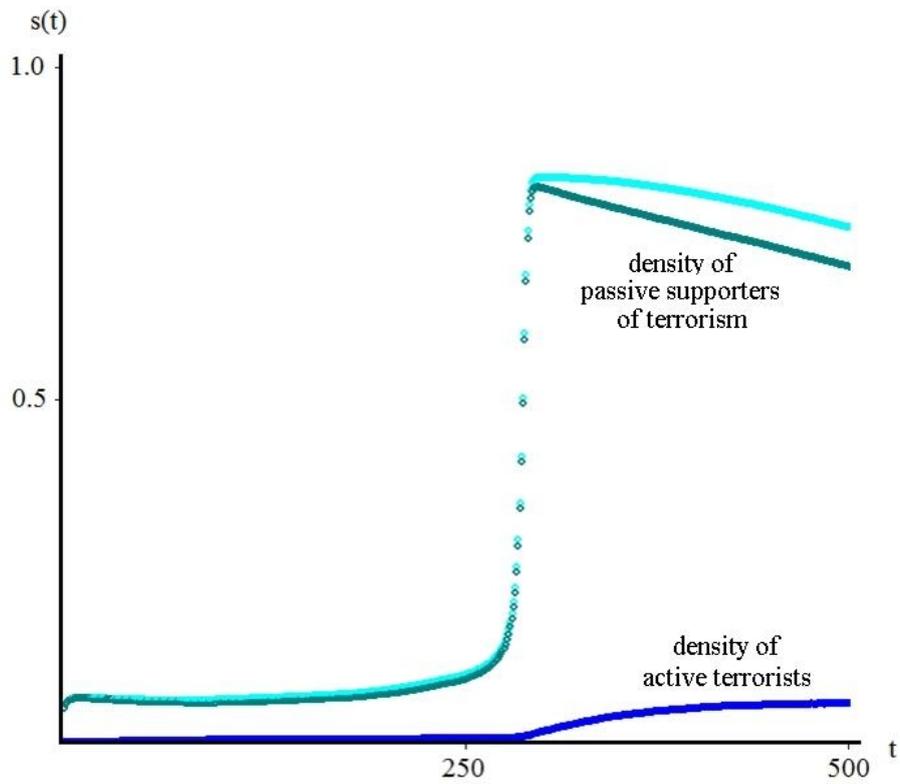


Random Networks, Threshold Models and Social Dynamics

Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Physik
der Universität Bielefeld
vorgelegt von

Sascha Delitzscher

20.10.2011



Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 4 |
| 1.1 | What is Sociophysics | 4 |
| 1.2 | Real-World Networks | 5 |
| 1.3 | Structure of the Thesis | 7 |
| 1.4 | Random Graphs and Network Theory | 8 |
| 1.4.1 | Erdős-Renyi Random Graphs | 9 |
| 1.4.2 | Scale-Free Networks | 10 |
| 1.5 | Threshold Models in Social Sciences and Economics | 11 |
| 1.5.1 | The Schelling Model of Segregation | 11 |
| 1.5.2 | The Granovetter Threshold Model of Collective Behavior | 13 |
| 1.5.3 | Complex Contagion | 14 |
| 1.5.4 | Information Cascades in Complex Networks | 16 |
| 1.5.5 | Adoption of Innovation | 16 |
| 2 | Inhomogeneous Sparse Random Graphs | 18 |
| 2.1 | Definitions and Properties | 18 |
| 2.2 | Additive and Multiplicative Coupling | 20 |
| 2.3 | The Phase Transition of Additive and Multiplicative Coupling | 21 |
| 3 | The Communication Index of Graphs | 24 |
| 3.1 | Basics | 24 |
| 3.1.1 | Definitions and Simple Properties | 25 |
| 3.1.2 | The CI and Computational Complexity | 27 |
| 3.1.3 | Inhomogeneous Sparse Random Graphs revisited | 28 |
| 3.2 | The Communication Index and Peer-to-Peer Networks | 28 |
| 3.2.1 | The CI as an a-priori Measure for Effectiveness | 29 |
| 3.2.2 | Paper 1: Optimizing Topology in Bit Torrent Based Networks | 30 |
| 3.3 | The Communication Index of Random Graphs | 32 |
| 3.3.1 | The Galton-Watson Process | 32 |
| 3.3.2 | The Erdős-Renyi Random Graph | 34 |
| 3.4 | Numerical Results | 36 |
| 3.4.1 | The Galton-Watson Process | 37 |
| 3.4.2 | The Erdős-Renyi Random Graph | 39 |
| 4 | Paper 2: Time-Ordered Information Processing on the Binary Tree | 41 |
| 4.1 | Overview | 41 |
| 4.2 | Results | 41 |
| 5 | Paper 3: Generalized Epidemic Processes and Threshold Percolation with Application to Knowledge Diffusion | 43 |
| 5.1 | Active and Passive Knowledge | 43 |
| 5.2 | Generalized Epidemic Processes | 44 |
| 5.3 | Results | 45 |

| | | |
|----------|---|-----------|
| 6 | Paper 4: Passive Supporters of Terrorism and Phase Transitions | 48 |
| 6.1 | The Model | 48 |
| 6.2 | Results and Conclusions | 49 |
| 7 | Danksagung | 52 |

1 Introduction

1.1 What is Sociophysics

Sociophysics is the branch of physics concerned with human behavior and society. This definition might seem a bit contradictory and strange to social scientists and physicists who never came in touch with this emerging field in nature science. The Oxford Dictionary defines physics as "the branch of science concerned with the nature and properties of matter and energy"¹, sociology as "the study of the development, structure, and functioning of human society"², while there is no entry for sociophysics, yet. This thesis and the sociophysical papers included may be seen as a step forward to change this and to elucidate how physical methods and techniques can help to understand human behavior.

As in every scientific theory it is almost impossible to date its birth. The oldest papers in sociophysics the author of this thesis is aware of are those of Weidlich, [41], and Callen and Shapiro, [8], from the early seventies, both found cited in [20]. Then it took almost ten years when Galam, joined by Gefen and Shapir, [21], came up with a paper where they modeled the process of strike in big companies using an Ising ferromagnetic model in an external reversing uniform field, and included a call to the creation of sociophysics.

All of the above examples show that physics is capable of even more than explaining the behavior of inanimate matter. Note that the aim of sociophysics is not to understand and explain the laws of individual behavior. Although individuals may be seen as particles driven by external forces, explaining typical behavior may fail due to lusts and instinct that drives individuals into irrational decisions. Nevertheless, these kinds of effect are belonging to the field of psychology rather than sociology, and sociophysics may not be the right approach to deal with such problems. But whenever a certain number of individuals congregate to achieve a mutual goal, unexpected and deviant behavior may get averaged out, and one of social sciences task is it to find models to explain collective behavior of individuals.

So, the situation is similar to the origin of statistical mechanics where physicists began to use methods of probability theory and statistics to explain the "collective behavior" of systems consisting of billions of single particles, e.g. gases, interacting with each other and resulting in a "global" behavior, all particles acting as a single entity, with single motions states becoming irrelevant to the outcome of measurements of global variables like temperature, velocity, or energy.

In this context, social science can be viewed as an addition to the experimental side of physics, providing measurements and falsification experiments. For instance, social scientists observed critical mass phenomena when using threshold models (see chapter 1.5), for most of these it can be shown that assuming a threshold in decision making implies a critical mass, i.e. if the number of decision takers succeeds a certain critical value almost every other individual will follow and also take that decision.

Of great importance in sociophysics are the underlying networks on which social dynamics are investigated. Usually, the social network of a society is identified with a graph $G(V, E)$, with vertex set V representing the individuals

¹<http://english.oxforddictionaries.com/definition/physics>

²<http://english.oxforddictionaries.com/definition/sociology>

of the society, and an edge set E whose elements represent social connections between the individuals. The graph G defined that way is not unique and depends sensitively on the definition of the individuals of a society and the definition of when they are connected. Both sets, V and E , may be introduced in various ways and may have a completely different topology depending on what definitions were chosen.

For instance, the vertex set of a society can be defined as the set of all people who can verify their identity as a civilian in a certain country, similarly it can be defined as the set of all people who are staying in that country at the time of the definition, or, more restrictive, it can be defined as the set of all people who are in hired labor.

While the elements of the vertex set are despite the manifold of possibilities precisely definable, for a real world society it is more complicated to define which vertices are linked to each other. The social network is designed to represent one or more special features of this society. According to choice and desired precision topologically very different networks may arise. For instance, the network of blood relation and the network of sexual contacts yield disjoint edge sets, in general. And none of these represents society and its inherent social interactions meaningfully.

Other examples for social networks are the movie actor database where any two actors are linked if they played together in a movie, and co-authorship networks where authors are linked if they have a joint publication. Nevertheless, there are many real-world networks where the vertex set does not consist of human beings, like the World Wide Web where the vertices are web documents connected with direct hyperlinks, and the Internet where the nodes are routers and computers which are connected if they are physically connected by wires and cables. All of these have common properties which get surveyed in the next section.

1.2 Real-World Networks

When finding similar behavior or characteristics in different systems physicists tend to ask "Is there a general principle which all these systems have in common and how can it be formalized?" With the development of faster and more powerful computers, and of new methods and algorithms in network theory, leading to new possibilities of evaluating large sets of data of real world networks, striking similarities between most of the investigated networks were uncovered. There are three main observables³ which most real networks seem to share: small average path length, power law tail of the degree distribution, and high cluster coefficient.

The cluster coefficient is a measure for how many neighbors of a node are neighbors as well, on average. Social networks tend to form cliques, as it is highly probable that two friends of an individual are also friends. Thus, the class of social networks within the real networks usually has a high cluster coefficient, magnitudes higher than random networks.

The small average path length is also known as the 'six degrees of separation', a popular hypothesis which states that everyone on earth is connected to everybody else by on average six steps along paths in the social network of

³See chapter 1.4 for definitions and vocabulary.

| Network | Size | $\langle k \rangle$ | γ_{out} | γ_{in} | Reference | Nr. |
|-----------------------------|------------------|---------------------|----------------|---------------|------------------------------------|-----|
| WWW | 325 729 | 4.51 | 2.45 | 2.1 | Albert, Jeong, and Barabási 1999 | 1 |
| WWW | 4×10^7 | 7 | 2.38 | 2.1 | Kumar <i>et al.</i> , 1999 | 2 |
| WWW | 2×10^8 | 7.5 | 2.72 | 2.1 | Broder <i>et al.</i> , 2000 | 3 |
| WWW, site | 260 000 | | | 1.94 | Huberman and Adamic, 2000 | 4 |
| Internet, domain* | 3015–4389 | 3.42–3.76 | 2.1–2.2 | 2.1–2.2 | Faloutsos, 1999 | 5 |
| Internet, router* | 3888 | 2.57 | 2.48 | 2.48 | Faloutsos, 1999 | 6 |
| Internet, router* | 150 000 | 2.66 | 2.4 | 2.4 | Govindan, 2000 | 7 |
| Movie actors* | 212 250 | 28.78 | 2.3 | 2.3 | Barabási and Albert, 1999 | 8 |
| Co-authors, SPIRES* | 56 627 | 173 | 1.2 | 1.2 | Newman, 2001b | 9 |
| Co-authors, neuro.* | 209 293 | 11.54 | 2.1 | 2.1 | Barabási <i>et al.</i> , 2001 | 10 |
| Co-authors, math.* | 70 975 | 3.9 | 2.5 | 2.5 | Barabási <i>et al.</i> , 2001 | 11 |
| Sexual contacts* | 2810 | | 3.4 | 3.4 | Liljeros <i>et al.</i> , 2001 | 12 |
| Metabolic, <i>E. coli</i> | 778 | 7.4 | 2.2 | 2.2 | Jeong <i>et al.</i> , 2000 | 13 |
| Protein, <i>S. cerev.</i> * | 1870 | 2.39 | 2.4 | 2.4 | Jeong, Mason, <i>et al.</i> , 2001 | 14 |
| Ythan estuary* | 134 | 8.7 | 1.05 | 1.05 | Montoya and Solé, 2000 | 14 |
| Silwood Park* | 154 | 4.75 | 1.13 | 1.13 | Montoya and Solé, 2000 | 16 |
| Citation | 783 339 | 8.57 | | 3 | Redner, 1998 | 17 |
| Phone call | 53×10^6 | 3.16 | 2.1 | 2.1 | Aiello <i>et al.</i> , 2000 | 18 |
| Words, co-occurrence* | 460 902 | 70.13 | 2.7 | 2.7 | Ferrer i Cancho and Solé, 2001 | 19 |
| Words, synonyms* | 22 311 | 13.48 | 2.8 | 2.8 | Yook <i>et al.</i> , 2001b | 20 |

Figure 1: Real world networks, $\langle k \rangle$ is the average degree, $\gamma_{in,out}$ is the exponent of the in-/out degree distribution. For networks which are undirected holds $\gamma_{in} = \gamma_{out}$. Figure is taken from [2].

earth's society. In 1969, Stanley Milgram conducted an experiment to test this hypothesis, [29]. He instructed 296 individuals of the United States to send a letter to a stockbroker in Boston, but with the additional rule that the letter must be sent only to persons which they know personally. Some information was given to the test persons, like full name, address, hometown, profession, so the subjects could guess which of their friends may be most appropriate to send the letter further on. Of the 296 only 64 reached their target in Boston, within this group the mean number of steps taken is 5.2, indicating a six degrees of separation effect on at least the United States.

Nevertheless, much criticism followed the publication of the results, for the test persons were not chosen completely random and the large fraction of letters which never reached their target leads to an underestimate of the actual average path length.

In 1998, Watts and Strogatz presented a theoretical model for random graphs with small average path length and high cluster coefficient which became very popular, [40]. They start with a regular ring lattice where each node is connected to a fixed number of nearest neighbors, then rewired each bond with a probability p . This procedure ensures the creation of shortcuts between originally distant vertices, yielding small path lengths, while the originally large cluster coefficient does not get significantly smaller.

What the Watts-Strogatz model can not provide is the power law tail of the degree distribution, as mentioned before, a remarkable property most real world networks seem to share, see fig. 1. The table depicted there shows a variety of networks which were investigated. All show the power law tail in the degree distribution which is separated into in-degree distribution and out-degree

distribution for directed networks. Note that independently of nature and of size, all exponents range between 1 and 3.4, indicating that there may be an underlying principle for the growth of such networks which not only explains the power law, but even why the exponents are distributed in such a small range.

Albert and Barabási used (and coined) preferential attachment to explain the structure of the World Wide Web and found a striking principle which leads to an inverse power law distribution. Unlike in the random graph models used by Erdős and Renyi, see chapter 1.4, or Watts and Strogatz, they used a graph growing procedure instead of working with a fixed number of vertices and rules for linking and rewiring. They considered a starting graph G_0 on n vertices and added one vertex per time step with a fixed number of edges which get linked to vertices in the existing graph with probability proportional to the degree of these vertices. They go on showing that this procedure leads to a power law degree distribution with exponent $\gamma = 3$.

Several refinements (like the Chung-Lu model, see chapter 1.4.2) and other mechanisms yielding power laws have been found since then, e.g. the cameo principle, [4]. For more information on real-world networks and phenomena see the detailed survey of Costa et al. [12].

1.3 Structure of the Thesis

This thesis proceeds with an introduction to random graphs and network theory to provide the vocabulary necessary to follow the presentation of all results later on. Since many of these are concerning threshold models, section 1 closes with an overview of threshold models which have already been proposed and investigated in social sciences and economics.

In section 2, two particular network structures of great importance in socio-physics are analyzed, network structures where links between nodes are formed in different ways, but may yield the same degree distribution. The following discussion makes use of a formalism developed by Bollobás, Janson, and Rioridan to treat inhomogeneous random graphs in great generality. This formalism gets introduced in the first subsection, and then used to derive a comparison of the phase transition of these two network structures.

