $\lambda < 3$) and the degree correlation is stronger than additive, many vertices are linked to the hubs via social dependencies and in turn also can get corrupt. Furthermore they may play a fatal role in increasing the weighted corruption density relevant for the mean field process as was explained at the end of section VII. The described situation is probably typical for strongly hierarchically organized countries or regional substructures e.g. systems with a dictatorial or monarchical tendency. In such societies a high prevalence of corruption seems almost unavoidable since the threshold b_0^c is close to zero. For democratic societies it seems therefore wise, to watch the behavior of hubs- whatever their social interpretation might be- more intensively than the "normal" part of the society.

IX. SUMMARY AND PERSPECTIVES

In this article we have presented a first study of the spread of corruption on scale free and highly clustered networks. One of the main observations so far is the strong dependence of the asymptotic dynamics on the initial number of corrupt individuals. This holds as well for the mean field process as for the local dynamics. Second there is a fatal resonance effect between global and local dynamics lowering dramatically the critical density of initial infection. As expected there is a positive correlation between clustering and spread of corruption respectively the critical initial density. Scale-freenes seems to play an important role for the corruption process for distributions with small exponent ($\lambda < 3$) and multiplicative degree correlation due to the high prevalence of infected hubs and the strong linkage of medium and low degree vertices to them. For higher exponents the dynamics is rather insensitive to the degree distribution. The strength of the degree correlation (from weak - additive till strong - multiplicative or even higher powers) in networks of social acquaintances seems to be related to the political and institutional structure of a society which favours liberal

organization forms as being less vulnerable to corruption.

There is a whole bunch of natural continuations or generalizations which have to be investigated next. Clearly a deeper understanding of the pure $\alpha-$ process and its phase transitions is necessary. The mathematical problem is already highly nontrivial on trees. The following short list gives a selection of natural generalizations and refinements:

- quenched disorder in all parameters
- inclusion of geographical or regional structure into the network
- inclusion of administrative or political substructures in which corruption typically will be realized
- evolving networks
- interaction between the corruption process and the network structure
- more heterogeneity in the social networks e.g. by incorporating family like structures or social profiles
- refined transition rules e.g. asymmetry between infecting and getting infected
- weighted networks
- different kinds and strength of corruption and their interplay
- economic impacts in a virtual population.

Besides the specific context of corruption dynamics there is a multitude of topics where the model presented in this paper could easily be adopted to. This includes so different themes as political opinion formation, social disorder processes, strategies for advertisement, doping usage, the spread of prejudices, migration dynamics, global terrorist networks and innovation processes. In all these examples one has a local and global dynamics very similar to the one described here. Of course there are differences. For instance in many mind formation problems the state space of individuals is rather complex and the local dynamics allows for many transitions not just 0-1 as in the corruption model. Furthermore aging phenomena and limits of resources could be included. But besides this addition of structure and complexity and the various interpretations there remains a good part of the findings of this work to be

true. There will be phase transitions in the initial density of certain properties and there can be resonance effects between the nonlinear global and local dynamics - both making the prediction of future difficult and challenging.

Acknowledgments

We would like to thank for the support of the Volkswagen Foundation, the DFG-Research Group 399 "Spectral Analysis, Asymptotic Distributions and Stochastic Dynamics" and the Austrian Research Center Seibersdorf, the latter also for providing us with the data set of EU-funded research projects. P. Martin especially thanks Klaus Geppert, Detlef Leenen, Christian Pestalozza and Albrecht Randelzhofer for stimulating discussions.

---

[1] R.Albert , A.-L. Barabasi : *Statistical Mechanics of Complex Networks*, Reviews of Modern Physics, **74**, 47 (2002), arXiv:cond-mat/0106096

[2] J. Andvig, O.-H. Fjeldstad, I. Amundsen, T. Sissener, T. Søreide : *Research on Corruption, a policy motivated survey,* Norwegian Agency for Development Co-operation, NORAD, Final report 2000

[3] B. Bannenberg: *Korruption in Deutschland und ihre strafrechtliche Kontrolle,* (2002), www.im.nrw.de

[4] M. Barber, A. Krueger, T. Krueger and T. Roediger-Schluga: *The Network of EU-Funded Collaborative R&D Projects,* Phys. Rev. E 73, 036132 (2006), arXiv: physics/0509119

[5] Ph. Blanchard, T. Krueger: *The "Cameo principle" and the origin of Scale-free graphs in social networks,* Journal of Statistical Physics, **114**, 5-6 (2004), arXiv: cond-mat/0302611

[6] Ph. Blanchard, C. Chang, T. Krueger: *Epidemic thresholds on scale-free graphs: the interplay between exponent and preferential choice*, Annales Henri Poincaré vol. 4, suppl.2, (2003), arXiv: cond-mat/0207319

[7] Ph. Blanchard, T. Krueger, A.Ruschhaupt: *Small world graphs by iteration of local edge formation,* Phys. Rev. E **71**, 046139, (2005) arXiv: cond-mat/0304563

[8] Ph. Blanchard, T. Krueger: *Random scale free intersection graphs and related bipartite structures,* in preparation

[9] Ph. Blanchard, S. Fortunato, T. Krueger: *How extremists impose the structure of social networks,* Phys.Rev.E, **71**, 056114 (2005), arXiv: cond-mat/0407434

[10] P. Dodds, D. Watts: *Universal behavior in a generalized model of contagion,* arXiv:cond-mat/0403699 (2004)

[11] Ch. Ellis, J. Fender: *Corruption and Transparency in a Growth Model*, preliminary draft (2003)

[12] J. Fill, E. Scheinerman, K. Singer-Cohen: *Random intersection graphs when m=(n): An equivalence theorem relating the evolution of the G(n, m, p) and G(n, p) models*, Random Structures and Algorithms **16**, 2 (2000)

[13] S. Galam: *Minority opinion spreading in random geometry,* Eur. Phys. J. B **25,** (2002)

[14] Ch. Borghesi, S. Galam: *Chaotic, staggered, and polarized dynamics in opinion forming: The contrarian effect,* Phys. Rev. E **73**, 066118 (2006)

[15] M. Karonski, E. Scheinerman, K. Singer-Cohen: *On random intersection graphs: the subgraph problem,* Combinatorics, Probability and Computing **8** (1999)

[16] P. Newman: *The spread of epidemic disease on networks,* Phys. Rev. E 66, 016128 (2002)

[17] P. Newman, J. Park: *Why social networks are different from other types of networks,* Phys. Rev. E **68**, 036122 (2003)

[18] R. Pastor-Satorras , A. Vespignani: *Epidemic Spreading in Scale-Free Networks*, Phys.Rev.Lett.**86,** 3200 (2001)

[19] S. Shi, T. Temzelides: *A Model of Bureaucracy and Corruption*, International Economic Review **45**, 3 (2004)

[20] T. Steinrücken: *Illegale Transaktionen und staatliches Handeln,* Deutscher Universitätsverlag 2003

[21] Transparency International: *CPI 2004*, http://www.transparency.org/cpi/2004/cpi2004.en.html

[22] F.Wirl : *Socio-economic typologies of bureaucratic corruption and implications*, Evolutionary Economics, **8** (1998)

CAMBO: Clustering by Adjacency Matrix Block Ordering

Andreas Krueger*

University of Bielefeld, Mathematical Physics

Abstract

Clustering results are often visualized as block-structured adjacency matrices. When the nodes are clustered and sorted by their cluster order, the adjacency matrix shows *blocks* of more-strongly connected subspaces along the matrix diagonal. The inspiring idea of our new algorithm was: Why not **directly** sort the nodes into such a block-structure?

We inductively developed a deterministic algorithm that uses a parametrized heuristic of mutual 'distances' of all nodes, reorders them by smallest distances in a linear chain, cuts between clusters at the highest distance jumps, and takes the one clustering with the best modularity as the end result. The three parameters influence the mixing of the direct connection weight A_{ij}, the two-step connections $(A^2)_{ij}$, the N_1-neighbourhood similarity, and the N_2-neighbourhood similarity. A proof-of-concept-implementation suitable for small networks is described. The algorithmic time complexity is $O(N^3)$ due to the matrix multiplication, we give a discussion of possible enhancements to the algorithm. The fruitfulness of this approach is shown through application to several networks: the Zachary Karate Club, where an unknown high-modularity 3-clustering could be found by our method; a set of 96 tumors that are clustered by their gene-similarity; and clustered topics of 27000 EU-funded R&D projects.

*Electronic address: akrueger@physik.uni-bielefeld.de,networks@AndreasKrueger.de

1. INTRODUCTION

1.1. Clustering

Clustering of networks into subgroups of highly interconnected nodes is one of the most interesting questions of network analysis. It helps to structure data by dividing a large network into smaller partitions with a high internal similarity and a low similarity to the nodes of the other partitions. With the great interest in network physics of the past decade ([2], [3], [4], [10], [24]) many new and efficient clustering algorithms have been found: Stochastic methods like the Potts model relaxation [18], deterministic methods like highest betweenness centrality cuttings [14], or matrix-multiplication-and-inflation fix points [22] [23] with the background of random walks, etc.. Some of the methods are divisive, i.e. cutting the whole network into smaller parts, some of them agglomerative, i.e. growing clusters starting from single nodes - our method is similiar to the latter type. Agglomerative as it *locally* looks for the next best candidate, but not completely local, because "best" related to *all previously chosen* nodes and with the global knowledge of all mutual line pseudo-distances - and with a *global* quality measure, Newman modularity.

For a comparison of clustering methods, see [9].

1.1.1. Clustering by matrix reordering

After clustering, for presenting the resulting clusters graphically, sometimes a matrix density plot in a *block matrix* structure is given. This paper presents a method that directly aims at generating such a block matrix. The idea came suddenly: Why should we use an "external" method to ultimately get a block matrix structure, if we can find a *direct way to get a block matrix structure*! We created and implemented such a method as a proof-of-concept.

The current network community in physics does not yet seem to have considered this strategy, but there is a history of related methods in computer science, which we discovered only recently: Sparse matrices are objects in many numerical algorithms like linear equation solving. They are usually not stored with all their zeros, but in more effective data structures. Reducing the "bandwidth" of such a sparse matrix by reordering the nodes helps to keep the memory usage even lower. In graph and matrix libraries like BOOST [5], the implemented

methods for reordering are Cuthill-MacKee [8], King [11] and Sloan [19]. They all take a given starting node (or find one semi-heuristically), and then identify an ordering of the nodes that keeps similar nodes close together. They are not targeted towards community clustering, however.

1.1.2. The CAMBO algorithm - short description

Our method follows a not-completely dissimilar strategy. We calculate pseudo-distances between all nodes once, choose a starting node, and then by locally minimizing some criterion, iteratively all the others. For each new node then, we have a "jump size" to all the previously chosen nodes. The highest jump size gives the first cutting, the second highest the second cutting, etc. By maximising the Newman criterion "modularity" [15], the "best" clustering is chosen among all those.

Or to translate this into the matrix picture, if the matrix is the adjacency matrix of such a network: We calculate distances between all lines once, choose a starting line, and then iteratively all the others, by locally minimizing some criterion. For each new line then, we have a 'jump size' to all the previously chosen lines. The highest jump size separates the most dissimliar blocks in that block matrix, the second highest the second most different blocks, etc. By maximising the Newman criterion "modularity" [15], the "best" clustering is chosen among all those.

The terms used for the matrix-line/network-node similarity work locally (direct, 2-step-connection) and globally (structural equivalence). The Newman Modularity has a local (uncut edges within clusters are good) and a global view (degree), too.

Neither of the criteria alone is sufficient for perfect clustering, so CAMBO tries to embrace all of them. Still, the last decision is always done by maximising the Newman Modularity, so (a) a "good" clustering with *low* Newman Modularity will *not* be found without detuning from the optimal parameter point, and (b) here is a request to the theoreticians: Please create more general network modularization observables than the now widely used "Newman modularity" (equation (5) below).

The storyline of this algorithm is *not* to introduce new reductionist concepts, rather to combine existing objects of the network landscape into a constructive scheme that has previously not been reported in the physics community!

2. A NEW CLUSTERING APPROACH

2.1. Matrix representation of networks

Let $G = (V, E)$ be an *undirected* graph with $N \in Z^+$ vertices $V \subseteq Z^+$ and $M \in Z^+$ edges $E \subseteq \binom{V}{2}$. The cardinality of the edges set E is bound by the full graph $M \leq \frac{N(N-1)}{2}$. A *weighted* graph carries for each edge $e = (v_1, v_2) \in \binom{V}{2}$ an edge function $w : \ E \to \mathbb{R}$, $w(e) = w(v_1, v_2)$.This real number can e.g. represent a connection or co-ocurrence strength between node v_1 and v_2 (v_i and v_j sometimes simply called nodes i and j). The unweighted case can then be recovered, if $w(e) \in \{0, 1\}, \forall e$. If we have an ordering $\{1, 2, ..., N\}$ of the node names into rows and column names, a symmetric matrix A, called the *adjacency matrix*, represents the whole network by putting the edge weight $w(e) = w(v_i, vj)$ into matrix element $A_{ij} = w(v_i, v_j)$. A row (or column, as for an undirected graph the matrix is symmetric) of that matrix, $(A_{ij})_i$ gives us the total information about all *direct* connections of node j to all other nodes i, that is the N_1-neighbourhood of j.

2.1.1. The N_1 and N_2 neighbourhoods: The matrices A and A^2

For simplicity, let us look at an *unweighted* network with a connection between two nodes ($A_{ij} = 1$) or no connection between two nodes ($A_{ij} = 0$). For a network with N nodes, the matrix A corresponds to a linear function in $\mathbb{Z}^N$ that maps the N Euclidean base vectors (like $e_i = (0, ..., \overset{i}{1}, .., 0)$ representing the node i) onto vectors n_i in which the vector components are 1 if a node is neighbour of i, and 0 if not:

$$A(e_i)^T = (A_{vw})(0, ..., \overset{i}{1}, .., 0)^T = (A_{1i}, A_{2i}, ..., A_{Ni})^T = (0, 1, 0, 0, 1, 1, .., 0)^T \qquad (1)$$

Applying this function A *twice* now gives the number of paths of length 2 between i and all other nodes, so to the next nearest neighbours:

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

which shows that sum of these $A_{ik}A_{kj}$ are the *number of 2-step-edges* between the two nodes i, j over all k (and thus the *number of triangles between i and j* if $A_{ij} = 1$). So A tells us about the N_1-neighbourhood, and A^2 about the N_2-neighbourhood of a node.

N.B.: What we call N_2-neighbourhood here, is meant to contain all those nodes that can be reached by travelling exactly 2 edges far - if the node degree is not zero, this includes the node itself (travelling any edge forwards and backwards *is* 2 steps), and if a node is part of a triangle then the other two corners are both in the N_1- and N_2-neighbourhood.

In the CAMBO algorithm, we compare all nodes mutually by their N_1- *and* N_2-*connection* - and additionally, by their N_1 and N_2 *neighbourhood-difference.*

2.1.2. Weighted

Let us now leave this simplification which was chosen for didactical purposes. If we extend the A_{ij} from a binary variable to *weighted* edges, most concepts can be transfered (e.g. "there is no/an edge between i and j (of strength 1)" to "there is no/an edge with strength A_{ij}", but it is highly unclear what triangles are, etc. To keep the 2-step links $(A^2)_{ij}$ still makes sense, because the larger the sum of all $A_{ik}A_{kj}$, the more and stronger 2-steps "ways" are between i and j, and their *multiplicative* combination $A_{ik}A_{kj}$ is similiar to combining two independent probabilities A_{ik} and A_{kj} to their joint probability.