Section 3 is devoted to the development of the theory of the communication index of graphs. The communication index is a measure for effectiveness of communication in graphs. Basic properties get proved, and an application to peer-to-peer networks is discussed, where under some simplifying assumptions the communication index provides the optimal network structure for such networks in terms of download latency. A joint paper has been published, the work contributed to this paper by the author of this thesis is summarized. The section closes with a calculation of the communication index for Galton-Watson trees, and with the presentation of some numerical results obtained due to the work of Andreas Krueger.

The second half of this thesis then provides summaries of the authors contribution to some publications, all of them attached to the dissertation. Section 4 introduces the concept of time-ordered information processing on networks, with calculations on the binary tree. In section 5 we turn our view to threshold models, using a generalized epidemic process to model the dissemination of knowledge in complex networks. Finally, section 6 presents a variation of this generalized epidemic process to apply the model to the spread of terrorism.

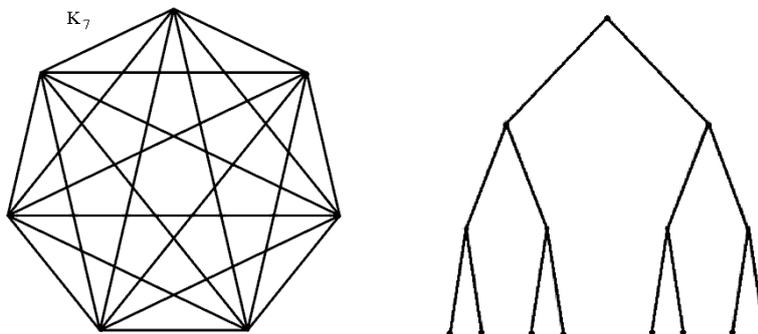


Figure 2: Left: The complete graph K_7 . Since each vertex $v \in V$ has the same, maximal degree $\deg(v) = 6$, the degree distribution is the Kronecker-Delta $p_k = \delta_{k6}$. Right: The binary tree on 15 vertices. The degree distribution is given by $p_2 = 1/15$, $p_3 = 14/15$, $p_k = 0$, $k \neq 2, 3$.

1.4 Random Graphs and Network Theory

This section is devoted to briefly introduce the vocabulary of random graph and network theory. For a more detailed overview see Durrett's book, [17], or the review paper of Albert and Barabási, [1]. With the Erdős-Renyi random graphs, small-world networks, and scale-free networks, the most important classes of random graphs and their properties are presented.

First of all, a graph $G = (V, E)$ (synonymously called network) is a pair, consisting of the so called *vertex set* V and the *edge set* E . In general, V is any finite set with n elements (also called *order* of G), the edge set E is a subset of $\binom{V}{2}$, the set of all unordered pairs (v, w) (sometimes denoted by $v \sim w$) of elements v, w of V . If $(v, w) \in E$ then v and w are called *neighbors*. For each vertex $v \in V$ the number $\deg(v) := |\{w \in V : v \sim w\}|$ of neighbors of v is called the *degree* of v . An important quantity to study in graphs and networks is the *degree distribution* $\{p_k\}$, where p_k is defined as the probability that a randomly drawn vertex has degree k . For non-random graphs the p_k are given by the fraction of nodes with degree k . For random graphs the definition of the p_k says that the degree is a random variable distributed with p_k . Note that the degree distribution of any realization of a random graph yields just a sample of the probability distribution, rather than the exact distribution itself.

Figure 2 shows two special cases of graphs, the *complete graph* on 7 vertices, where the edge set consists of every possible edge, and the *binary tree*. Trees are graphs with n vertices, $n - 1$ edges, and without cycles, i.e. it is not possible to find any closed path between two vertices.

Other important quantities are distance and diameter. The *distance* between two vertices $v, w \in V$ is defined as the length of the shortest path between v, w . The *diameter* of a graph is then defined as the maximum distance between two vertices. There is an issue with graphs which are not connected. A graph is *connected* if there is a path along edges between any two vertices. For each $x \in V$ the subset $C_x \subset V$ of all vertices which are connected to x is called the *component* of x . In case of non-connected graphs the distance of vertices of

disjoint components may be defined as infinity or as the maximum diameter of all components of G (which are, by definition, connected).

1.4.1 Erdős-Renyi Random Graphs

The first to study random graphs were Hungarian mathematicians Erdős and Renyi. In their famous articles [18], [19], they construct random graphs in two different ways. At first they consider the set of all graphs with n nodes and m edges, denoted by $G(n, m)$, and study properties of any graph G drawn uniformly at random from $G(n, m)$. The other, nowadays more common variant is to define the probability space formed by all graphs with n vertices, and where every edge is present with a given probability p . The set of such graphs is denoted by $G(n, p)$. Note that for many graph properties, e.g. connectedness, the $G(n, p)$ and $G(n, \binom{n}{2}p)$ asymptotically ($n \rightarrow \infty$) yield the same results.

The degree distribution of graphs from $G(n, p)$ follows a binomial distribution. Since the probability that a vertex has exactly k edges is given by $p^k(1-p)^{n-1-k}$, and there are $\binom{n-1}{k}$ ways to arrange these edges, the degree distribution is given by

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

Hence, the average degree of all vertices is given by $\langle k \rangle = np$. Note that for $n \rightarrow \infty$, with $np \rightarrow c$ for some constant $c > 0$, the binomial distribution approaches a Poisson distribution with mean c .

Also Erdős-Renyi random graphs show a critical phenomenon. Choosing $p = p(n)$ to be dependent of the graph size, one can consider the limit behavior of $G(n, p(n))$ for $n \rightarrow \infty$. A certain property Q is said to hold *almost surely* if the probability of having Q tends to 1 as n tends to infinity. An important property of random graphs to be studied is the appearance of certain subgraphs. Consider a random graph $G(n, p)$ and a graph F consisting of k vertices and l edges. Since the k vertices can be chosen in $\binom{n}{k}$ ways and arranged in $k!$ ways (up to isomorphism), the expected number $E(k, l)$ of such subgraphs F is proportional to

$$E(k, l) \propto \binom{n}{k} k! p^l \propto (pn^{k/l})^l, \quad (1)$$

indicating that it is sufficient to choose $p(n) \propto n^{-k/l}$ to obtain a nonzero number E of such subgraphs.

Two important special cases should be mentioned. In case of $l = k - 1$, $E(k, k - 1)$ is the expected number of trees on k vertices. From (1) it can be seen that when $k = l$, corresponding to graphs which are cycles, the probability of having cycles and trees of all orders jumps discontinuously from 0 to 1.

Another classic question is that of the emergence of a giant connected component which also exhibits a phase transition. For any $G(n, p)$ denote by $c = np$ the average degree and consider $p = p(n) = c/n$, and large n . If one thinks of the random graph as grown from branching processes with each vertex as a branching root, then for $c < 1$ each of these branching processes dies out quickly and the resulting graph consists of rather small components. On the other hand, if $c > 1$, there is a non-zero probability that the branching process survives for

infinity, and combine each other to yield a giant connected component. It can be shown that for $c < 1$ none of the single components of $G(n, c/n)$ is larger than $O(\log n)$ vertices, while for $c > 1$ there is almost surely a giant component of size $O(n)$, and the second largest component is of size $O(\log n)$. Rigorous proofs can be found in the book of Bollobás on random graphs, see [6].

1.4.2 Scale-Free Networks

As mentioned earlier, many real world networks have the property of being scale-free, i.e. rather than the Binomial distribution of Erdős-Renyi random graphs their degree distribution $\{p_k\}$ follows a power law $p_k = ck^{-\gamma}$, at least for large degrees k . Thus, alternative concepts are needed to model real world networks.

A popular construction of a random graph model which leads to a power law degree distribution was done by Barabási and Albert in [2]. They used an approach different from the existing paradigm that used a fixed number of vertices and a couple of rules to randomly connect them. In their model the random graph grows by adding new vertices at a fixed rate to the network. These new vertices are assumed to have initially a number of m edges which connect to the other vertices already in the network. Also different from existing models they assumed that the probability that a new vertex connects to another vertex is not independent of the degree of the "older" vertex. They observed that most real networks exhibit *preferential attachment*, a connection rule which states that the probability of connection is directly proportional to the degree of the older vertex. Formally, the probability $P(d_v)$ that a new vertex connects one of its m edges to a node v with degree d_v is given by

$$P(d_v) = \frac{d_v}{\sum_{v \in V_t} d_v}. \quad (2)$$

The sum ranges over all vertices which are at the moment t in the vertex set V_t , the newly introduced vertex is excluded. In several different ways, using continuum theory, master equation, and rate equation, they derive the degree distribution from equation (2) to be (asymptotically, $t \rightarrow \infty$)

$$p_k \propto 2m^2 k^{-3}.$$

The obvious drawback of this result, the only value for the exponent which can be derived is 3, can be adjusted by changing the model in the following way, as done by Chung and Lu, [11]. Choose a probability p and an initial graph G_0 - usually consisting of one vertex with a self loop to yield degree 1 - and consider the following two operations: The vertex-step, which gets performed with probability p , where one vertex gets added to the graph and connected with exactly one another vertex proportional to its degree. Otherwise perform the edge-step, where a new edge is formed between independently drawn vertices with probability proportional to their degree.

Chung and Lu show that this procedure yields a power law degree distribution $p_k = k^{-\beta}$, with

$$\beta = 2 + \frac{p}{2-p} \in [2, 3].$$

For more information on this topic see [7].

1.5 Threshold Models in Social Sciences and Economics

This section is devoted to give a brief overview of how threshold models emerged and were used in social sciences and economics. For the sake of linguistic precision the descriptions given here follow in part the statements made in the cited publications. All figures are of the own work of the authors thesis.

All these models have in common that individuals are given an attribute, the so called *threshold*, which measures the (un-)willingness of that individual to take a certain binary decision. Threshold models of collective behavior postulate that an individual engages in a behavior based on the number of other individuals already engaged in that behavior. In the early 1970s, scientists tried to find an approach to explain crowd behavior, superior to existing models which were based on arguments of institutionalized norms and values. These were lacking aspects of how individual preferences in norms and values are interacting and aggregating to finally produce the collective behavior observed.

Notice that although all models presented here have an individual threshold condition in common, some use absolute values for the threshold and some use fractions (of a well-defined neighborhood). The main difference is that models using absolute values only take account of how many neighbors have already taken the decision regardless of how many did not.

1.5.1 The Schelling Model of Segregation

Schelling, [34] [35] [36], came up with a model of residential segregation using the idea of behavioral thresholds which are thresholds of individuals for moving out of the neighborhood as a function of the number of neighbors of their own color, which also do so. He discusses a spatial linear model to describe how segregation is generated by individuals living next to next in a street with preferences in how a satisfying neighborhood should look like. He assumes that there is a dichotomy dividing households in two classes, namely black and white⁴.

Thus, a street populated with black or white households can be described by a string x of 0's and 1's of length n , where n is the number of houses, $x = x_1 \cdots x_n \in \{0, 1\}^n$. The *neighborhood* of x_i is defined as the four next nearest neighbors on each side, making a total of 8 neighbors for each number not too close to the end of the string. For such numbers, the neighborhood is just defined as the four neighbors toward the center plus the one, two or three outboard neighbors. A number x_i is said to be *satisfied* if at least half of the neighbors have the same value as x_i . Note that this is exactly what we call a (fractional) threshold. Clearly, the fractional nature of the thresholds here make no difference, since each individual has the same number of neighbors and in this case each fractional threshold can be translated into an absolute threshold.

The moving algorithm works as follows. Each unsatisfied x_i will move to the next place, where the satisfaction condition is met. Next here means taking the least number of steps. Schelling does not clarify what should happen if there are two satisfactory new positions in the string, one left of x_i and one right, at the same distance. We assume that numbers then prefer the boundary, thus moving away from the center. Movement of the x_i happens in order from left to right, whenever another x_j turns from unsatisfied to satisfied it will not

⁴One could also label these two classes with "0" and "1" or green and blue, the labeling black and white has historical reasons for Schelling's analysis of mixed racial neighborhoods.

```

      . . . . .
init: 00110110001001000111100000010000010011011010111000

      . . . . .
1.it: 011111000000000001111110000000000000011111111100000

      . . . . .
2.it: 111110000000000001111110000000000000011111111100000

```

Figure 3: Illustration of the Schelling model on a chain ("Linear Distribution") of $n = 50$ binary numbers $x_i \in \{0, 1\}, i = 1, \dots, n$. Each number represents an individual of one out of two colors, with densities here $P(x_i = 0) = 0.6$ and $P(x_i = 1) = 0.4$. The first row shows a random sample of a Bernoulli process with $p = 0.5$, dotted numbers are dissatisfied with their neighborhood. The second row shows the first iteration of the moving algorithm. All but the most left numbers are satisfied now, thus a second iteration is necessary, depicted in the third row, where all numbers are satisfied, which constitutes an equilibrium state of the system.

move. Any x_j that turns from satisfied to unsatisfied waits until all initially unsatisfied numbers have moved. Then a second iteration starts with the same rules as before.

Figure 3 shows an example of the algorithm for a binary string of length 50. The string is a random sample taken from a Bernoulli process with $p = 0.5$. It can be observed that already after one single step, all but one x_i are satisfied and the structure of x has changed from random to clustered. Thus, the algorithm produces local segregation. By varying the size of the neighborhood he finds that cluster size correlates with neighborhood size, i.e. the larger neighborhood is defined (with threshold kept fixed at 50 percent) the larger the resulting non mixed clusters are.

Furthermore, Schelling mentions the phenomenon of the *tipping point*, which describes the behavior of some neighborhoods to 'tip' from being mainly white to mainly black. Among sociologists and economists, see [24], [28], it is widely believed that there is a critical value in the number of residents of different color which move into a monochromatic neighborhood and causing original residents to move out. Mayer, [28] observed tipping in a neighborhood of about 700 single family homes. A few houses were sold to black people in 1955. 'The selling of the third house convinced people that the neighborhood was destined to become mixed.' A year later 40 houses had been sold to black people and everyone defined the neighborhood as mixed. Still it was unclear whether the neighborhood would become completely black. But two years later the percentage had gone above 50% and the end result was no longer questioned, Schelling writes.

In physical terms, a phase transition (from pure white to pure black) occurred and a critical mass was needed. Critical mass phenomena are striking features of threshold models and can be derived from the individual thresholds and network dependent quantities.

1.5.2 The Granovetter Threshold Model of Collective Behavior

Granovetter [22] generalized Schelling's ansatz to describe collective behavior using a binary decision model. Individuals choose between two alternatives, with the final decision depending on how many others made which choice. He mentions rioting as an example, where individuals tend to join in more likely when there are already many other rioters, since the probability of being apprehended is smaller the larger the number of involved is.

He defines the threshold as the proportion of the group an individual would have to see join an action before it would do so. It is formed by some combination of costs and benefits which (at least rational) individuals are assumed always to estimate. The point where the perceived benefits of taking the decision exceed the perceived costs then equals the threshold. In terms of rioting a "radical" would have a low threshold, a "conservative" would have a higher threshold. He mentions that the threshold model may be used not only to model collective behavior in riots, but to describe diffusion of innovations, spread of rumors and diseases, striking, voting and many more.

Formally, Granovetter assumes that all individuals may have perception of each other⁵ and everyone has a threshold x , distributed by a function $f(x)$ with cumulative distribution function $F(x)$. The proportion of the population who have joined the riot at time t is denoted by $r(t)$. Thus, the process is described by a difference equation

$$r(t+1) = F(r(t)), \quad (3)$$

with all zero-threshold individuals initially infected as initial condition.