2.1.3. Structural equivalence

Think about the pictures of blocks in clustered matrices: What will create blocks of strong connection (bright colour) along the diagonal, and weak connection far off the diagonal (dark colour)? The more similiar two lines are, the closer they should be in the matrix, because then their overlap will be kept in their common block, and the dark part will be similiar, too, because these two lines are similarly weak connected to all the other nodes of all the other clusters. Here is where a "structural equivalence" (term from sociology) enters: Two nodes are structurally equivalent, if their neighbourhood is identical, so if they are connected to the same nodes. If two nodes are in the same cluster, they should probably share a lot of their "dark" parts in the matrix.

structural equivalence of i and j can be expressed by: $\sum_{\substack{k=1 \\ k\neq i,j}}^{N} |A_{ik} - A_{jk}|$

If this term is zero, the nodes i and j are completely structurally equivalent.

2.2. A line distances heuristic for a parametrized deterministic clustering

In order to identify *similiar* nodes during a clustering attempt, we create a 3-parameter heuristic d'_{ij} weight that mixes the direct connection weight A_{ij} of two nodes i and j, the 2-step weight $(A^2)_{ij}$ of two nodes i and j, and the differences of their N_1- and N_2-neighbourhoods - the resulting d'_{ij} is a constructive proxy for the (negative) network "clustering strength" d'_{ij} between the two nodes i and j:

$$\begin{aligned} d'_{ij}(\tau', \beta', \gamma') &= -A_{ij} - \tau'(A^2)_{ij} + \beta' \sum_{\substack{k=1 \\ k\neq i,j}}^{N} |A_{ik} - A_{jk}| + \gamma' \sum_{\substack{k=1 \\ k\neq i,j}}^{N} \left|(A^2)_{ik} - (A^2)_{jk}\right| \\ &= -A_{ij} - \tau'(A^2)_{ij} + \beta' B'_{ij} + \gamma' C'_{ij} \end{aligned} \tag{3}$$

Structural equivalence means that the neighbourhood of two nodes are identical, and the *greater the difference between* the N_1- (and N_2-) neighbourhoods of two nodes i and j, the *greater* are the two sums $B'_{ij} = \sum_{k=1,\ k\neq i,j}^{N} |A_{ik} - A_{jk}|$ and $C'_{ij} = \sum_{k=1,\ k\neq i,j}^{N} |(A^2)_{ik} - (A^2)_{jk}|$, and thus **the greater** this pseudo-distance d'_{ij} between node i and j. The stronger the direct connection A_{ij} or the 2-step connection $(A^2)_{ij}$ between them, **the lesser** is that pseudo-distance d'_{ij}, therefore the minus sign. B'_{ij} and C'_{ij} are calculated only once to save computing time.

2.2.1. Normalizing the four terms was unsuccessful

You might want to skip this chapter; it is documenting what is implemented, but it is not crucial to understand the CAMBO method itself - and by the nature of the two sorts of terms (values and *differences* of values), the summation just could not be successful, which we only found out afterwards.

At the very beginning, there was the hope that a parameter set (τ, β, γ) for the *best* clustering of the matrix A_{ij} can stand for the whole network, and thus make completely different networks comparable to each other, by only looking at their optimal (τ, β, γ). So we tried to normalize all four terms in (3). This normalization is not crucial for the method itself, yet still in the programmed version we constructed the pseudo-distance slightly differently from equation (3) above.

The main obstacle is that the connection weights in A_{ij} and $(A^2)_{ij}$ are by their nature very different from the differences of N_1- and N_2-neighbourhoods; the first are values, the second differences of values. We still tried to normalize both in a way that was hoped to be reasonably summable.

Given is an adjacency matrix $A = A_{ij}$ with an irrelevant diagonal (The diagonal A_{ii} in an adjacency matrix represent self-loops that are irrelevant to clustering).The averages of the non-diagonal elements are called α_1 and α_2:

$\alpha_1 = \frac{1}{N^2-N} \sum\limits_{\substack{i,j=1 \\ i\neq j}}^{N} A_{ij}$ and $\alpha_2 = \frac{1}{N^2-N} \sum\limits_{\substack{i,j=1 \\ i\neq j}}^{N} (A^2)_{ij}$

Then we normalize the matrices A_{ij} and $(A^2)_{ij}$ to $\widetilde{A}_{ij}$ and $(\widetilde{A^2})_{ij}$ so that each has a *standard deviation of 1.* The transformed matrix $\widetilde{A}_{ij}$

$$\widetilde{A}_{ij} = (A_{ij}-\mu)/\sigma \qquad \text{with} \qquad \mu = \frac{1}{N^2-N} \sum_{\substack{i,j=1 \\ i\neq j}}^{N} A_{ij} \qquad \text{and} \qquad \sigma^2 = \frac{1}{N^2-N} \sum_{\substack{i,j=1 \\ i\neq j}}^{N} (A_{ij}-\mu)^2$$

then has a variance $\widetilde{\sigma}^2 = 1$. The plan was that (for "well-behaving" distributions of matrix elements) we would have an *expectation value of 1 for the differences of matrix elements,* and then by dividing with $\frac{1}{N-2}$, these contributions would be of the same order of magnitude as the connection weights.

The pseudo-distance $d_{ij}(\tau, \beta, \gamma)$ then looks like this, and this is the way it is actually implemented

$$\begin{aligned} d_{ij}(\tau,\beta,\gamma) &= -\frac{1}{\alpha_1} A_{ij} - \frac{\tau}{\alpha_2} (A^2)_{ij} \\ &\quad + \frac{\beta}{N-2} \sum_{\substack{k=1 \\ k\neq i,j}}^{N} \left| \widetilde{A}_{ik} - \widetilde{A}_{jk} \right| \\ &\quad + \frac{\gamma}{N-2} \sum_{\substack{k=1 \\ k\neq i,j}}^{N} \left| (\widetilde{A^2})_{ik} - (\widetilde{A^2})_{jk} \right| \\ &= -\frac{1}{\alpha_1} A_{ij} - \tau \frac{1}{\alpha_2} (A^2)_{ij} + \beta \cdot B_{ij} + \gamma \cdot C_{ij} \end{aligned} \tag{4}$$

with $B_{ij} = \sum\limits_{\substack{k=1 \\ k\neq i,j}}^{N} \left| \widetilde{A}_{ik} - \widetilde{A}_{jk} \right|$ and $C_{ij} = \sum\limits_{\substack{k=1 \\ k\neq i,j}}^{N} \left| (\widetilde{A^2})_{ik} - (\widetilde{A^2})_{jk} \right|$.

This $\widetilde{\ }$-matrix transformation, and the $\frac{1}{N-2}$ factors in (4) are **not** crucial for the CAMBO-method itself, but they were introduced with the intention to compare the (τ, β, γ) combination of the optimal clustering among different networks. The hope was that the optimal

clustering of a network tells us a great deal about the internal properties of that network, and the optimal parameter set (τ, β, γ) is a proxy for a certain way to cluster that network - so that in the end, on an abstract (τ, β, γ)-level, completely different networks can be set into relation to each other. This was not successful due to the different nature of the ingredients of (4).

2.2.2. Reparametrization?

At this level of understanding, the search for a "perfect" parametrization is not yet finished (partly due to the still small number of test cases) and could be topic of a subsequent paper. As soon as dozens of networks are clustered, and their optimal (τ, β, γ) compared, we could create a better parametrization and check it in reality. We hoped to condense the network clustering property into a small-dimensional vector like (τ, β, γ), and to get a 'feeling' for the clustered network itself, by looking at the optimal (τ, β, γ). Unfortunately at the moment, the parameters are still size dependent.

One possibility would be to use only the *non-zero* elements for averaging α_1 and α_2 because for unweighted or almost-unweighted networks (with a lot of zeroes in the adjacency matrix) a simple sum is misleading - and also using the same standardization for the sums B_{ij} and C_{ij}, instead of the $\sqrt{\sigma^2} \to 1$, could be worth trying.

Moreover, clearer from the mathematical viewpoint is probably the negative value $s_{ij} = -d_{ij}$ which could be called 'similarity' of the nodes i and j.

Translated back into the language of pure mathematics all this might seem a little obscure, but in the end, all these choices were made during *programming the objects* that were used in the algorithms - computer programs and mathematics sometimes look very differently.

2.3. The modularity of a clustering

At several stages of the algorithm (and through experimenting with the many versions in the development phase of this now 5000 lines of code), we need a criterion to decide how *"good"* a clustering is, in order to identify the (locally) *optimal* clustering. A clustered network can be *cut* into partitions, by keeping only the edges *within* the clusters - Loosely speaking, the fewer existing edges of the total network that are cut by a certain clustering,

the better that clustering is, and the more non-existing edges are cut, the better the clustering is. We tried several modularity measures like the "Boutin-Hascoet-Modularization" [6], the "van-Dongen-efficiency" [22], [23] and the "Newman modularity" [15]. Because it subjectively produced the clearest results visually, we ultimately chose the Newman modularity Q, knowing about its drawbacks:

$$Q(A_{ij}, k_i, c_i) = \frac{1}{2M}\sum_{v,w}^{N}(A_{vw} - \frac{k_v k_w}{2M})\delta(c_v, c_w) \tag{5}$$

with the degree k_i of node i, M the number of edges ($2M = \sum k_i$), and c_i the cluster number of node i, so that the sum (5) is only over all nodes v and w that are in the same cluster. The $k_v k_w$ term shows one drawback: The implicit null-model is a multiplicative ($k_v \cdot k_w$), e.g. not an additive ($k_v + k_w$) degree-degree correlation, which might be misleading for networks with a strong deviation from a multiplicative degree-correlation.

2.4. The CAMBO clustering algorithm

The following 7 steps describe the CAMBO algorithm:

(1) The adjacency matrix A_{ij} is created from the given network, then normalized, squared, and subtracted. As the resulting A_{ij} and A^2_{ij}, B_{ij} and C_{ij} are identical for all later runs, they are stored on harddisk for later use in each (τ, β, γ) step. Now, in each of such steps, we create a pseudo-distance matrix $d_{ij}(\tau, \beta, \gamma)$ calculated by (4).

(2) At first it seemed logical, that the method depends on the starting line, and we experimented with several choices (highest/lowest degree, closest pair of lines), but with a scan of *all* starting lines, we found out that the resulting clusterings had identical modularities Q for our test cases, so now we choose any (the first) line as the starting line.
(3) How to find the next line? Here also, we experimented with several strategies (closest to last one, closest to average over all previously taken, ...), and identified the best strategy to be:

$\rightarrow$ Choose the next line i among the unchosen that has the lowest pseudo-distance d_{ij} to *any* of the previously chosen lines j.

This lowest pseudo-distance is stored as the "step" into the $(N-1)$ line differences S_i, used later because the highest of those will be corners in the to be block-ordered matrix.

(S_i = lowest line difference of line i to any of the previous lines) (Please see remark (1) below).
(4) After choosing all lines sequentially, the matrix A is reordered into that sequence - into a shape that maximizes the blocks along the diagonal, because similiar nodes follow each other.
(5) As we have kept the line difference steps S_i, we can now use that information to identify the ideal positions for *cuts* between clusters; the position i of the highest line difference jump $\max_i S_i$ being the first candidate for a cutting, the second for a second cutting, etc.- this is done by hashmapping all jumps to their positions, and sorting the jump distances in falling order. Now in this order, the cutting into clusters is done and each modularity (5) for each clustering $\mathcal{C}''$ is stored. $Q(\mathcal{C}'') = Q(\#cuts)$. See figure 2 as an example. (Please see remark (2) below).
(6) The highest modularity among all these clusterings $\mathcal{C}''$ "wins", and the respective clustering $\mathcal{C}'$ is the result for this (τ, β, γ) -parameter combination, $\mathcal{C}' = \mathcal{C}(\tau, \beta, \gamma)$.
(7) Finally, the (τ, β, γ) parameter plane is scanned to find the overall best modularity Q, which corresponds to the best possible clustering $\mathcal{C}$ with $\mathcal{C} = arg\ \max_{\mathcal{C}'}\ Q(\mathcal{C}')$.

Remarks:

(1) Whenever there are *several* next lines j with identical d_{ij}, the heuristics at the moment is to choose the one (first) $k$with the highest A_{ij}direct connection weight.

(2) the cutting process interatively cuts into more clusters, the cut points chosen are the lines with (decreasing) line difference steps S_i. If there are several next lines with identical $S_i$then all of them are chosen at the same time in any order, was a very first version. Now this has been improved by a modularity calculation for all subsets of these equal-S_i-cuttings. However, subsets tend to quickly become many: If there are 3 elements in the set, 7 subsets have to be checked, but if there are 8, already 255 possible subsets have to be checked (and thus modularities calculated). This is the reason why this improved algorithm dramatically slows down for unweighted networks, and for the (τ, β, γ)= (0,0,0) case, because then the d_{ij}contains many identical entries.

2.4.1. Time complexity of the algorithm

Due to the matrix multiplication, our algorithm cannot perform better than $O(N^3)$. On the other hand, the squared matrix is always saved to disk, so many different clustering

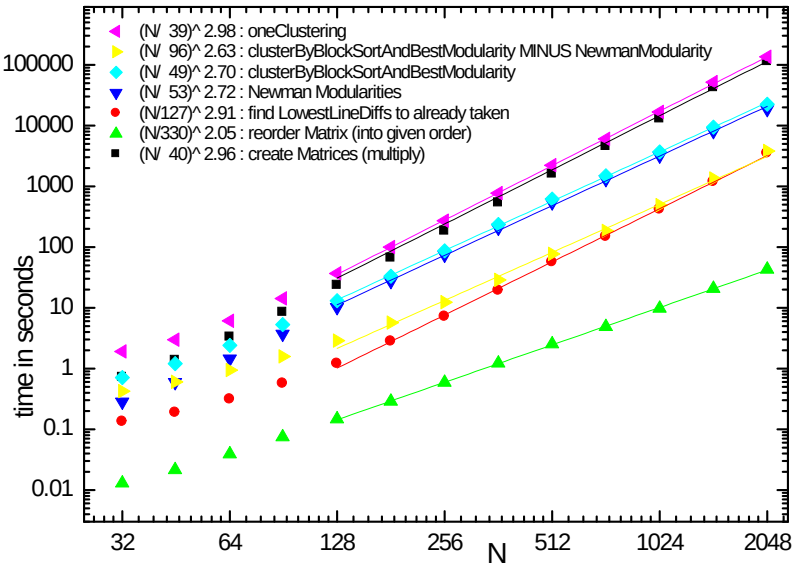


FIG. 1: Runtime measurement for *one* clustering of an ER-random graph, divided into the most important and time consuming subroutines. The overall time complexity is $O(N^3)$.

attempts can recycle these matrices, they have to be prepared only once. One idea to decrease the time (by a constant factor) would be to implement the matrix multiplication in a faster programming language than Python.

For figure 1 we have created ER-random graphs with $M = 5N$ and N up to 2048, and measured the contributions of subroutines to the total duration. Keep in mind that these times are for finding the best clustering in *one* (τ, β, γ)-point, so to locate the globally best clustering, the (τ, β, γ)-parameter space has to be scanned, too.

The calculation of the Newman modularities is done for each and every clustering, and it takes up a lot of time (second after the matrix multiplication), so any idea to reduce this time will largely improve the overall performance of the CAMBO algorithm.

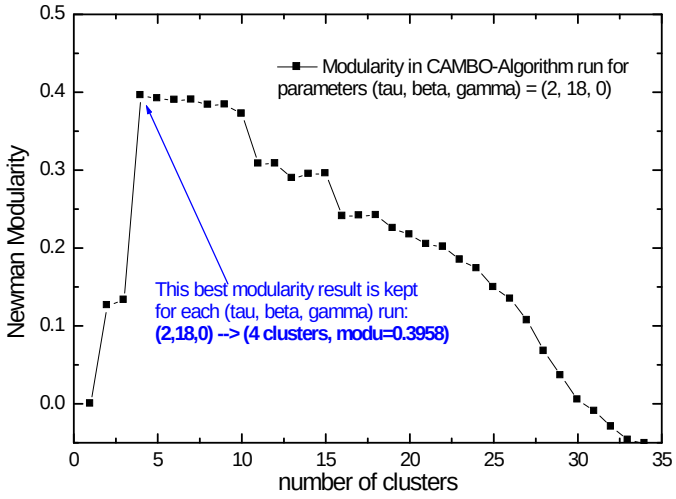


FIG. 2: One block sorting of ZACHC matrix, using the line difference parameters $(\tau, \beta, \gamma = 2, 18, 0)$. Shown are modularities for increasing number of clusters, cut in the order of decreasing line difference steps. You can clearly see that there is a maximum modularity for 4 clusters, which is kept as the end result for the $(\tau, \beta, \gamma = 2, 18, 0)$-run.