To find equilibrium states of the process one has to solve equation (3) and set $r(t+1) = r(t)$, or equivalently finding the fixed points of F .

Thereafter the case where f is a normal distribution is discussed, since normal distributions are characteristic of populations where no strong tendencies of any kind to distort a distribution of preferences away from its regular variation about some central tendency. The results, obtained due to the work of Bob Phillips, have a surprising property. Considering the equilibrium fixed point r_e of F as a function of the standard deviation s of the normal distribution, $r_e = r_e(s)$ increases gradually up to a small number, then suddenly jumps to a value near 1, see figure 4.

There is a critical value s_c for the standard deviation at which the equilibrium density of rioters jumps discontinuously from almost zero to almost one. That the density reaches neither zero nor one is due to finite size effects. Since thresholds are normally distributed but can not be negative, any individual who would get a negative threshold assigned gets a zero threshold instead. Threshold zero individuals count as initially rioting. With a too small standard deviation there are not enough initially rioting individuals and not enough low threshold individuals to join the riot. Thus, the critical standard deviation constitutes some sort of optimal distribution of thresholds for the system, given the shape (here: normal) of the distribution curve. If the standard deviation is too large, then the same pathology as before with negative threshold individuals happens. Any individual with threshold higher than 100 is considered to be a threshold 100 individual and therefore immune to rioting. Too many initially immune individuals are the reason for the curve in figure 4 to be monotonically decreasing.

⁵In physical terms, the underlying network is a complete graph.

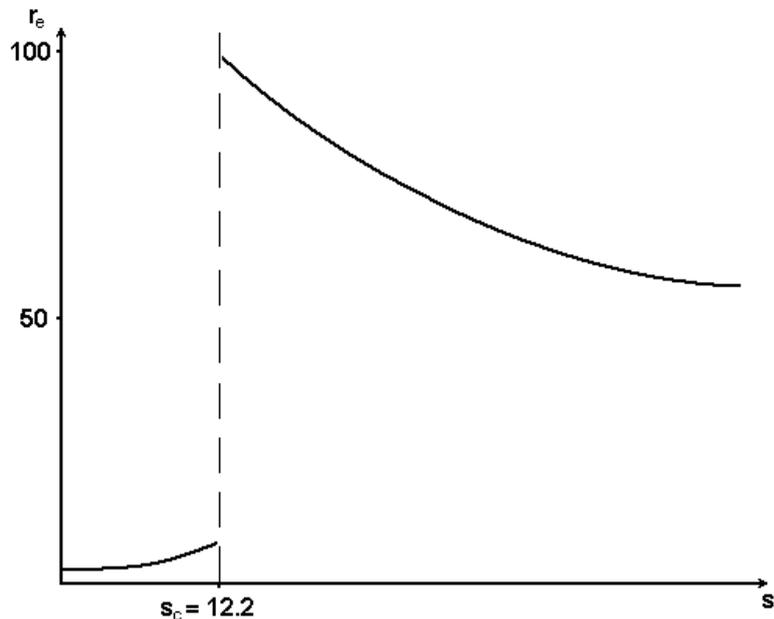


Figure 4: Equilibrium number r_e of rioters plotted against standard deviation s of normal distributions of thresholds with mean 25, $N = 100$. If the standard deviation is too small there will not be enough low threshold individuals needed to obtain a high number of rioters. If s is too wide there will be too much of high threshold individuals which most likely never riot.

He also discusses some issues of the threshold model access to explain collective behavior. One crucial assumption of the sociological setup is that individuals act rational and try to maximize the benefits from their decision. He points out, that with game theoretic arguments one could obtain a final state of the system, which maximizes the benefits of all actors but which is different from the equilibrium found using the threshold model. A more detailed discussion of this can be found in [3].

1.5.3 Complex Contagion

Unlike Schelling and Granovetter who discussed threshold models on rather simple network structures, Centola and Macy used developments about the structure of real world networks, i.e. the small world principle, [9]. The classic formulation of this principle comes from Watts and Strogatz, [40], where they demonstrate that the rate of propagation of simple contagion processes on a clustered network can be significantly increased by randomly rewiring a few ties within a cluster to establish bridges between clusters.

Simple contagion describes diffusion processes on networks where one single contact with a contagious node may be sufficient to get also contagious. Examples are diseases or the spread of information. In contrast to that, there is complex contagion where contact with multiple contagious sources is needed to

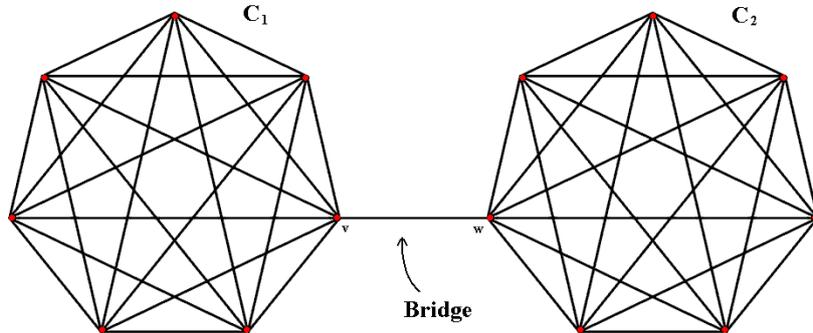


Figure 5: The bridge between two clusters constitutes a long tie which is (structurally) weak. None of the vertices in cluster C_2 but w can help v succeed the threshold. Thus, the spread of complex contagion from C_2 to C_1 is very unlikely to happen, in contrast to simple contagion.

become contagious. Each vertex in the network has a threshold which describes how many of its neighbors at least have to be contagious.

Centola and Macy identify at least four mechanisms that might explain the origin of complex contagion and thresholds: strategic complementarity, credibility, legitimacy, and emotional exchange. Strategic complementarity means that individuals take costs and benefits into account before they make their decision. This is exactly what Granovetter used to motivate his threshold model of collective behavior. Credibility is an important attribute of innovations, the higher the credibility of a new product the higher will be the chances of adopting it. Credibility rises as friends of individuals adopt. Spread of rumors or urban legends also relies on credibility, since hearing the same story from different people makes it seem more likely to be true. Closely related is legitimacy, individuals tend to adopt an innovation not before legitimacy is high enough, with legitimacy rising the more neighbors adopt. Emotional contagion describes the amplification of expressive and symbolic impulses in human behavior in spatially and socially concentrated gatherings.

In their work they were putting Granovetter's hypothesis of the *strength of weak ties* to the test in case of complex contagion. In [23], Granovetter demonstrated that the so called weak ties in a network accelerate the spread of disease, i.e. simple contagion. Weak ties are edges between vertices which had a large distance to each other if this edge would be removed. This is why they are called long ties as well. Considering complex contagion on small world networks, Centola and Macy found that long ties do not have the same effect on the spread of contagion, in some cases they may even be slowing down the process of dissemination. Figure 5 illustrates this effect. A bridge between two clusters connects vertices which do not have any neighbor in common. Say one cluster is contagious with every vertex infected, and the other one is not. Then there is only one link to the uninfected cluster which is contagious, which is insufficient to attain the threshold.

Also, Centola and Macy mention critical mass as a special feature of complex contagion processes but do not cover this topic.

1.5.4 Information Cascades in Complex Networks

In [39], Watts discusses information cascades on a sparse random network of interacting agents whose decisions are determined by the actions of their neighborhood according to a threshold rule. A cascade appears whenever a set of initially infected agents helps the so called *vulnerable* agents to attain their threshold who then activate other agents with higher threshold and so on. In Watts model there is given a degree distribution $\{p_k\}$ and fractional thresholds ϕ drawn from a given distribution $f(\phi)$, on a network of infinitely many agents. An Agent is called vulnerable if their threshold ϕ is smaller than the inverse of their degree k , i.e. $k \leq 1/\phi$, such that one infected neighbor suffices to attain the threshold. Hence, an agent with k neighbors is vulnerable with probability $\rho_k = \int_0^{1/k} f(\phi)d\phi$. With that, Watts derives the *cascade condition*

$$\sum_k k(k-1)\rho_k p_k = z, \quad (4)$$

where z denotes the average degree of the network. This condition is a criticality condition in the sense that it marks the transition between two regimes or phases. Whenever the left side of equation (4) is smaller than z all clusters of vulnerable agents are small in size and no cascades can occur. Whenever it is greater than z the typical size of clusters of vulnerable agents is infinite, so there is a positive probability that one of these percolates.

For sparse graphs Watts finds that the propagation of cascades is limited by the global connectivity, while in the case of a dense graph cascade propagation is limited by the high threshold agents. Further simulation results indicate that a heterogeneous threshold distribution increases the probability of global cascades.

1.5.5 Adoption of Innovation

An interesting application of threshold models with regard to knowledge diffusion is presented in [38], namely the diffusion of innovations. There, diffusion of innovation is the process by which members of a society adopt an innovation and thresholds describe the reluctance to adopt a new idea or try a new product. In contrast to former ideas as defined by Granovetter, Valente assigns a non-trivial topology to the society in which innovations disseminate. He defines the *social network* as the pattern of friendship, advice, communication or support which exists among the members of a social system.

Since behavior of individuals is based on how many others took a certain decision, one has to distinguish between global and local thresholds. Valente calls global thresholds *collective behavior thresholds* and mentions the problem of the inability of individuals to monitor the behavior of all the other individuals in the system to determine whether the threshold is succeeded or not. Thus, he suggests *adoption thresholds*, which only take neighbors into account. The proportion of adopters in the neighborhood is called *exposure* and the threshold is defined as the exposure at the time of adoption.

An important feature of this model is the categorization of adopters based on innovativeness as measured by time-of-adoption, following Rogers, [33]. Adopters are classified as early adopters, early majority, late majority, and laggards. *Early*

adopters are individuals whose time-of-adoption is greater than one standard deviation earlier than the average time-of-adoption. The *early and late majorities* are individuals whose time-of-adoption is bounded by one standard deviation earlier and later than the average. *Laggards* are those individuals who adopted later than one standard deviation from the mean.

In the same manner Valente classifies the personal network threshold adopters. *Very low threshold* individuals have a threshold one standard deviation lower than the average threshold. *Low and high threshold* individuals have thresholds bounded by one standard deviation less than and greater than the average. *Very high threshold* individuals have thresholds one standard deviation greater than the average.

With these classifications three datasets (physicians in Illinois, Brazilian farmers, Korean women) are analyzed which meet the requirements of providing data on time-of-adoption and social network ties. As defined above, thresholds of individuals are determined by the exposure at the time-of-adoption. An unsurprising but nevertheless significant finding then is that in all three datasets early adopters tend to have very low thresholds, while laggards are those individuals with very high thresholds.

2 Inhomogeneous Sparse Random Graphs

In [5], Bollobás, Janson and Riordan developed a formalism to treat inhomogeneous sparse random graphs in great generality. The class $G(n, p_{xy})$ of random graphs they discuss is defined by edge probabilities $p_{xy} = \min\{\kappa(x, y), 1\}$, where κ is some symmetric nonnegative integrable function on $S \times S$, and S is a separable metric space equipped with a probability measure μ , forming the so called *ground space* (S, μ) . They derive a wide variety of results which can be seen as a unification and generalization of various known results from random graph theory. The main result is a theorem about the existence, uniqueness, and size of the giant component in inhomogeneous sparse random graphs. Inhomogeneous means that the probability of the presence of edges is not the same for each edge, as for instance in Erdős-Renyi random graphs where each edge is present with the same probability p . Sparse means that the average number of edges scales linearly with the number of vertices.

2.1 Definitions and Properties

In order to define the edge probabilities well it has to be made clear what the set of vertices consists of.

Definition 2.1 *A vertex space \mathcal{V} is a triple $(S, \mu, \{\mathbf{x}_n\}_{n \geq 1})$, where (S, μ) is a ground space and, for each $n \geq 1$, \mathbf{x}_n is a random sequence (x_1, \dots, x_n) , such that for each μ -continuity set⁶ A*

$$\#\{i : x_i \in A\}/n \rightarrow \mu(A), \quad (5)$$

in probability.

Note that, writing δ_x for the measure consisting of a point mass of weight 1 at x , equation (5) can be written as $\nu_n(A) \xrightarrow{p} \mu(A)$, where $\nu_n := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ is the empirical distribution of \mathbf{x}_n .

Given a kernel κ on a vertex space $\mathcal{V} = (S, \mu, \{\mathbf{x}_n\}_{n \geq 1})$, define the random graph $G^\mathcal{V}(n, \kappa)$ as the random graph $G^\mathcal{V}(n, p_{ij})$, where

$$p_{ij} = \min\{\kappa(x_i, x_j)/n, 1\},$$

yielding a random graph on n vertices and edges $i \sim j$ present with probability p_{ij} . One may interpret the members x_j of the sequence \mathbf{x}_n as the strength of a certain property of the vertex j . Edges between vertices i and j are then formed according to the values of the kernel κ on x_i and x_j .

The degree distribution P_k for a random graph $G^\mathcal{V}(n, \kappa)$ converges to a superposition of Poisson distributions.

$$P_k \rightarrow \int \frac{\lambda(x)^k}{k!} e^{-\lambda(x)} d\mu(x), \quad (6)$$

where the convergence is in probability and $\lambda(x) = \int \kappa(x, y) d\mu(y)$. Note that, depending on the choices of vertex space and kernel, various distributions may arise, even power laws, as Bollobás, Janson, Riordan showed.

⁶A set $A \subset S$ is a μ -continuity set if A is measurable and its boundary is a measure zero set.

Further on, given a kernel κ on a ground space (S, μ) , an integral operator T_κ on (S, μ) is defined by

$$(T_\kappa f)(x) = \int_S \kappa(x, y) d\mu(y),$$

for any measurable function f such that the integral is defined for a.e. $x \in S$. Also, an operator norm is defined

$$\|T_\kappa\| = \sup\{\|T_\kappa f\|_2 : f \geq 0, \|f\|_2 \leq 1\}, \quad (7)$$

coinciding with the usual L^2 operator norm of T when being finite.

To study the component structure of $G(n, \kappa)$, they used the multi-type Galton-Watson branching process with type space S , where a vertex of type $x \in S$ gives birth to a set of children distributed as a Poisson process on S with intensity $\kappa(x, y) d\mu(y)$. Let $\rho(\kappa, x)$ be the probability that the branching process starting in a vertex of type x survives for infinity. Thus,

$$\rho(\kappa) := \int_S \rho(\kappa, x) d\mu(x)$$

is the overall survival probability of the branching process.

To state the main result of [5], the conclusion about the existence, size, and uniqueness of the giant component of the random graphs $G(n, \kappa)$, we need to introduce the notion of *graphical* kernels. These are designed to avoid that the graph is determined by the behavior of κ on a measure zero set.

Definition 2.2 *A kernel κ is graphical on a vertex space $\mathcal{V} = (S, \mu, \{\mathbf{x}_n\}_{n \geq 1})$ if the following conditions hold:*

- κ is continuous a.e. on $S \times S$;
- $\kappa \in L^1(S \times S, \mu \times \mu)$;

•

$$\frac{1}{n} \mathbb{E}e(G^\mathcal{V}(n, \kappa)) \rightarrow \frac{1}{2} \int \int_{S^2} \kappa(x, y) d\mu(x) d\mu(y) = \frac{\|\kappa\|_1}{2}.$$

Also it is convenient to allow the kernel to be dependent on n , to include models such as Erdős-Renyi random graphs $G(n, p)$ where the particular case $p = p(n) = c/n$ is of great importance. This leads to the following definition.

Definition 2.3 *Let κ be a graphical kernel on a vertex space $\mathcal{V} = (S, \mu, \{\mathbf{x}_n\}_{n \geq 1})$. A sequence $\{\kappa_n\}$ of kernels on (S, μ) is graphical on \mathcal{V} with limit κ if, for a.e. $(y, z) \in S^2$, $y_n \rightarrow y$ and $z_n \rightarrow z$ imply that $\kappa_n(y_n, z_n) \rightarrow \kappa(y, z)$ and*

$$\frac{1}{n} \mathbb{E}e(G^\mathcal{V}(n, \kappa_n)) \rightarrow \frac{1}{2} \int \int_{S^2} \kappa(x, y) d\mu(x) d\mu(y).$$

We are now ready to state the main result. Denote by $C_1(G^\mathcal{V}(n, \kappa_n))$ the largest component of $G^\mathcal{V}(n, \kappa_n)$, and by O_p and o_p the usual Landau symbols, with the subscript p indicating convergence in probability. Also, denote by $f = \Theta(g)$ if $f = O(g)$ and $g = O(f)$. Then the following holds:

Theorem 2.4 ([5]) *Let κ_n be a graphical sequence of kernels on a vertex space \mathcal{V} with limit κ .*

- *if $\|T_\kappa\| \leq 1$, then $C_1(G^\mathcal{V}(n, \kappa_n)) = o_p(n)$, while if $\|T_\kappa\| > 1$, then $C_1(G^\mathcal{V}(n, \kappa_n)) = \Theta(n)$ almost surely.*

- *For any $\epsilon > 0$,*

$$\frac{1}{n}C_1(G^\mathcal{V}(n, \kappa_n)) \leq \rho(\kappa) + \epsilon,$$

almost surely.

- *If κ is irreducible⁷, then*

$$\frac{1}{n}C_1(G^\mathcal{V}(n, \kappa_n)) \rightarrow \rho(\kappa),$$

in probability.

In all cases $\rho(\kappa) < \mu(S)$; furthermore, $\rho(\kappa) > 0$ if and only if $\|T_\kappa\| > 1$.

This theorem can be seen as a generalization of classical finite-type branching processes where the largest eigenvalue determines the criticality condition.

2.2 Additive and Multiplicative Coupling

Of particular importance in sociophysics are the two kernels κ_* and κ_+ , defined by $\kappa_*(x, y) = \psi(x)\psi(y)$, for some nonnegative integrable function ψ , and $\kappa_+(x, y) = \phi(x) + \phi(y)$, for some nonnegative integrable function ϕ . Most commonly used is $\psi(x) = \phi(x) = x$, for all x , with S being a finite interval in the real numbers and μ being any probability distribution on S , yielding that the probability that a vertex v_x with degree x and a vertex v_y with degree y form a link with probability proportional to the product xy , or the sum $x + y$, respectively. The interest in these two kernels arises not only from their algebraic simplicity, but from their interpretation in terms of social systems. To see this, we outline in short how networks with kernels $\kappa_{*,+}$ can be grown from a simple algorithm which admits such an interpretation.

So, chose n vertices with given degree x_1, \dots, x_n . To construct the edge set for either case, we assume that the degree of each vertex v_x is the sum of an in-degree and an out-degree, similar to the situation in directed graphs. Since vertices v_x and v_y can form an edge only if one of the half-edges contributes to the in-degree and the other half-edge to the out-degree, the probability $P(v_x \sim v_y) = p_{xy}$ of finding randomly drawn vertices v_x and v_x linked is given by $P(v_x \sim v_y) = P(x_{out})P(y_{in}) + P(x_{in})P(y_{out})$, where the subscripts *out/in* represent out- and in-degree of the vertices.

To obtain multiplicative coupling, for each vertex with degree x choose $x_{out} = x_{in} = x/2$. The fractional ansatz with factor 1/2 ensures that the graph constructed in this way will not end up with 'open' half-edges, at least asymptotically in the case of $n \rightarrow \infty$. It follows that

⁷A kernel κ is irreducible if $A \subset S$ and $\kappa = 0$ a.e. on $A \times (A \setminus S)$ implies $\mu(A) = 0$ or $\mu(A \setminus S) = 0$.

$$\begin{aligned}
P(v_x \sim v_y) &\propto \frac{x}{2} \frac{y}{2} + \frac{x}{2} \frac{y}{2} \\
&\propto xy.
\end{aligned}$$

To obtain additive coupling assume that the out-degree of each vertex is bounded by a constant d . This yields

$$\begin{aligned}
P(v_x \sim v_y) &\propto d(y-d) + (x-d)d \\
&= d(x+y) - 2d^2 \\
&\propto x+y.
\end{aligned}$$

Hence, individuals in a network with multiplicative coupling tend to create new links on their own, since half of their degree are outgoing links which can be interpreted as requests for new neighbors. Such behavior can be observed, for instance, in economies of scarcity like the former GDR, or Cuba, where individuals need private connections to professionals to carry out specific problems, e.g. mechanics, tradespeople or people with access to a certain rare supply.

On the other hand, individuals in networks with additive coupling have a bound for outgoing links, whenever they reached a certain number of neighbors they are not requesting any new ones, though still are accepting requests.

Also, mixtures are possible, there are interpolating kernels, for instance

$$\begin{aligned}
\kappa_\alpha(x, y) &= \alpha xy + (1 - \alpha)(x + y), \text{ or} \\
\kappa_\beta(x, y) &= (xy)^\beta (x + y)^{1-\beta}.
\end{aligned}$$

For both holds

$$\begin{aligned}
\kappa_0 &= \kappa_+, \\
\kappa_1 &= \kappa_*.
\end{aligned}$$

In what follows we will see that it is possible to choose κ_* and κ_+ such that the resulting asymptotic degree distributions of $G(n, \kappa_*)$ and $G(n, \kappa_+)$ are the same. Nevertheless, any two (random) graphs with the same degree distribution may be topologically very different. We focus on the emergence of the giant component, showing that the phase transition in additively coupled networks always happens 'later' than in multiplicatively coupled systems.

2.3 The Phase Transition of Additive and Multiplicative Coupling

Recall, that the norm of the operator $\|T_\kappa\|$ is defined as in (7). For differentiable kernels, this is equivalent to the constrained variational problem

$$\int (G(x, f(x)) - \lambda N(x, f(x))) dx = \text{extr!} \tag{8}$$

with $G(x, f(x)) = (T_\kappa f)^2(x)$ and constraint $N(x, f(x)) = f^2(x) - 1$. Since $G - \lambda N$ does not depend on any derivatives of f , the Euler-Lagrange equations of variational calculus yield $G(x, f(x)) - \lambda N(x, f(x)) = \text{const}$. A simple calculation

shows, that the constant on the right hand side can always be chosen to be zero. Therefore we get

$$(T_\kappa f)^2(x) - \lambda f^2(x) = 0$$

what is equivalent to

$$f(x) = \frac{1}{\sqrt{\lambda}}(T_\kappa f)(x). \quad (9)$$

Now Let g be the function that attains the maximum, i.e. a solution of (8). Since

$$0 = \int (T_\kappa g)^2(x) dx - \int \lambda g^2(x) dx = \|T_\kappa\|^2 - \lambda,$$

we calculate that

$$\|T_\kappa\| = \sqrt{\lambda}.$$

Thus, it is possible to determine the operator norm of T_κ by computing the Lagrangian multiplier λ . This way of calculating may not be the most general way, but yields a convenient way to calculate the phase transition of multiplicative and additive kernels.

Let $\kappa_*(x, y) = \psi(x)\psi(y)$ for some nonnegative function $\psi \in L^1(\mathcal{S})$. We use equation (9) to compute $\|T_{\kappa_*}\|$ for the multiplicative kernel κ_* . Defining $I(f) = \int f(t)\psi(t)d\mu(t)$ gives $T_{\kappa_*}f(x) = I(f)\psi(x)$ and inserting (9) in $I(f)$ gives

$$I(f) = \frac{I(f)}{\sqrt{\lambda}} \|\psi\|_2^2,$$

from where we calculate $\|T_{\kappa_*}\|$ to be

$$\|T_{\kappa_*}\| = \sqrt{\lambda} = \|\psi\|_2^2.$$

Now let $\kappa_+(x, y) = \phi(x) + \phi(y)$ for some nonnegative function $\phi \in L^1(\mathcal{S})$. Define $I_1(f) = \int f(t)d\mu(t)$ and $I_2(f) = \int \phi(t)f(t)d\mu(t)$. Then $T_{\kappa_+}f(x) = I_1(f)\phi(x) + I_2(f)$, inserting (9) in $I_1(f)$ and $I_2(f)$ yields the following linear system of equations.

$$\begin{aligned} \sqrt{\lambda}I_1(f) &= \|\phi\|_1 I_1 + I_2 \\ \sqrt{\lambda}I_2(f) &= \|\phi\|_2^2 I_1 + \|\phi\|_1 I_2, \end{aligned}$$

with solution

$$\|T_{\kappa_+}\| = \sqrt{\lambda} = \|\phi\|_2 + \|\phi\|_1.$$

To compare multiplicative and additive kernels and their phase transitions, we want to choose the functions ψ and ϕ , such that κ_* and κ_+ yield the same degree distribution. The next lemma shows, that this is always possible.

Lemma 2.5 Let $\kappa_*(x, y) = \psi(x)\psi(y)$ be a multiplicative kernel and $\kappa_+(x, y) = \phi(x) + \phi(y)$ an additive kernel. If

$$\phi(x) = \|\psi\|_1\psi(x) - \frac{1}{2}\|\psi\|_1^2, \quad (10)$$

then the random graphs $G(n, \kappa_*)$ and $G(n, \kappa_+)$ have the same degree distribution.

Proof. From (6), we see, that we need to calculate $\lambda_*(x) = \int \kappa_*(x, y)d\mu(y)$, which is $\lambda_*(x) = \|\psi\|_1\psi(x)$. Choosing $\phi(x)$ as in (10) yields

$$\begin{aligned} \lambda_+(x) &= \int \kappa_+(x, y)d\mu(y) \\ &= \int (\|\psi\|_1(\psi(x) + \psi(y)) - \|\psi\|_1^2)d\mu(y) \\ &= \|\psi\|_1\psi(x) + \|\psi\|_1 \int \psi(y)d\mu(y) - \|\psi\|_1^2 \\ &= \lambda_*(x) \end{aligned}$$

Since the degree distribution is determined by a superposition of Poisson distributions with means $\lambda(x)$, the resulting asymptotic degree distributions must be the same.

□

Notice, that due to the desired positivity of ϕ equation (10) imposes the condition $\psi(x) \geq \|\psi\|_1/2$, almost surely, on ψ .

To compare the phase transitions of $G(n, \kappa_*)$ and $G(n, \kappa_+)$, we need to compute $\|T_{\kappa_*}\|$ and $\|T_{\kappa_+}\|$. We have shown that $\|T_{\kappa_*}\| = \|\psi\|_2^2$ and $\|T_{\kappa_+}\| = \|\phi\|_1 + \|\phi\|_2$. Choosing ϕ according to equation (10) yields

$$\begin{aligned} \|\phi\|_1 &= \frac{1}{2}\|\psi\|_1^2, \\ \|\phi\|_2 &= \|\psi\|_1\sqrt{\|\psi\|_2^2 - \frac{3}{4}\|\psi\|_1^2}, \end{aligned}$$

and the norm $\|T_{\kappa_+}\|$ is given by

$$\|T_{\kappa_+}\| = \|\psi\|_1\sqrt{\|\psi\|_2^2 - \frac{3}{4}\|\psi\|_1^2} + \frac{1}{2}\|\psi\|_1^2.$$

Figure 6 shows the behavior of $\|T_{\kappa_+}\|$, depending on $\|\psi\|_1$, where $\|\psi\|_2$ is set to the critical value 1. The plotted function is defined by $y = x\sqrt{1 - (3/4)x^2} + 1/2x^2$, where $x = \|\psi\|_1$ and $y = \|T_{\kappa_+}\|$. It can be seen, that y is increasing from $y = 0$ to a maximum value $y_{\max} = 1$. All values of $\|T_{\kappa_+}\|$ are below the line $y = 1$ and belong to a phase transition of $G(n, \kappa_+)$ which is smaller than the phase transition of $G(n, \kappa_*)$, thus, describing the subcritical case of $G(n, \kappa_+)$.

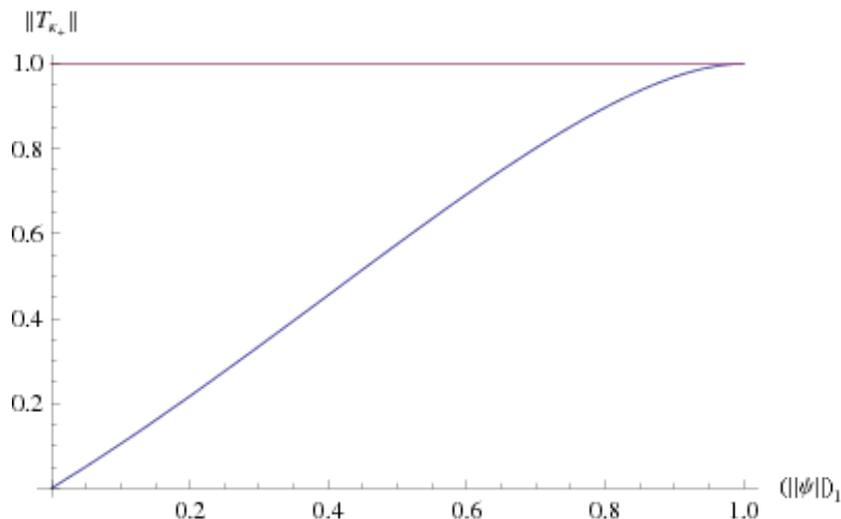


Figure 6: The behavior of $\|T_{\kappa_*}\|$, when $\|T_{\kappa_*}\|$ is set to the critical value of the phase transition $\|T_{\kappa_*}\| = 1$ (red line). $\|T_{\kappa_*}\| = x\sqrt{1 - (3/4)x^2} + 1/2x^2$, depending on $x = \|\psi\|_1$ (blue line)

3 The Communication Index of Graphs

3.1 Basics

A simple way of characterizing structural properties of a graph or a network is to use only its degree distribution, the ordered sequence of relative frequencies of degrees. One could use any kind of entropy, information or complexity measure to describe and distinguish certain families of graphs only by their degree distribution. A survey about graph entropy can be found in [37], complexity of graphs in [32].

Unsurprisingly, these measures turn out to be not sensitive enough to distinguish all kinds of graphs. In any case, there exist graphs, which we would consider to be very similar, for which these distance measures yield large values.

A promising approach is done by Dehmer in [13] and Dehmer and Emmert-Streib in [14], where probability distributions corresponding to networks are constructed in a non trivial manner. These distributions allow the investigation of the entropy functional and therefore an analytical treatment of notions like information or similarity for graphs.

While their focus is on finding probability measures on graphs, our aim is to associate weights with the edges, which reflect structural properties of the graph. It turns out, that the edge weights constructed in this paper result in a uniform distribution for both, regular graphs and star graphs, thus maximizing the entropy in the set of connected graphs.

Besides such entropic considerations, we use these edge weights to derive a measure for the effectiveness of communication on complex networks. The first chapter is devoted to the introduction of the communication index for graphs and its relevance as an effectiveness measure for communication. We then turn our view to an application of the communication index in the field of

peer-to-peer networks. With the help of the communication index an optimal network structure can be obtained for networks with a given distribution of bandwidths of the peers. Further on, we calculate the distribution of edge weights for the Galton-Watson tree and its communication index. This result translates to Erdős-Renyi random graphs. Finally we made some simulations to calculate the communication index, its edge weight distribution and its node weight distribution, on Galton-Watson trees and Erdős-Renyi random graphs to illustrate the behavior the quantity on simple random graphs.