3. EXAMPLE RUNS: ZACHARY KARATE, EU-PROJECTS TOPICAL NETWORK, GENETICALLY SIMILIAR TUMORS

3.1. example 1: Zachary Karate Club splitting

This is data collected from the members of a university karate club. The ZACHC matrix indicates the relative strength of the associations (number of situations in and outside the club in which interactions occurred). Wayne Zachary ([25], [21]) 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.

The mean non-diagonal elements of A and A^2 are $\alpha_1 = 0.412$ and $\alpha_2 = 8.301$. Let's look at the $(\tau, \beta, \gamma) = (2, 18, 0)$ clustering run, figure 2 shows the modularity over the number of clusters. Recall: cuts are done in decreasing order of line difference steps. We can clearly see a maximum modularity $Q = 0.3958$ at 4 clusters, and keep these results for $(\tau, \beta, \gamma) = (2, 18, 0)$.

Now we scan all reasonable (τ, β, γ)-parameters. In figure 3, we see the parameter dependence of the modularity, to the left as a selection of series of $Q_\beta(\tau)$ functions for several

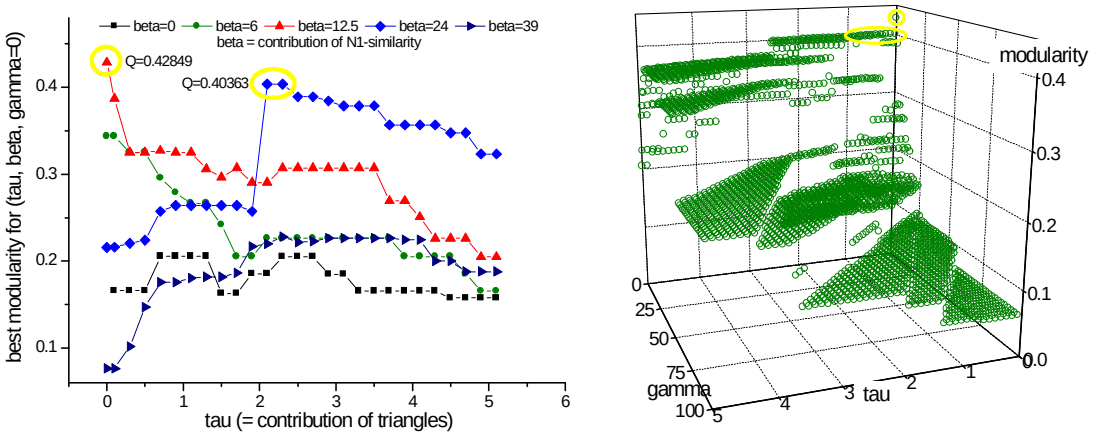


FIG. 3: Modularity Q vs (τ, β, γ) that were scanned. Left: $Q_\beta(\tau)$ for 5 different β, and $\gamma = 0$. Right: $Q_\beta(\tau, \gamma)$ for $\beta = 12.5$

selected β, to the right as a 3-dimensional mountain plot $Q(\tau, \gamma)$ with $\beta = 12.5$. The best modularities for the 3-cluster and (also empirical) 2-cluster split are circled in yellow.

In figures [4] and [5] we see some examples of these best clusterings, with the adjacency matrix density plot attached. **PIC#01**: The clustering with the **overall best modularity 0.4285** is found at $(\tau, \beta, \gamma) = (0, 12.5, 0)$; it splits into 3 clusters. **PIC#02**: The second best clustering (modularity 0.4036, with 2 clusters) **which is the empirically found splitting of the club into two groups**, is found for example at $(\tau, \beta, \gamma) = (1, 19, 0)$, and at many other (τ, β, γ)-points, e.g. $(1, 5, 10)$ or $(1, 0, 20)$. **PIC#03** gives an interesting 4-cluster solution with Q=0.3958 (we already know that from figure [2]), in which the right half of the club would split into three clusters. Look at the red nodes v12, v18, v20, v22 - their are **not** *connected to each other, but the CAMBO algorithm puts them into one cluster* for $\tau, \beta = 2, 18$ - their 2–step-connection is strong enough and N_1-neighbourhood (structural equivalence) is similiar enough to declare them into one cluster. Because the 4 nodes are not directly connected mutually, this *a good example for completely new clusterings that can be found by our algorithm.* **PIC#04** with modularity 0.3847 shows two clusters, but this time node v9 and v31 went to the other halved Karateclub. The modularity difference is only 4.7%, but the fate of the now-two clubs would have been very different, if v9 and v31 went to the other side. **PIC#05** with Q=0.3781 and 7 clusters shows that nodes v13, v10, v29 are in weaker group connection to the others, if second next neighbours and the N_1-

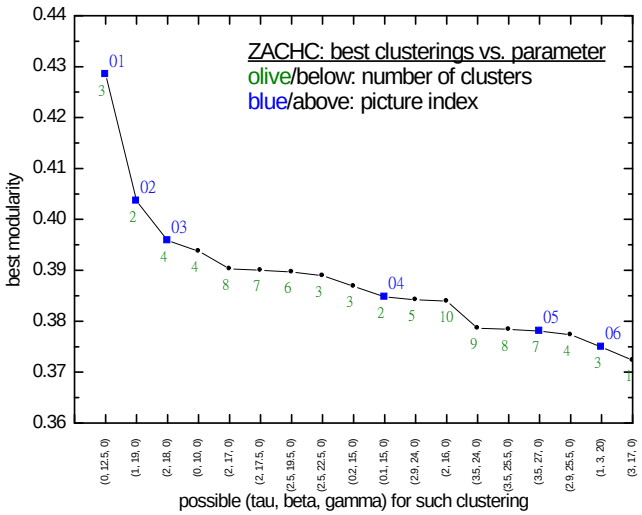


FIG. 4: The best clusterings (highest modularities) on the scanned (τ, β, γ)-manifold, with one representative given as x-axis-labels. Olive numbers below are the number-of-clusters. Blue numbers above are picture indices for figure [5].

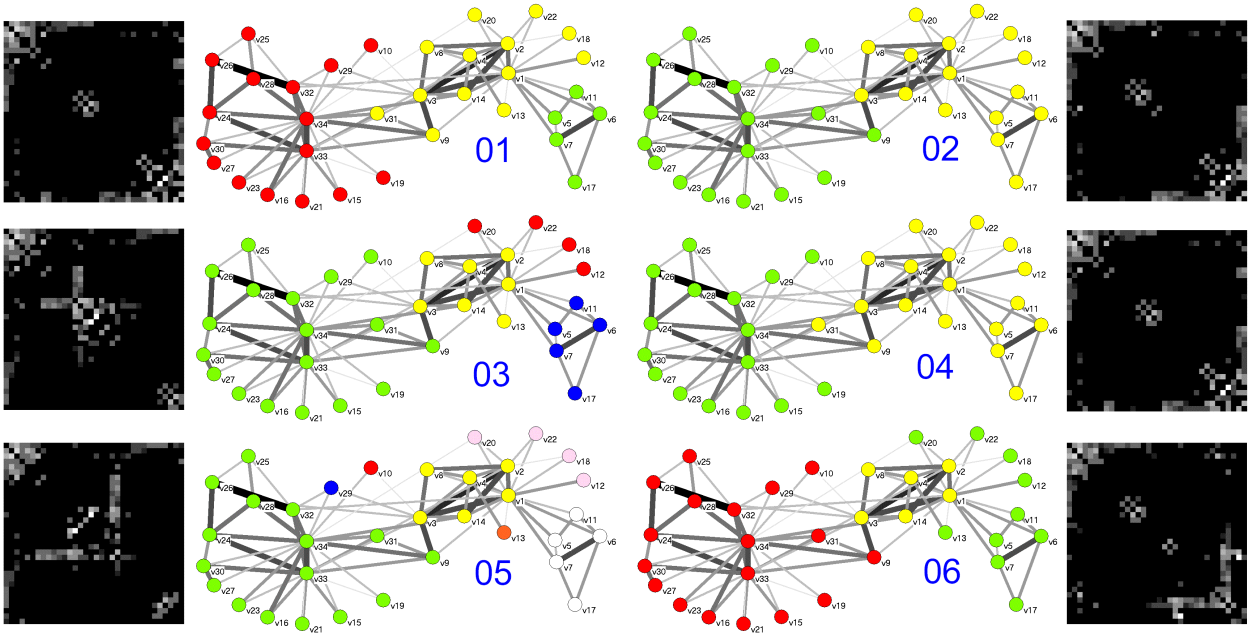


FIG. 5: Some clusterings of figure [4] made with Pajek [16]. #01 is the best modularity solution, #02 is the second best - this is how the Karate Club actually splitted into two groups.

neighbourhood are taken into account more strongly. A variation of PIC#03 is **PIC#06**; with $(\tau, \beta, \gamma) = (1, 3, 20)$. The γ (the N2-neighbourhood structural equivalence) - still part of the program mainly for historical reasons- is sometimes an intereresting dimension to look at, because we could not find *this* clustering for any (τ, β with $\gamma = 0$).

Summary: The CAMBO algorithm finds the *empirical* split of the Zachary Karate Club as the *second best* modularity solution. Because of the terms of higher order connectivity (in this case the β-term, structural-N_1-equivalence $B_{ij} = \sum_{\substack{k=1 \\ k \neq i,j}}^{N} \left| \widetilde{A}_{ik} - \widetilde{A}_{jk} \right|$) *we could identify a completely new clustering of ZACHC into 3 clusters*, with an even higher modularity of 0.4285, that has not previously been reported.

3.2. example 2: EU-projects topical network

In the EU project NEMO [13], we study the European Framework Programmes (FP) on Research and Technological Development [12]. In FP1-FP4 about 27000 projects, each described (among other data) by some of 45 given subject indices (e.g. Agriculture, Nuclear Fusion, Aerospace, ...) were the raw data. Whenever two of these 45 topics co-occurred, the edge weight between these 2 topic nodes was increased by one, e.g. "24 Scientific Research" and "25 Social Aspects" were named together in 3418 projects: $w(v24, v25) = 3418$; while "9 Agriculture" and "25 Social Aspects" were named together in 111 projects: $w(v9, v25) = 111$. Now that is enough to have a weighted network of topic-to-topic relations - which we have analyzed using the CAMBO algorithm (see figure [6]).

This network is nice for illustrational purposes as you can draw your own conclusions about the goodness of the clustering because you know all node names. The main result is that there are no real surprises, so 27000 projects cluster those topics like the common sense would do it. However, the *highest* modularity solution is the "best", but actually not the most instructive because it has only 6 clusters - and further subclustering is quite interesting, so we detuned slightly from the Newman-optimal clustering to a 2%-lower modularity. We also included this 10-clusters solution, which has the sixth best modularity among all found clusterings.

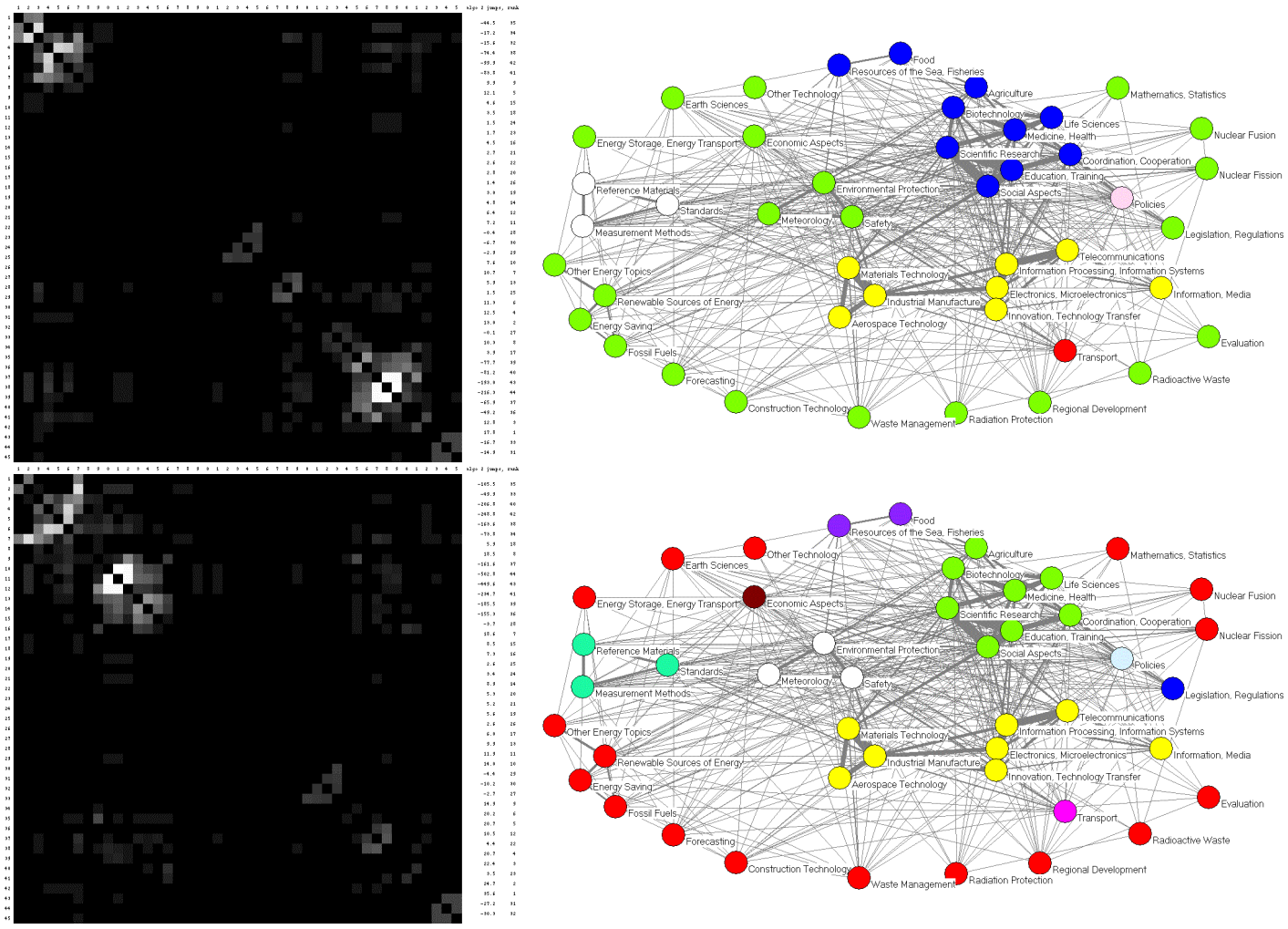


FIG. 6: Two good clusterings of co-ocurrence of topics in EU-research-projects. TOP: best clustering: $(\tau, \beta, \gamma = 10, 100, 0)$, 6 clusters, Q=0.34281. BOTTOM: 6th-best clustering $(\tau, \beta, \gamma = 25, 200, 0)$, 10 clusters, Q=0.33596

3.3. Example 3: genes and tumors

The idea for this CAMBO algorithm was born during an investigation of the Alizadeh *et. al.* [1]-lymphoma-tumor-dataset of microarray gene expression levels. From the 96×4026 tumor×gene information we generated a 96×96 tumor×tumor network by "weighted projection" (topic of a forthcoming paper) from all the gene expression levels, then analyzed this network of tumor-tumor-similarity. The best clustering so far with Q=0.06654 and 14 clusters was found at $(\tau, \beta, \gamma) = (0, 9, 5)$, the adjacency matrix is shown in figure 7 - you can