3.1.1 Definitions and Simple Properties

We start by introducing the concept of the communication index of a graph. To do so, we set $B_v = \{e \in E : e = (v, w), w \in V\}$ for each vertex $v \in V$. Thus, B_v is the set of outgoing links from v .

Definition 3.1 Let $G = (V, E)$ be a graph on n vertices, i.e. $\#V = n$. For each edge $e \in E$, define

$$\gamma(e) = \min\left\{\frac{1}{\deg(v)} : v \in e\right\} \quad (11)$$

We call $\gamma(e)$ the communication weight of the edge e . For each vertex $v \in V$, define the communication strength γ_v by

$$\gamma_v = \sum_{e \in B_v} \gamma(e).$$

The number

$$\gamma(G) = \frac{1}{n} \sum_{v \in V} \gamma_v$$

is called communication index of G .

Figure 7 shows an example of a graph equipped with communication weights. Each edge is equipped with its communication weight and each node is equipped with its communication strength (numbers in circles), which is the sum of all communication weights on the edges adjacent to the node.

The intuitive meaning of the communication weights becomes clear as follows: Suppose any vertex $v \in V$ communicates in some way with its neighbors and none of them is being favored. So, v spends an equal amount of time t for communication with each of them, namely $t_v = 1/\deg(v)$. Since any neighbor w of v has neighbors to communicate itself, two nodes can only communicate an amount of time, which equals $\min\{t_v, t_w\} = \gamma((v, w))$. The communication index $\gamma(G)$ may then be seen as a measure of efficiency of the communication on G .

Since the degree of every communicating node is always greater or equal 1, $\deg(v) \geq 1$ for all $v \in V$, one may as well interpret the $\gamma(e)$ as the probability that a communication between the vertices v and w , which constitute the link $e = (v, w)$, will be realized⁸. Then $\gamma(G)$ is a measure for the ability of the graph G to originate communication.

⁸Here we assume, that vertices with zero degree do not communicate at all.

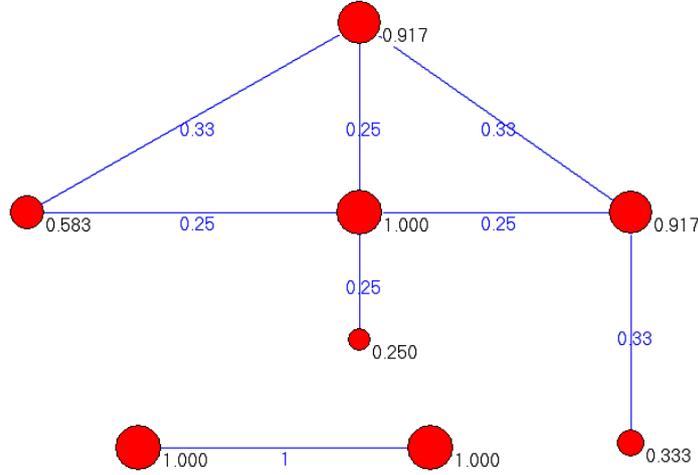


Figure 7: Visualization of the communication index scheme. This network with two components has $N=8$ nodes and $M=8$ edges, so an average degree of 2. The communication weight (blue numbers) on each edge is calculated by $\min(\frac{1}{\deg(x)}, \frac{1}{\deg(y)})$ of the end-degrees $\deg(x)$ and $\deg(y)$. Here the average communication edge-weight is 0.375. Then all communication edges adjacent to a node are summed up, for each node ("communication strength" - black numbers, and proportional sizes of the nodes). The node-average communication-node-sum is **0.75**, which we call **communication index** of this network.

We now prove some elementary properties of the communication index. Notice that there is an additivity property for γ , namely if G consists of k components G_k , then $n\gamma(G) = \sum_j \gamma(G_j)$. Hence, there is no loss in generality if we assume graphs to be connected.

Lemma 3.2 *Let G be a connected graph on $n > 1$ vertices. Then*

- (i) $1/(n-1) \leq \gamma_v \leq 1$ for all $v \in V$.
- (ii) Vertices v with maximal degree fulfill $\gamma_v = 1$.
- (iii) $\gamma(G) \geq \frac{2}{n}$, with equality iff G is a star graph.
- (iv) $\gamma(G) \leq 1$, with equality iff G is a regular graph with positive degree.

Proof.

- (i) The lower bound is trivial, since $\gamma(e) \geq 1/(n-1)$ for all $e \in E$. The upper bound is a consequence of the standard estimate

$$\gamma_v = \sum_{e \in B_v} \gamma(e) \leq \deg(v) \max_{e \in B_v} \gamma(e) = \frac{\deg(v)}{\deg(w)} \leq 1,$$

since $\max_e \gamma(e) = 1/\deg(w)$ for some $w \in V$ implies $\deg(v) \leq \deg(w)$.

- (ii) Consider a vertex $v \in V$ with $\deg(v) \geq \deg(w)$ for all $w \in V$. Then $\gamma(e) = 1/\deg(v)$ for all $e \in B_v$ and

$$\gamma_v = \sum_{e \in B_v} \frac{1}{\deg(v)} = \deg(v) \frac{1}{\deg(v)} = 1.$$

- (iii) This follows directly from the definition of $\gamma(G)$, from the lower bound for the γ_v and from the fact that there is at least one vertex v' with $\gamma_i = 1$:

$$\begin{aligned} n\gamma(G) &= \sum_{v \in V} \gamma_v = 1 + \sum_{v \neq v'} \gamma_v \\ &\geq 1 + \frac{1}{n-1}(n-1), \end{aligned}$$

where equality in the last line holds if and only if the value of γ_v of each of the remaining $(n-1)$ vertices equals $1/(n-1)$. But this is possible only if those vertices are connected to the vertex v' of maximal degree and not connected to any other vertex. Thus, equality holds if and only if G is a star.

- (iv) This follows directly from the definition of $\gamma(G)$ and from the upper bound for γ_v . To avoid the case of the empty graph, we assume $\deg(v) > 0$ for all $v \in V$.

$$\gamma(G) = \frac{1}{n} \sum_{v \in V} \gamma_v \leq \frac{1}{n} \sum_{v \in V} 1 = 1,$$

with, since G is connected, equality if $\gamma_v = 1$ for all $v \in V$. This is the case if and only if all vertices have maximal degree, therefore G must be a regular graph.

□

This lemma justifies the description of γ as a measure for the effectiveness of a communication process on a graph. Intuitively, one would classify the star graph not as an adequate network for communication. Identifying the vertices with people and edges with some relationship between them (for instance they may know each other by name), the lemma can be illustrated as follows. The central person of the star knows all other people and will divide its time up in equal parts. Thus, they will have not much time to communicate to a single person. The only person with a saturated communication is the central one.

The other extremal case is the one, where every node has the same number of neighbors. Anyone on the network has the same amount of time for each neighbor, therefore none of them will end up with dispensable time.

3.1.2 The CI and Computational Complexity

Let G be a star graph, G' be a regular graph, denote the communication edge weights of G with $\gamma_G(e)$, of G' with $\gamma_{G'}(e)$. One key observation is that in both cases the communication edge weights are uniformly distributed and the distributions $P(\gamma_G(e) = 1/k)$ and $P(\gamma_{G'}(e) = 1/k)$ yield zero entropy. In that

sense, these graphs can be considered as maximally non-random, in agreement with intuition. Note that, in contrast, entropy of the degree distribution yields minimum entropy only for regular graphs, not star graphs, although their computational complexity is the same, namely $O(\log n)$.

3.1.3 Inhomogeneous Sparse Random Graphs revisited

To analyze the Communication Index of such graphs, we adapt the definition of the communication weights in the following way. For each edge (v, w) that is present in a realization of a $G(n, p_{xy})$, we set

$$\gamma(v, w) = \min\left\{\frac{1}{\mathbb{E} \deg v}, \frac{1}{\mathbb{E} \deg w}\right\},$$

where \mathbb{E} denotes the (ensemble) mean. In terms of the B-J-R formalism, given a kernel κ on a ground space (S, μ) and a vertex v of type x , this mean is given by $\mathbb{E} \deg v = \lambda(x) = \int_S \kappa(x, y) d\mu(y)$.

With $\lambda(x) \geq 1$ for all x it is reasonable to interpret $\gamma(x, y)$ as the probability that communication between the types x and y comes about. This gives rise to a new random graph $G(n, \tilde{p}_{ij})$, with $\tilde{p}_{xy} = \min\{\kappa(x, y)\gamma(x, y), 1\}$. For instance, if we choose a multiplicative coupling of the vertices, i.e. $\kappa(x, y) = c\psi(x)\psi(y)$, for some integrable function ψ and $c > 0$, then $\lambda(x) = c\|\psi\|_1\psi(x)$. Thus, the communication kernel $\tilde{\kappa}(x, y) = \kappa(x, y)\gamma(x, y)$ is given by

$$\begin{aligned} \tilde{\kappa}(x, y) &= c\psi(x)\psi(y) \min\left\{\frac{1}{c\|\psi\|_1\psi(x)}, \frac{1}{c\|\psi\|_1\psi(y)}\right\} \\ &= \frac{c\psi(x)\psi(y)}{c\|\psi\|_1 \max\{\psi(x), \psi(y)\}} \\ &= \frac{1}{\|\psi\|_1} \min\{\psi(x), \psi(y)\}. \end{aligned}$$

Note that the tunable constant c cancels out and the structural behavior of the communication graph $G(n, \tilde{p}_{xy})$ depends only on the choice of ψ and of the ground space. Also, for $S = (0, 1]$ equipped with the Lebesgue measure, and $\psi(x) = x$, this is Dubin's model, see [5, 16.1.] for more details. It was shown in [] (see also [] and []) that in this case, the phase transition 'is of infinite order', i.e. has infinite exponent.

3.2 The Communication Index and Peer-to-Peer Networks

There has been an increasing interest in the field of peer-to-peer computer networks recently. In a peer-to-peer computer network, every node has the same capabilities and responsibilities, there are no distinguished nodes. This is in contrast to client/server architectures, where the roles of the client and server nodes differ.

Our focus is on peer-to-peer networks of unstructured type, such as Bit Torrent, meaning there is no specific network topology to be formed by the participating peers. Peer-to-peer networks such as Bit Torrent are increasingly used for disseminating potentially large files from a server to many end users via the Internet. The key idea is to divide the file into many equally-sized

parts and then let users download each part from another user who has already downloaded it.

A traditional Bit Torrent system has elements called Trackers whose main purpose is to enable peers to find each other. The Bit Torrent Tracker randomly assigns a new (entering) user a set of peers that are already in the system to communicate with.

3.2.1 The CI as an a-priori Measure for Effectiveness

The CI introduced and discussed in section 3.1.1 uses the degree of vertices in a network to measure time utilization and effectiveness of communication. In peer-to-peer networks, vertices are communicating with each other, but constrained by upload and download capacities. If we modify the definition of the communication to take into account that there are natural limits on how much data can flow along a connection between to links in the network, we can view the CI as an a-priori measure for effectiveness in peer-to-peer networks.

Assume that a vertex v with capacity $C_v > 0$ enters the network with $d_v > 0$ links and distributes its capacity among its neighbors equally, i.e. assigning an amount of C_v/d_v of its capacity to each neighbor to use it for exchange of pieces of a file. Then v acquires $\gamma_t(v)$ pieces per time step t (or at least a number proportional to that quantity), with

$$\gamma_t(v) = \sum_{w:v \sim_t w} \min\left\{\frac{C_v}{d_v}, \frac{C_w}{d_w}\right\},$$

where the sum ranges over all neighbors of v at time t (denoted by $v \sim_t w$ here).

Neglecting the last piece problem⁹, v is acquiring pieces until it has L pieces, formally

$$\sum_{t=1}^{T(v)} \gamma_t(v) = L,$$

with $T(v)$ being the download latency. Assuming that v acquires the same amount of pieces at each time step, i.e. $\gamma_t(v) = \gamma(v)$, for the download latency of a peer v we get

$$L = \sum_{t=1}^{T(v)} \gamma(v) = T(v)\gamma(v) \Rightarrow T(v) = \frac{L}{\gamma(v)},$$

showing the inverse proportionality of the CI $\gamma(v)$ of a peer and its download latency $T(v)$.

Even without the strong assumption that a peer acquires the same amount of pieces all over the time, one can obtain the CI as an upper bound via the standard estimation $\sum_{t=1}^{T(v)} \gamma(v) \geq \min_t\{\gamma_t(v)\}T(v)$, since $\gamma_t(v)$ cannot be zero for non-zero capacities, showing that peers want to maximize the smallest of their values γ_t .

⁹Using common unchoking algorithms the last piece(s) of a file may be harder to get than other pieces, elongating the download process. This is the so called last piece problem.

3.2.2 Paper 1: Optimizing Topology in Bit Torrent Based Networks

In [10] we have shown how the communication index can help optimizing topology in bit torrent networks. The static network of a Bit Torrent system is modeled as a connected regular graph $G(V, E)$, where the set of vertices V represents the peers in the network and the set edges E the connections established among the peers.

Assume that a vertex can be one of n types $0 < x_1 < \dots < x_n$ (which we want to interpret as bandwidths), which are distributed with density ρ_i , i.e. there is a fraction of ρ_i vertices of each type i in V . The fraction q_{ij} of edges of type (x_i, x_j) fulfills the constraint

$$\rho_i = q_{ii} + \frac{1}{2} \sum_{k=1}^{i-1} q_{ik} + \frac{1}{2} \sum_{k=i+1}^n q_{ik}, \quad (12)$$

just stating that the probability of finding a peer with bandwidth x_i equals the probability of finding one attached to an edge of type (x_i, x_i) plus half of the probability that it is attached to an edge of type (x_i, x_j) , $j \neq i$. This is true for any regular graph, since the total number of edges scales linearly with the total number of vertices.

Whenever two peers, one with bandwidth x_i , the other one with bandwidth x_j , want to exchange files, the quantity $\min\{x_i, x_j\}$ is decisive for the speed of exchange, since the peer with the lower bandwidth is not able to upload or download faster than its bandwidth allows. To adopt the communication index here, we need to customize the concept a bit. Due to optimistic unchoking in Bit Torrent networks, it may happen that with a certain probability p_{ij} the larger of both bandwidths gets realized, see [10] for details on this.

Thus, we define the communication edge weights to be random variables γ with the following distributions

$$\begin{aligned} P(\gamma(x_i, x_i) = x_i) &= 1 \\ P(\gamma(x_i, x_j) = \max\{x_i, x_j\}) &= p_{ij} \\ P(\gamma(x_i, x_j) = \min\{x_i, x_j\}) &= 1 - p_{ij}. \end{aligned}$$

Then, the mean edge weight is given by

$$E(\gamma) = \sum_{i=1}^n x_i q_{ii} + \sum_{i,j:i < j} (\max\{x_i, x_j\} p_{ij} q_{ij} + \min\{x_i, x_j\} (1 - p_{ij}) q_{ij})$$

Eliminating q_{ii} using equation (12) and using that $\Delta_{ji} := x_j - x_i > 0$ for $i < j$ yields

$$E(\gamma) = \bar{x} + \sum_{i,j:i < j} \Delta_{ji} (p_{ij} - \frac{1}{2}) q_{ij}, \quad (13)$$

with \bar{x} being the mean value of the types.

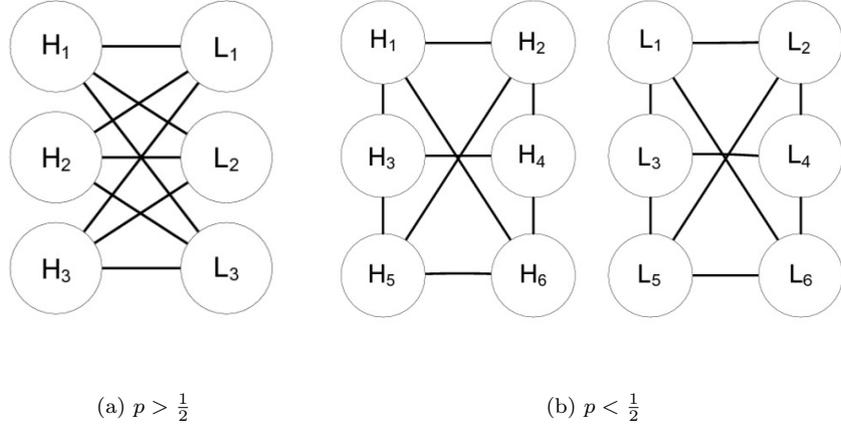


Figure 8: The connectivities of high and low bandwidth nodes, that maximizes the edge weight in a regular graphs, for two cases $p > \frac{1}{2}$ (figure 8(a)) and $p < \frac{1}{2}$ (figure 8(b)). For $p > \frac{1}{2}$, the graph forms a bipartite graph, where no peers of high or low bandwidth are directly connected to each other, on the other hand when $p < \frac{1}{2}$, the high and low bandwidth peers gets isolated from each other to maximize the edge weight.

Now assume $n = 2$, i.e. consider two types $x_1 =: x_L < x_H := x_2$ of different bandwidth, say high (H) and low (L). Then the mean value E of γ becomes a one dimensional linear function, with the fraction x of mixed edges (x_H, x_L) as variable,

$$E(x) = \text{const.} + (x_H - x_L)(p - 1/2)x.$$

Since E is a linear function, the maximum value gets attained on the boundary of the domain of E , which is defined by the constraints $\rho_H = x_H + \frac{x}{2}, \rho_L = x_L + \frac{x}{2}$. Thus, there are two cases we need to investigate, namely $p < 1/2$ and $p > 1/2$. In case of $p = 1/2$ any topology is equivalently effective.

If $p < 1/2$, the slope of E is negative, therefore we want to minimize the value of x . This already indicates the optimal strategy in distributing the bandwidths among peers in the network: Avoid edges with different types of peers attached, cluster peers of the same bandwidth together.

Nevertheless, in any connected graph the solution $x = 0$ (and therefore $x_H = \rho_H, \rho_L = x_L$) is not possible. Since one wants to maximize the number of bonds of the same type and minimize mixed bonds, one needs to find a partition (V_1, V_2) of the vertex set V , with $\rho_L = |V_1|/|V|, \rho_H = |V_2|/|V|$, and such that the number of edges from V_1 to V_2 is minimal. This is exactly the problem of finding a minimum cut in a graph, given the size of the partition.

If $p > 1/2$, the slope of E is positive, therefore we want to choose x maximal, i.e. avoiding edges with peers of different type. Note that due to the constraints, the solution $x = 1$ is only possible if $\rho_H = \rho_L = 1/2$. Similar to the problem with $p < 1/2$ one needs to find a cut in the graph, but here a maximum cut. Figure 8 illustrates the two cases.

Optimizing Topology in Bit Torrent Based Networks

J. Chandra, S. Delitzscher, N. Ganguly, A. Jhunjhunwala, T. Krueger, N. Sharma,

Computer Communications Workshops (INFOCOM WKSHPs), 2011 IEEE Conference on, Issue Date: 10-15 April 2011, 888 - 893

Abstract: In this paper, we discuss the importance of the network connectivities of the peers in Bit Torrent based systems in determining the download performance of the peers. In this context, assuming that the fraction of the peers of each bandwidth are known, we derive optimal connectivities of the peers that help to improve the average latency of the peers. We represent the topology of a Bit Torrent based system as a weighted graph, where the average edge weight of the graph directly relates to the download latency of the peers. We formulate the average edge weight of the whole system as a linear function of the fraction of the edges that connect peers of different bandwidth and derive the topology that maximizes the average edge weight of the network. Simulation results based on the Bit Torrent protocol validates the fact that in the optimal topology, peers have 13% better download latency as compared to topologies formed in the normal Bit Torrent based systems. Further the obtained topology also improves the fairness of the system as compared to normal Bit Torrent significantly.

Find Paper 1 attached to the framework text.

3.3 The Communication Index of Random Graphs

When considering random graphs, the communication index becomes a random variable. One natural way to investigate the behavior of the communication index of random graphs is to consider random trees as realizations of certain branching processes first, and then generalizing it to the class of Erdős-Renyi random graphs.

To obtain analytical results, we calculate the distribution of the communication index edge weights for the supercritical Galton-Watson branching process, conditioned on not dying out. We go on calculating the

3.3.1 The Galton-Watson Process

Let ξ_i^t , $i, t \geq 0$, be i.i.d. nonnegative integer-valued random variables. Define a sequence Z_t , $t \geq 0$ by $Z_0 = 1$ and

$$Z_{t+1} = \begin{cases} \xi_1^{t+1} + \dots + \xi_{Z_t}^{t+1} & \text{if } Z_t > 0 \\ 0 & \text{if } Z_t = 0 \end{cases}$$

The sequence $\{Z_t\}$ is called a *Galton-Watson process*. The idea behind the definition is that Z_t is the number of people in the t -th generation of a branching process, and each member of the t th generation gives birth independently to an identically distributed number of children. The family $\psi_k = P(\xi_i^t = k)$ is called the *offspring distribution*. It is a well known fact, that the process dies out

with probability 1 if $\mu = \mathbb{E}\xi_i^t < 1$, in this case the process is called *subcritical*. There is a positive survival probability if $\mu > 1$, in this case the process is called *supercritical*. For more details, see for example Chapter 4 of [16].

As usual, we associate for each $t = 0, 1, 2, \dots$, a random graph G_t with the process, in the following way. The graph G_0 is the graph consisting only of the root vertex associated with $Z_0 = 1$. The graph G_1 consists of the root and $\xi_1^0 = Z_1$ vertices adjacent to the root. The graph $G_t = (V_t, E_t)$ is the tree consisting of the vertex set V_t , which fulfills $\#V_t = n(t) = \sum_{\tau=0}^t Z_\tau$, and of the edge set E_t , where $e = (v, w) \in E_t$ whenever v is a child of w or w is a child of v . Every edge $e \in E_t$ of G_t may be equipped with the communication weights $\gamma(e)$ as the minimum value of the inverse degrees of two adjacent vertices belonging to e .

Note that vertices of the last generation all have degree 1. So, to avoid a bias in the calculations we assign to each vertex on the t -th generation virtually its degree at time $t + 1$, but do not count any of these virtual edges. The communication index of G_t is then given by

$$\gamma(G_t) = \frac{2}{n(t)} \sum_{e \in E_t} \gamma(e).$$

We are now interested in the distribution $P(\gamma(e) = 1/k)$ of the communication weights $\gamma(e)$. If we want to relate the offspring distribution ψ_k of the Galton-Watson process to the degree distribution of the graphs G_t , we need to ensure that the process can not die out. To do so, we treat the supercritical case first, where $\mu > 1$. Nevertheless, there is a positive probability of extinction q , thus we need to condition the probabilities on $Z_n > 0$. In that case, $p_k = P(\deg(v) = k | Z_n > 0) \rightarrow \psi_{k-1}$ as n goes to infinity. The probability of extinction q is strictly less than 1 and positive if and only if $\psi_0 > 0$. Thus, we impose $\psi_0 = 0$ and derive a result on the distribution of the communication weights and may then use a transformation of Harris, see [25], to reduce the case $q > 0$ to $q = 0$.

Furthermore, notice that in this context choosing randomly an edge $e = (v, w)$, connecting a vertex w with its unique parent v , is equivalent to choosing randomly one vertex w and then considering the link between w and its parent v . The probability that this link is attached to a vertex v of degree k is then, as t goes to infinity, given by $\mu^{-1}(k-1)\psi_{k-1}$. Thus, the asymptotic joint probability $p_{kk'}$ that an edge is attached to a vertex v of degree k' and a parental vertex w of degree k is given by

$$p_{kk'} = \frac{(k-1)p_k p_{k'}}{\mu}.$$

From this joint probability we can find the distribution of the communication weights.

Lemma 3.3 *Let $\{Z_t\}$ be a supercritical Galton-Watson process with offspring distribution ψ_k and $\psi_0 = 0$. Set $P(\deg(v) = k) = p_k$, then, as t goes to infinity,*

$$P\left(\gamma(e) = \frac{1}{k}\right) = \frac{p_k}{\mu} \left((k-1) \sum_{j=1}^k p_j + \sum_{j=1}^{k-1} (j-1)p_j \right).$$

Proof. The probability, that a communication weight $\gamma(e)$ of an edge e takes the value $1/k$ for some k can be evaluated by using elementary probability calculus.

$$\begin{aligned}
P\left(\gamma(e) = \frac{1}{k}\right) &= \sum_{j < k} p_{kj} + \sum_{j < k} p_{jk} + p_{kk} \\
&= \sum_{j=1}^{k-1} \frac{(k-1)p_k p_j}{\mu} + \sum_{j=1}^{k-1} \frac{(j-1)p_j p_k}{\mu} \\
&\quad + \frac{(k-1)p_k^2}{\mu} \\
&= \frac{p_k}{\mu} \left((k-1) \sum_{j=1}^k p_j + \sum_{j=1}^{k-1} (j-1)p_j \right)
\end{aligned}$$

□

3.3.2 The Erdős-Renyi Random Graph

To derive the distribution of edge weights, note that the joint degree distribution $p_{kk'}$ of finding a vertex of degree k and a vertex of degree k' attached to a randomly drawn edge of a $G(n, c/n)$ is given by

$$p_{kk'} = \frac{kk'}{c^2} p_k p_{k'} = p_{k'k},$$

where $\{p_k\}$, is the usual (Poisson-) degree distribution given by $p_k = \exp(-c)c^k/k!$.

Lemma 3.4 *Let $G(n, p)$ be an Erdős-Renyi random graph with $p = c/n$. Then the distribution $P(\gamma(e) = 1/k)$ of communication edge weights is given by*

$$P\left(\gamma(e) = \frac{1}{k}\right) = \exp(-2c) \frac{c^{2k-2}}{(k-1)!} \left(\frac{1}{(k-1)!} + 2 \sum_{j=1}^{k-1} \frac{c^{j-k}}{(j-1)!} \right).$$

Proof.

$$\begin{aligned}
P\left(\gamma(e) = \frac{1}{k}\right) &= p_{kk} + \sum_{j < k} p_{kj} + \sum_{j < k} p_{jk} \\
&= p_{kk} + 2 \sum_{j=1}^{k-1} p_{kj} \\
&= \frac{k^2}{c^2} p_k^2 + 2 \sum_{j=1}^{k-1} \frac{kj}{c^2} p_k p_j \\
&= \frac{k}{c^2} p_k \left(k p_k + 2 \sum_{j=1}^{k-1} j p_j \right).
\end{aligned}$$

Inserting $p_k = \exp(-c)c^k/k!$ gives

$$\begin{aligned}
P\left(\gamma(e) = \frac{1}{k}\right) &= \frac{k}{c^2} \frac{c^k}{k!} \exp(-c) \left(k \frac{c^k}{k!} \exp(-c) + 2 \sum_{j=1}^{k-1} j \frac{c^j}{j!} \exp(-c) \right) \\
&= \exp(-2c) \frac{c^{2k-2}}{(k-1)!} \left(\frac{1}{(k-1)!} + 2 \sum_{j=1}^{k-1} \frac{c^{j-k}}{(j-1)!} \right).
\end{aligned}$$

□

With that we calculate the expected value $\mathbb{E}\gamma(e)$ of a randomly drawn edge e .

$$\begin{aligned}
\mathbb{E}\gamma(e) &= \sum_{k=1}^{\infty} \frac{1}{k} P\left(\gamma(e) = \frac{1}{k}\right) \\
&= \exp(-2c) \sum_{k=1}^{\infty} \frac{c^{2k-2}}{k!} \left(\frac{1}{(k-1)!} + 2 \sum_{j=1}^{k-1} \frac{c^{j-k}}{(j-1)!} \right). \quad (14)
\end{aligned}$$

In the definition of the communication index vertices with zero degree have been assumed to be not communicating at all and have therefore been excluded. However, for the Erdős-Renyi random graph each vertex v has a certain probability of having degree zero, namely $P(\deg(v) = 0) = (1 - c/n)^{n-1} \approx \exp(-c)$. Thus, on average there are $n \exp(-c)$ vertices with zero degree and $n^* = (1 - \exp(-c))n$ vertices with non-zero degree. The exclusion of vertices with zero degree leads to a refined definition of the communication index where the normalization uses n^* rather than n ,

$$\gamma(G(n, p)) = \frac{2}{n^*} \sum_{e \in E} \gamma(e).$$

We now derive a formula to calculate the communication index of Erdős-Renyi random graphs. This shows that, for large n , the communication index is given by average degree times average edge weight (times the zero-degree correction term).

Theorem 3.5 *For the Erdős-Renyi random graph $G(n, p)$, with $p = c/n$, holds*

$$\gamma(G(n, p)) \rightarrow \frac{c\mathbb{E}\gamma(e)}{1 - \exp(-c)}, \quad (15)$$

where convergence is in probability and $\mathbb{E}(\gamma(e))$ denotes the expected value of the communication edge weight of a randomly drawn edge.

Proof. For the $G(n, p)$, $p = c/n$, the expected number $\mathbb{E}\#E$ of edges is given by $\mathbb{E}\#E = \binom{n}{2}p = c(n-1)/2$. From the theory of Erdős-Renyi random graphs it is well known that, for any $\epsilon > 0$, the probability of finding the actual number $\#E$ of edges in the interval $[c(n-1)/2 - n^{\frac{1}{2}+\epsilon}, c(n-1)/2 + n^{\frac{1}{2}+\epsilon}]$ tends to 1 as n tends to infinity. Furthermore, the probability of finding $\#E$ outside this interval decreases exponentially with n^ϵ .

Hence, for the expected value of the communication index holds

$$\begin{aligned}
\mathbb{E}\gamma(G(n, p)) &= \mathbb{E}\left(\frac{2}{n^*} \sum_{e \in E} \gamma(e)\right) \\
&= \frac{2}{n^*} \mathbb{E} \sum_{e \in E} \gamma(e) \\
&\leq \frac{2}{n^*} \sum_{i=1}^{c(n-1)/2 + n^{\frac{1}{2} + \epsilon}} \mathbb{E}\gamma(e) \\
&= \frac{2}{n^*} \left(\frac{(n-1)}{2}c + n^{\frac{1}{2} + \epsilon}\right) \mathbb{E}\gamma(e) \\
&= \frac{2}{n(1 - \exp(-c))} \left(\frac{(n-1)}{2}c + n^{\frac{1}{2} + \epsilon}\right) \mathbb{E}\gamma(e) \\
&\rightarrow \frac{c\mathbb{E}\gamma(e)}{1 - \exp(-c)}
\end{aligned}$$

The lower bound works analogously with $c(n-1)/2 - n^{\frac{1}{2} + \epsilon}$ for the number of edges instead of $c(n-1)/2 + n^{\frac{1}{2} + \epsilon}$.

Finally, using the Markov inequality and $\gamma(e) \leq 1$, for all edges e , yields, for any $\delta > 0$,

$$\begin{aligned}
P(|\gamma(G(n, p)) - \mathbb{E}\gamma(G(n, p))| \geq \delta) &\leq \frac{\text{Var}(\gamma(G(n, p)))}{\delta^2} \\
&= \frac{4 \text{Var}(\sum_{e \in E} \gamma(e))}{n^2 \delta^2} \\
&\leq \frac{4(c(n-1)/2 + n^{\frac{1}{2} + \epsilon})}{n^2 \delta^2} \rightarrow 0,
\end{aligned}$$

where we used the bound $\#E \leq c(n-1)/2 + n^{\frac{1}{2} + \epsilon}$ in the last line.