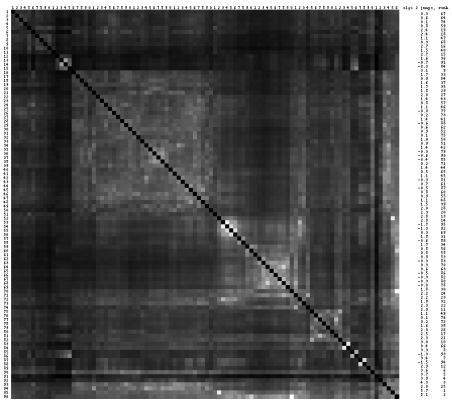


FIG. 7: Blocked adjacency matrix of tumor-tumor-similarity from Alizadeh et al. (see text).

cluster	clSize	name
1	15	ARRY22X___SUDHL5@5
1	15	ARRY85X___OCI_Ly3@0
1	15	ARRY84X___OCI_Ly10@0
1	15	ARRY83X___DLCL-0042@0
1	15	ARRY17X___WSU1@5
1	15	ARRY88X___DLCL-0052@0
1	15	ARRY20X___OCI_Ly12@5
1	15	ARRY19X___U937@5
1	15	ARRY21X___OCI_Ly13.2@5
1	15	ARRY18X___Jurkat@5
1	15	ARRY16X___OCI_Ly1@0
1	15	ARRY10X___Blood_T_cells_Adult_Na\xefve_CD4+_Unstimulated@4
1	15	ARRY11X___Blood_T_cells_Adult_Naive_CD4+_I+P_Stimulated@4
1	15	ARRY12X___Cord_Blood_T_cells_Neonatal_Naive_I+P_Stimulated@4
1	15	ARRY13X___Blood_T_cells_Neonatal_Na\xefve_CD4+_Unstimulated@4
2	36	ARRY59X___DLCL-0010@0
2	36	ARRY58X___DLCL-0037@0
2	36	ARRY60X___DLCL-0015@0
2	36	ARRY73X___DLCL-0033@0
2	36	ARRY61X___DLCL-0026@0
2	36	ARRY67X___DLCL-0002@0
2	36	ARRY86X___DLCL-0051@0
2	36	ARRY87X___DLCL-0034@0
2	36	ARRY89X___DLCL-0008@0
2	36	ARRY81X___DLCL-0031@0
2	36	ARRY82X___DLCL-0007@0
2	36	ARRY80X___DLCL-0036_OCT@0
2	36	ARRY78X___DLCL-0012@0
2	36	ARRY57X___DLCL-0018@0
2	36	ARRY70X___DLCL-0003@0
2	36	ARRY75X___DLCL-0040@0
2	36	ARRY77X___DLCL-0028@0
2	36	ARRY72X___DLCL-0048@0
2	36	ARRY68X___DLCL-0016@0
2	36	ARRY62X___DLCL-0005@0
2	36	ARRY63X___DLCL-0023@0
2	36	ARRY65X___DLCL-0024@0
2	36	ARRY64X___DLCL-0027@0
2	36	ARRY71X___DLCL-0014@0
2	36	ARRY54X___DLCL-0039@0
2	36	ARRY53X___Tonsil@2
2	36	ARRY69X___DLCL-0020@0
2	36	ARRY51X___DLCL-0006@0
2	36	ARRY95X___DLCL-0030@0
2	36	ARRY94X___DLCL-0004@0
2	36	ARRY74X___DLCL-0025@0
2	36	ARRY50X___DLCL-0032@0
2	36	ARRY93X___DLCL-0029@0
2	36	ARRY49X___DLCL-0011@0
2	36	ARRY52X___DLCL-0049@0
2	36	ARRY23X___DLCL-0041@0
3	23	ARRY24X___FL-9@6
3	23	ARRY27X___FL-10_CD19+@6
3	23	ARRY28X___FL-10@6
3	23	ARRY29X___FL-11@6
3	23	ARRY30X___FL-11_CD19+@6
3	23	ARRY31X___FL-6_CD19+@6
3	23	ARRY32X___FL-5_CD19+@6
3	23	ARRY36X___Cord_Blood_B_cells@7
3	23	ARRY35X___Blood_B_cells@7
3	23	ARRY37X___CLL-60@8
3	23	ARRY45X___CLL-13@8
3	23	["ARRY43X___CLL-71_Richter's@8"]
3	23	ARRY41X___CLL-51@8
3	23	ARRY39X___CLL-9@8
3	23	ARRY38X___CLL-68@8
3	23	ARRY42X___CLL-65@8
3	23	ARRY40X___CLL-14@8
3	23	ARRY44X___CLL-71@8
3	23	ARRY33X___Blood_B_cells_memory_CD27+@7
3	23	ARRY47X___CLL-52@8
3	23	ARRY26X___FL-12_CD19+@6
3	23	ARRY48X___DLCL-0009@0
3	23	ARRY46X___CLL-39@8
4	8	ARRY4X___Blood_B_cells_anti-IgM+IL-4_24h@3
4	8	ARRY8X___Blood_B_cells_anti-IgM+CD40L_6h@3
4	8	ARRY7X___Blood_B_cells_anti-IgM_6h@3
4	8	ARRY9X___Blood_B_cells_anti-IgM+CD40L+IL-4_6h@3
4	8	ARRY5X___Blood_B_cells_anti-IgM+CD40L+IL-4_24h@3
4	8	ARRY2X___Blood_B_cells_anti-IgM+CD40L_24h@3
4	8	ARRY6X___Blood_B_cells_anti-IgM+IL-4_6h@3
4	8	ARRY3X___Blood_B_cells_anti-IgM_24h@3
5	2	ARRY0X___Blood_B_cells_anti-IgM+CD40L_low_48h@3
5	2	ARRY1X___Blood_B_cells_anti-IgM+CD40L_high_48h@3
6	2	ARRY15X___Thymic_T_cells_Fetal_CD4+_I+P_Stimulated@4
6	2	ARRY14X___Thymic_T_cells_Fetal_CD4+_Unstimulated@4
7	2	ARRY92X___Tonsil_GC_B@1
7	2	ARRY91X___Tonsil_GC_Centroblasts@1
8	1	ARRY90X___SUDHL6@0
9	1	ARRY79X___DLCL-0021@0
10	1	ARRY76X___DLCL-0017@0
11	1	ARRY55X___Lymph_Node@2
12	2	ARRY56X___DLCL-0001@0
12	2	ARRY66X___DLCL-0013@0
13	1	ARRY25X___FL-9_CD19+@6
14	1	ARRY34X___Blood_B_cells_naive_CD27-@7

FIG. 8: : Q=0.06654 with 14 clusters.

clearly see the block-structure, and the clustering of tumors into mutually similiar tumors of 14 different tumor types is given in figure 8. Due to our limited understanding of genetics, we cannot definitely say, if our best clustering is a good clustering, but two arguments suggest it: (a) similiar tumor-*names* are clustered together, and (b) a Fruchterman-Reingold relaxation

algorithm (for visualization in Pajek [16]) places nodes *close together* that actually end up in the same cluster of our clustering.

4. SUMMARY AND OUTLOOK

We have introduced a novel approach towards clustering a network by directly ordering its adjacency matrix into a block structure. We have defined an algorithm, called CAMBO, based on this approach, and have demonstrated its efficacy through several examples. Currently, at this level of automatism, the algorithm is working properly, but is at first and foremost a proof-of-concept, and can still be optimized. Some ideas for future improvements include:

- The parameters (τ, β, γ) are exhaustively scanned to identify the best $\mathcal{C}$ among all $\mathcal{C}'$ clusterings, in order not to miss any best modularity $Q(\tau, \beta, \gamma)$. There are more goal-orientend methods (simple hill-climbing, evolutionary optimization, etc.) which will search for the best Q in the (τ, β, γ)-modularity landscape with much less control points.

- The $N(N-1)/2$ line pseudo-distances d_{ij} are now calculated for *all* possible pairs (i, j), regardless how close or far nodes i and j are. Perhaps there is a way to divide-and-conquer the whole network?