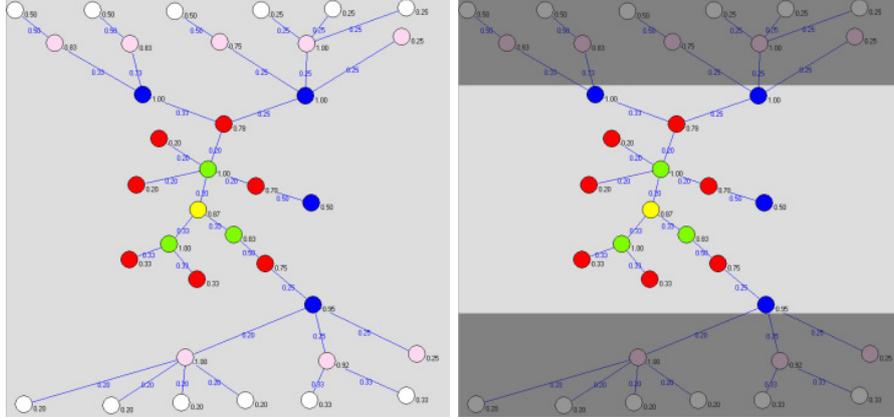
This shows that

$$\gamma(G(n, p)) \rightarrow \mathbb{E}\gamma(G(n, p)) = \frac{c\mathbb{E}\gamma(e)}{1 - \exp(-c)}.$$

□

3.4 Numerical Results

All figures presented in this section are due to the work of Andreas Krueger. He used an application written in Python to implement the communication index on the Galton-Watson process and on Erdős-Renyi random graphs. The main observation for the Galton-Watson trees is that the communication index, defined as the average node sum, is the average degree times the average edge weight.



(a) Galton-Watson sample

(b) Inner nodes of the sample

Figure 9: Figure 9(a) shows a realization of a Galton-Watson tree with a Poisson(1.5)-offspring-distribution from generation 1 (yellow) to generation 6 (white nodes). Figure 9(b) shows a visualization of the counting scheme to avoid finite size effects. The white nodes are an *artificially last* generation in this finite-size realization (they all have a degree 1), so their edges and their neighboring nodes (pink) have finite-size effects in their communication weights and sums. We thus do the statistics of the communication weights and sums on Galton-Watson trees only up to the third last generation, and call these remaining nodes *inner nodes* (yellow, green, red and blue).

3.4.1 The Galton-Watson Process

Since our result is only valid in the supercritical case and simulations can not include an infinity of generations, we used the following procedure. Given a probability distribution with mean greater one there will be a finite probability that the process survives for infinity. To be precise, we chose a Poisson distribution $Po(\lambda)$ with mean $\lambda = 4$. The computer calculated the first $T = 10$ generations of the process, then interrupted, ending up on a random tree with $\sum_{t=0}^T Z_t$ vertices. To simulate a sample subgraph of the infinite tree, the leaves on the 10th generation are ignored, since they all have degree 1 and would adulterate the resulting communication weights of the edges they belong to. The leaves are used as dummies to produce edges from generation 9 to 10, so the degree of the vertices of generation 9 is correct. For the same reason, we ignore the values of the communication strength of all vertices of the 10th generation.

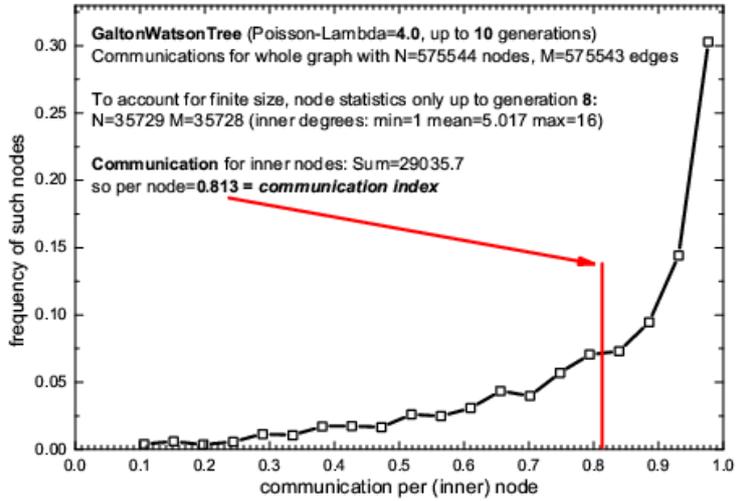


Figure 10: Histogram of communication strengths (node sums) of the *inner nodes* (see figure 9(b)) on a Galton-Watson tree with offspring distribution Poisson(4). The **communication index** is **0.813**.

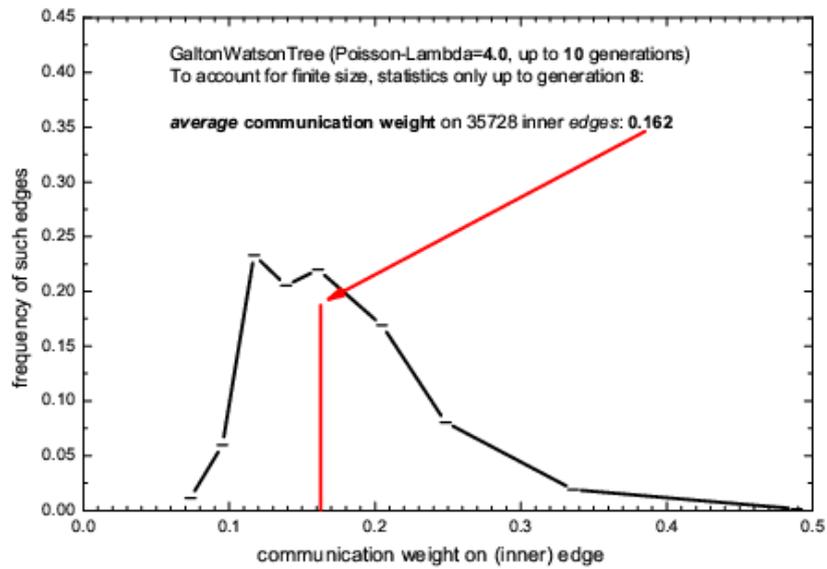


Figure 11: Histogram of the communication edge weights that lead to the figure 10.

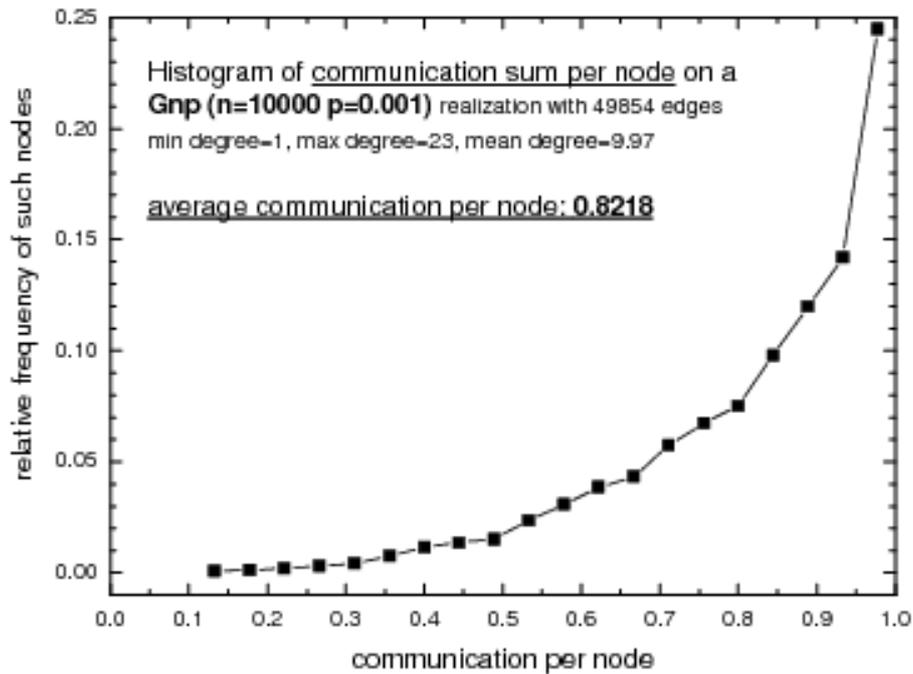


Figure 12: Erdős-Renyi graphs: Communication node-sums histogram on a Gnp realization with $n=10000$, $p=1/1000$. The **communication index** for this network is **0.8218**. Formula (15) predicts $\gamma(G(n, 10/n)) = 0.8228$.

3.4.2 The Erdős-Renyi Random Graph

Here a $G(n, p)$ network generator was used to compute some realizations of Erdős-Renyi random graphs. Then, the distribution of the edge weights was calculated and the communication index for these sample graphs. Figure 13 shows the statistics for the communication weights of the Erdős-Renyi random graph. Plotted on a log-lin scale, as done in figure 14, shows a linear behavior of the communication weights for small values of k . While the edge weight statistics of the Erdős-Renyi random graph differs from the edge weight statistics of the Galton-Watson process, the node statistics do not. We find the same behavior for the node statistics in figure 10 and 12.

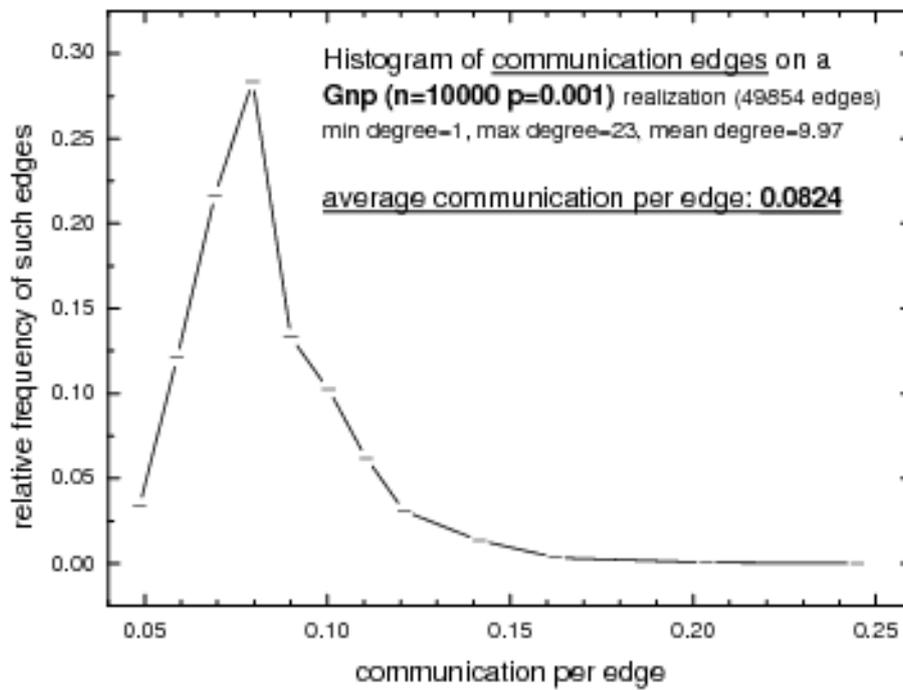


Figure 13: Communication edge-weights histogram of the communication edges that lead to figure 12.

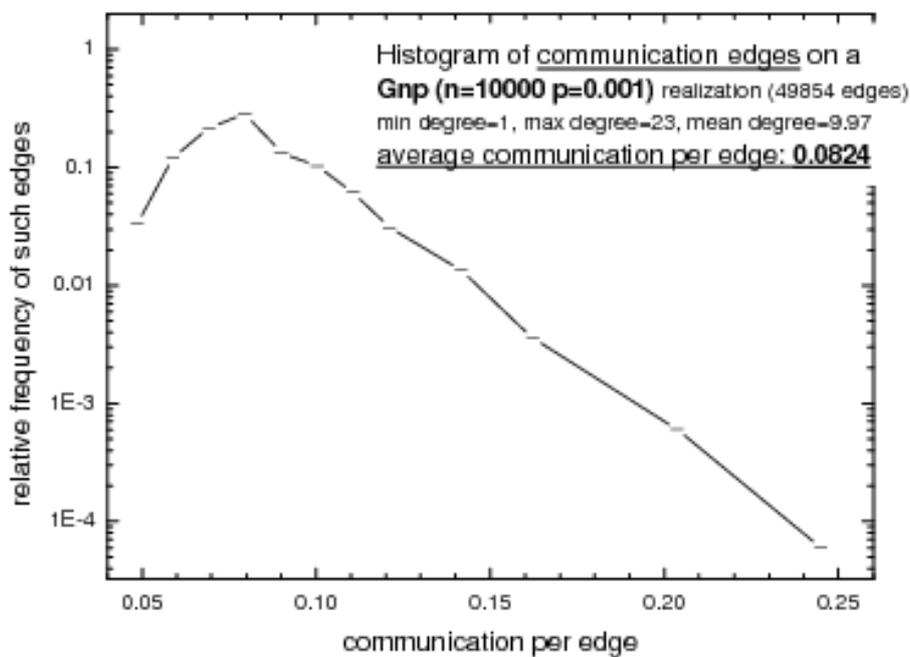


Figure 14: Same as figure 13, but plotted on a log-lin scale.

4 Paper 2: Time-Ordered Information Processing on the Binary Tree

The analysis of communication and its effectiveness on graphs so far was considered as static and taking place simultaneously. Each vertex was assumed to communicate to each of its neighbors at the same time step. In this section we want to outline the work done in paper [15] which sheds a new light on communication modeling by refining the time resolution of communication. Doing so reveals certain effects on the dissemination of information on networks and on the de facto network structure, which can be seen in the following simple example.

Consider three individuals, a supervisor and two subordinates, the supervisor giving orders to both employees, the employees coordinating themselves among each other. Hence, in terms of communication networks they form a triangle. The supervisor gives orders to the first employee in the first half of the day, to the second one in the second half. This leads to the situation that both subordinates can not communicate to each other, since whenever one has free time the other one is occupied in the chiefs bureau. The resulting communication network is a tree of order three rather than a triangle.

This indicates that it is of importance to investigate carefully which networks are capable of communication at all, and which properties they have - or not have.

4.1 Overview

The example above already carries the ideas leading to a formalization of the problem. The most important node in the network (here: the one with most neighbors) defines the time resolution in which communication per classical time step takes place, i.e. how many sub-steps one step has. Then, every vertex enumerates its edges with numbers from 1 to the max-degree, interpreted as the order in which a vertex communicates to its neighbors. This is what we call a *schedule* for a vertex. If this can be done in a compatible way, i.e. there is a schedule for each vertex such that there is no edge in the graph with two different numbers, we speak of a *graph schedule*.

When interpreting the numbers as the order at which communication within one time step takes place, it becomes clear that information can only process along paths of increasing numbers, the *admissible paths*. With this handful of definitions some natural questions can be raised: What are the graphs which have a graph schedule and what graphs have none? Given one vertex as a seeder of information, what is the range of communication per time step, i.e. what is the maximal length of an admissible path? How many vertices gain the information per time step, and how long does it take until each vertex in the graph is informed?

Not all of these questions are getting addressed in the paper. We only treat the problem on the binary tree, where there not only a graph schedule exists, the process is also independent of the particular choice of schedule.

4.2 Results

The number $n(g, t)$ of vertices can be calculated as follows.