- The (τ, β, γ) do not seem to be completely independent, it may be possible to reduce the number of parameters. However, we have seen through examples that all 3 parameters seem to be important in some cases (e.g. the $(\tau, \beta, \gamma) = (0, 9, 5)$ solution of the tumors, and the $(\tau, \beta, \gamma) = (1, 3, 20)$ clustering of the Karate club), so none can simply be excluded at the moment.

- While cutting along the decreasing line difference jumps (see figure 2) to check all clusterings $\mathcal{C}''$ for one (τ, β, γ) parameter point to find the clustering $\mathcal{C}'$ with the best modularity $Q'(\tau, \beta, \gamma)$, it will save time to cut only up to a given maximal number of clusters, or down to a given modularity threshold. Moreover, the modularity can probably be recycled when each additional cut is done, because only the node-pairs that now fall in different clusters have to be removed from the modularity sum (5).

- Already calculated modularities could be cached to avoid recalculating what is already known. Modularity is (expensively) calculated from network A and clustering C. Many $(\tau, \beta, \gamma)-$points will result in the same clustering. Looking them up from, e.g., a hash-table (with the clustering as the key) instead of recalculation saves time. See figure 1 for the huge impact of the modularity calculation on the overall time complexity, for larger systems it is the second most time consuming routine (after the $O(N^3)$ matrix multiplication).

- Perhaps an additional (binary) parameter besides (τ, β, γ) may be useful, to be able to set the contribution of the direct connection A_{ij} to zero in some cases?
  Moreover, there was the suggestion [20] to use the triangle matrix $(A_{ij} \cdot \sum_k A_{ik}A_{kj})_{ij}$ which is *zero* if there are *no triangles* between i and j. Will that be better for clustering than the contributions A and A^2?

- To account for the structural equivalence of two nodes, the N_1- and N_2-difference terms (B_{ij} and C_{ij}) are added to d_{ij} (see formula (4)). A possible alternative would be to instead add the *correlation matrix* — the variance-normalized *covariance matrix* — with the two lines X and Y of the adjacency matrix treated as if they were random variables:
  $Cor(X,Y) = \frac{\sum(X-E(X))(Y-E(Y))}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\sum E(XY)-E(X)E(Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$.

- *Repulsion networks:* In principle, the method should not only be able to cluster attraction networks, in which the edge weight corresponds to the mutual attraction of the linked nodes, but also "repulsion networks", in which the edges represent the dislike, hate, or repulsion of the linked nodes. Just set the parameters α_1 and α_2 to negative values, to account for *higher* line pseudo-distance d_{ij} if the edge weight A_{ij} is higher; the difference of the A_{ij} and A^2_{ij} line distances -and thus the difference of the N_1- and N_2-neighbourhoods- will cluster "similarly hating" nodes together.

- *Several 'best' clusterings*: To have exactly one best clustering for a network makes it easier to use that result for further analysis of higher abstraction, so we do look at the overall best modularity clustering with a feeling of a certain singularity. On the other hand, in most of the cases, the second best clusterings are not much worse. Most high modularity points on the 3-parameter manifold do give an important insight into

the network; several clusterings of almost same modularity show different "realities", by dividing into sharp and non-overlapping partitions. There are ideas to combine all those "best" clusterings into one "fuzzy" clustering, in which each node carries a probability to be counted into one or the other clustering.

Summary: The CAMBO algorithm is a new $O(N^3)$-slow but well-working *deterministic* clustering algorithm, with *easily interpretable parameters* for mixing the direct connection weight, 2-step-paths influence, and N_1- and N_2-structural equivalence into a heuristic for ordering the nodes, and by Newman modularity maximization, for finding a best clustering. An implementation, consisting of approximately 5000 lines of Python [17] code, will be made available online [7]. Please mail to me, if you use or improve the idea of this algorithm.

Acknowledgement 1 *This work has been supported in part by the European FP6-NEST-Adventure Programme, contract number 028875. Andreas Krueger would like to thank: The Volkswagen Foundation, and the Systems Research unit of the Austrian Research Center, for financing this PhD; Tyll Krueger and Philippe Blanchard for our always-inspiring working atmosphere; Madeleine Sirugue-Collin, Pierre Chiappetta for fruitful discussions about genes and tumors, Michael Barber by far not only, but especially for persuading me to use Python, Thomas Roediger-Schluga for pursuing the idea of a project between social and natural sciences, and Hanne for being the good soul of our department.*

[1] Alizadeh et. al.: *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*, Nature **403**: 503-511 (2000), http://llmpp.nih.gov/lymphoma

[2] A.L.Barabasi, R.Albert, *Emergence of Scaling in Random Networks*, Science 286, 509 (**1999**)

[3] A.L.Barabasi, R.Albert, R.Albert, *The diameter of the world-wide web*, **1999**, Nature (London) 401, 130-131; cond-mat/9907038

[4] R.Albert , A.-L. Barabasi : *Statistical Mechanics of Complex Networks*, Reviews of Modern Physics, 74, 47 (**2002**), arXiv:cond-mat/0106096

[5] *BOOST*, collection of C++ libraries, www.boost.org

[6] F.Boutin, M.Hascoet, *Cluster Validity Indices for Graph Partitioning*, Proceedings of the Conference on Information Visualization IV'**2004**, http://www.lirmm.fr/~mountaz/Publi/iv2004_boutin_hascoet.pdf

[7] *CAMBO*-Website will be linked from http://www.AndreasKrueger.de/networks

[8] E. Cuthill and J. McKee, *Reducing the bandwidth of sparse symmetric matrices.* Proceedings of the 24th National Conference of the ACM, **1969**, recited after: http://www.boost.org/libs/graph/doc/bibliography.html#cuthill69:reducing_bandwith

[9] L. Danon, J. Duch, A. Diaz-Guilera, A. Arenas: *Comparing community structure identification.* Journal of Statistical Mechanics: Theory and Experiment, **2005**

[10] S.N. Dorogovtsev, J.F.F. Mendes, *The shortest path to complex networks*, **2004,** arXiv:cond-mat/0404593

[11] King, I. P., *An automatic reordering scheme for simultaneous equations derived from network analysis.*, Int. J. Numer. Methods Engrg. 2, pp. 523-533, **1970**, recited after: http://www.boost.org/libs/graph/doc/bibliography.html#king70

[12] M. Barber, A. Krueger, T. Krueger and T. Roediger-Schluga: *The Network of EU-Funded Collaborative R&D Projects,* Phys. Rev. E 73, 036132 (**2006**), arXiv: physics/0509119

[13] *NEMO - Network Models, Governance and R&D collaboration networks*, EU-project contract#028875, http://www.nemo-net.eu

[14] M.E.Newman, M.Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E 69, 026113 (**2004**), arXiv:cond-mat/0308217

[15] Clauset, Newman, Moore, *Finding community structure in very large networks*, Phys. Rev. E

70, 066111 (**2004**), arxiv:condmat/0408187

[16] V.Batagelj, A.Mrvar, *Pajek - Program for Large Network Analysis*, http://vlado.fmf.uni-lj.si/pub/networks/pajek

[17] http://www.python.org

[18] J. Reichardt, S. Bornholdt, *Detecting fuzzy community structures in complex networks with a Potts model*, Phys. Rev. Lett. 93, 218701 (**2004**), arxiv:cond-mat/0402349

[19] S. W. Sloan, *An algorithm for profile and wavefront reduction of sparse matrices*, Int. j. numer. methods eng., 23, 239 - 251 (**1986**), recited after: http://www.boost.org/libs/graph/doc/sloan_ordering.htm

[20] Tyll Krueger, personal communication, 8.6.**2007**

[21] *UCINET 6 - Social Network Analysis Software*, http://www.analytictech.com/ucinet/ucinet.htm

[22] Stijn van Dongen, *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, May **2000**, http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm

[23] Stijn van Dongen, *MCL = Markov Cluster Algorithm*, http://micans.org/mcl

[24] D.J.Watts, S.H.Strogatz, *Collective dynamics of 'small-world' networks*, Nature 393, 440 (**1998**)

[25] W.Zachary (**1977**). An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33, 452-473.