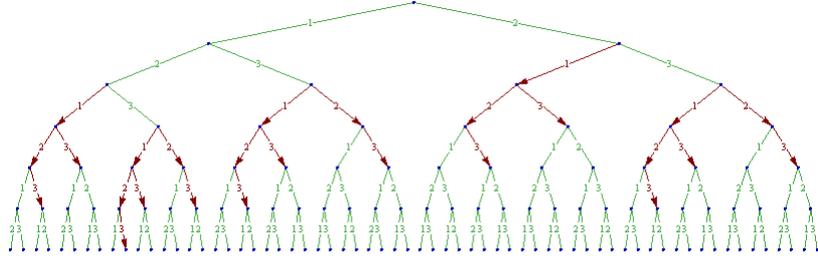


Figure 15: The binary tree with the root as the sender of information. The admissible paths going from all vertices which were infected by the root at $t = 1$. Admissible paths which infect new vertices at $t = 2$ are rendered red. Of the possible 62 vertices on the first 5 generations of the tree, only 31 get the information spread by the root.

Theorem 4.1 *The number $n(g, t)$ of vertices on generation g , that get infected at time t , is given by $n(g, t) = \frac{1}{g!} \frac{d^g}{dx^g} \Phi(x, t) |_{x=0}$, where $\Phi(x, t)$ is the solution of*

$$\begin{aligned} \Phi(x, t) &= (3+x)x^2\Phi(x, t-1) + x^3\Phi(x, t-2) & (16) \\ \Phi(x, 1) &= 2x + 3x^2 + x^3 \\ \Phi(x, 2) &= x^2 + 7x^3 + 11x^4 + 6x^5 + x^6. \end{aligned}$$

Such a recurrence relation with initial conditions may be derived for any kind of graph which admits a schedule. Notice that the right side of (16) carries structural information of the graph while the initial conditions reflect the influence of the particular schedule.

From the above we calculate the total number $n(t)$ of vertices which gain information at time t and the overall number $\rho(t)$ of vertices with information.

Lemma 4.2

$$\begin{aligned} n(t) &= \left(1 - \frac{1}{\sqrt{5}}\right) (2 - \sqrt{5})^t + \left(1 + \frac{1}{\sqrt{5}}\right) (2 + \sqrt{5})^t \\ \rho(t) &= -2 + \left(1 - \frac{2}{\sqrt{5}}\right) (2 - \sqrt{5})^t + \left(1 + \frac{2}{\sqrt{5}}\right) (2 + \sqrt{5})^t. \end{aligned}$$

Time-Ordered Information-Processing on the Binary Tree
S. Delitzscher
to be published

Abstract: In this paper we study the effects on information processing on a graph that come in to play, when a certain time ordering is introduced on that graph. If the communication between vertices does not take place simultaneously but consecutively, the set of possible paths on which information may proceed becomes an interesting entity to investigate.

Find Paper 2 attached to the framework text.

5 Paper 3: Generalized Epidemic Processes and Threshold Percolation with Application to Knowledge Diffusion

With the growing importance of the Internet and its virtual social networks like Facebook or Twitter, personal contact with people and their opinions became much more frequent and therefore more contagious. Individuals do not need to meet friends anymore to take influence or to get influenced, they only need an Internet connection, which nowadays is possible even without computers.

Hence, individuals can easily gain access to large databases of information and entertainment (e.g. Wikipedia, Youtube) and share it online with other individuals who are connected via any social network platform. The entities spread in such a way are examples of what we want to call knowledge diffusion. The main issue in modeling knowledge and the spread of knowledge is that presently there is no single agreed academic definition of knowledge itself. The Oxford English Dictionary [30] defines knowledge the following way:

Knowledge:

- *facts, information, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject*
- *awareness or familiarity gained by experience of a fact or situation*

Since we are not interested in modeling the origination of knowledge, but rather in dissemination of knowledge, we lay our focus on the interactions of bearers of knowledge to characterize knowledge and knowledge spread.

5.1 Active and Passive Knowledge

No matter what knowledge may be or where how it originates, from the definition of knowledge given above it follows that individuals may have different levels of knowledge. Both items include the notion of experience, and different levels of experience imply different levels of understanding and different capabilities of using the knowledge learned or acquired. In addition, this is not only a simple classification, it is a hierarchy with correlation between distinct levels. Highly experienced individuals teach knowledge to lower experienced individuals, e.g. by writing books, teaching classes or giving talks, all of it optionally shared on the Internet to grant public access. Lower experienced individuals consume, gain experience, adopt knowledge, and level up, see figure 16.

To include this circumstance, we divide the state of having knowledge into two sub-states, namely having passive knowledge and having active knowledge. Having passive knowledge here means low level experience, having the ability (and the will¹⁰) to talk about it in a more or less serious way. In contrast, an individual with active knowledge has good expertise and the capability of applying this knowledge or even of developing it further.

¹⁰Individuals who have a certain knowledge, but for personal, occupational or whatever reasons do not want to communicate this, are not contagious and can not be infected. They are immune.

Circuit of Experience

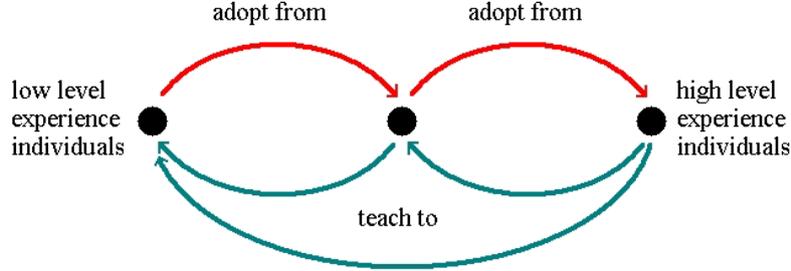


Figure 16: The circuit of experience: High level experienced individuals teach knowledge to mid-level and low-level experienced individuals. Successful adoption helps individuals gain experience: they level up.

Hence, similar to standard epidemiology classifications, in our model we assume that there are four levels or states of having knowledge. These states are: immune, susceptible, having passive knowledge, and having active knowledge. For instance, say this knowledge is the twelve-tone technique in musical composition. Having passive knowledge about it may be joyful listening to music composed with this technique. Nevertheless, it may not enable an individual to compose a successful twelve-tone piece of music itself. For such a project individuals would need active knowledge.

5.2 Generalized Epidemic Processes

When the infection states are introduced, we go on specifying the diffusion process we chose to model knowledge diffusion on a network. Since we defined passive knowledge as the ability and will to communicate, any vertex with passive or active knowledge is assumed to have a certain probability to infect its neighbors with passive (not active!) knowledge. For each vertex v this simple contagion process yields a probability $P^\epsilon(v) = 1 - (1 - \epsilon_p)^{N_p(v)}(1 - \epsilon_a)^{N_a(v)}$, where ϵ_p (ϵ_a) is the probability that a contagious vertex infects another vertex with passive knowledge and $N_p(v)$ ($N_a(v)$) is the number of neighbors of v with passive (active) knowledge.

However, it is not sufficient to describe knowledge diffusion as a simple contagion. It has the property that repetition increases the probability of adoption. Getting in contact with a certain kind of knowledge sufficiently often may change an individuals mind, making it curious to find out more, ending up infected with that knowledge. That individual attained a *threshold*. Hence, we assume that every vertex v has a threshold $\Delta(v)$, and the probability of adopting passive knowledge jumps from (generally very small) $P^\epsilon(v)$ to some high probability α . This combination of simple contagion and complex contagion is what we call *generalized epidemic processes*. Notice that both kinds of contagion are included as special cases, simple contagion is obtained by choosing $\Delta(v) = n$ for all v , where n denotes the overall number of vertices in the network and

the threshold can never be attained. Complex contagion can be obtained by choosing $\epsilon_a = \epsilon_p = 0$.

Further on, we include the possibility of getting infected with knowledge by network independent processes, for instance via Television, Internet or any kind of mass media. Since the supply of certain knowledge in mass media depends on how widespread this knowledge is, we assume that the probability of getting infected by mass media is a function of the densities of already infected vertices, again distinguishing the influence of vertices with active knowledge which mainly provide the content for reports about this knowledge, and vertices with passive knowledge which by their demands provide the market for such reports. Formally, $P^\beta(v) = \min\{\beta_p f(x) + \beta_a f(y)\}$, where $\beta_{a,p}$ are tunable parameters, f is a function on the unit interval, and x (y) is the density of vertices with passive (active) knowledge.

Finally, we included the possibility of forgetting knowledge. At each time step for each vertex with knowledge there is a probability γ that the vertex forgets and transitions to the susceptible state again. Notice that the particular choice of γ is related to the kind of knowledge rather than to individuals. The parameter reflects the relevance of the knowledge, its complexity and difficulty to learn and keep. For instance the parameter would be small or even zero for the knowledge about the date of the terrorist attack on the World Trade Center, due to the catchy internationally well known phrase "9/11".

5.3 Results

Various special cases of the model are discussed analytically. First of all we considered the pure complex contagion process, i.e. the simple contagion parameters $\epsilon_{p,a}$ and mean field parameters $\beta_{p,a}$ were chosen to be zero. Notice that independently of our work, Janson et al. discussed the same process in [26].

To evaluate the equilibrium states of the complex contagion process we assumed that each vertex v has the same threshold $\Delta = \Delta(v)$. On an Erdős-Reyni random graph $G(n, p)$, $p = c/n$, for the mean density s_t of infected vertices at time t holds

$$s_{t+1} = 1 - (1 - s_0)e^{-cs_t} \sum_{k=0}^{\Delta-1} \frac{(cs_t)^k}{k!},$$

as $n \rightarrow \infty$. From the resulting fixed point equation we found that there are either three fixed points or just one in the domain $[0, 1]$. In case of three fixed points the one closest to a_0 is the dynamical stable one. The phase transition occurs in the case when the two smallest fixed points coincide to form an indifferent fixed point with slope 1. From that condition, $f'(x) = 1$, we calculate the critical density a_0^c which is given by

$$a_0^c = 1 - \frac{e^{-\frac{1}{2}(1-c+\sqrt{c^2-3-2c})}}{c(c-1-\sqrt{c^2-3-2c})}. \quad (17)$$

Near that value a long transient behavior of the system can be observed, until the system reaches the final high prevalence fixed point.

We also discussed the pure mean field process including forgetting, i.e. the parameters α and ϵ of the local process are set to zero. We discuss two special

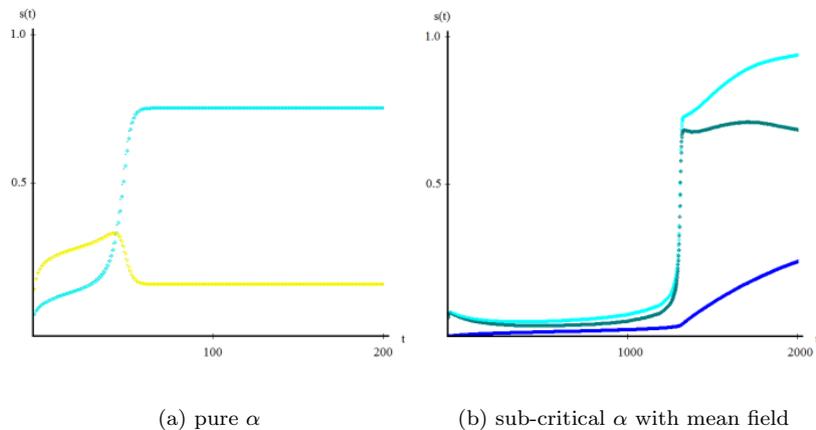


Figure 17: The α -process on $G(n, c/n)$ with $n = 100'000$, $c = 3.5$ and $\Delta = 2$ for all vertices. 17(a) shows the pure α -process for an initial infection $s_0 = 0.06$ ($a_0^c \approx 0.058$). The turquoise colored line is the total density of infected vertices, yellow shows the density of semi-infected vertices. 17(b) shows a sub-critical α -process, $s_0 = 0.05$, with an activated mean-field process for active vertices, $\beta_2 = 0.2$. The rate ρ of production of vertices with active knowledge is $\rho = 0.001$. The turquoise line is the total density of infected vertices, blue the density of vertices with passive knowledge, green the density of vertices with active knowledge.

cases of the mean field process function f , namely quadratic proportionality $f(x) = x^2$, and linear proportionality $f(x) = x$. In both cases we again observe critical values for the initial density a_0 . Also we discussed the influence of a set of vertices which is almost 'immune' to forgetting. Such subsets can exist due to the local threshold process when the infection probability α above the threshold is set to 1. Then every vertex with at least Δ (we chose $\Delta = 2$) infected neighbors will immediately get reinfected after forgetting. We show that the existence of such an immune set lowers the value of the critical initial density. In combination with the α -process which may create such subsets, this means that starting with a value below the critical initial density for both may yield a phase transition.

Figure 17 illustrates this effect. While fig. 17(a) shows a pure overcritical α -process on a $G(n, p)$ with critical initial density $a_0^c \approx 0.058$, fig. 17(b) shows a sub-critical α -process with an activated mean field process. Notice that in the beginning the number of infected vertices even decreases. Then the system remains in a transient state, some vertices learn, some forget, until there is a configuration of infected vertices which is almost immune, helping the system to succeed in that way lowered critical density.

Finally, the idea of subsets which are immune to forgetting can be used to determine the size of the k -cores¹¹ of any network. For each vertex choose the threshold $\Delta = k$, then initially infect all vertices and run the simulation

¹¹The k -core of a network is a maximal connected subgraph in which all vertices have degree at least k .

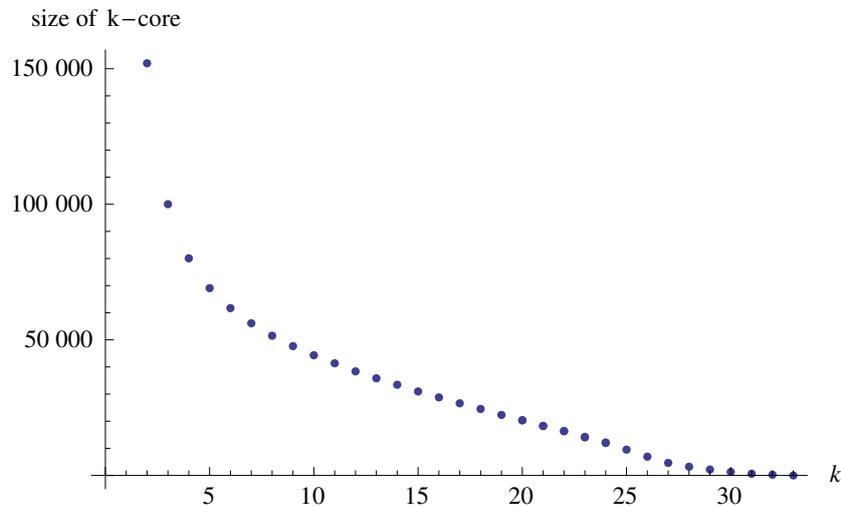


Figure 18: Size of the k -cores of the StudiVz network

only with the α -process, with $\alpha = 1$, and with forgetting turned on. Each vertex with less than k neighbors will forget sooner or later with possibility of reinfection. Each vertex with k or more neighbors will be reinfected immediately after forgetting, so the equilibrium state of the network is the state where all vertices in the k -core are infected, the remaining vertices uninfected. Figure 18 shows the sizes of all k -cores of the Studi-Vz network, an Internet social network of students enrolled at universities in Bielefeld, see [27] for more details on the properties of this network.

Generalized Epidemic Processes and Threshold Percolation with Application to Knowledge Diffusion

P. Blanchard, S. Delitzscher, G. Hiller, T. Krueger, R. Siegmund-Schultze
to be published

Abstract: In this paper we use a generalized epidemic process to model knowledge diffusion in social networks. We find that, using results from threshold percolation theory, there can happen a phase transition in the number of vertices with a particular knowledge if the number of initially knowing vertices is larger than some critical value. In sociological literature this value is often referred to as critical mass, which now can be explained with the existence of thresholds for each vertex in the network.

Find Paper 3 attached to the framework text.

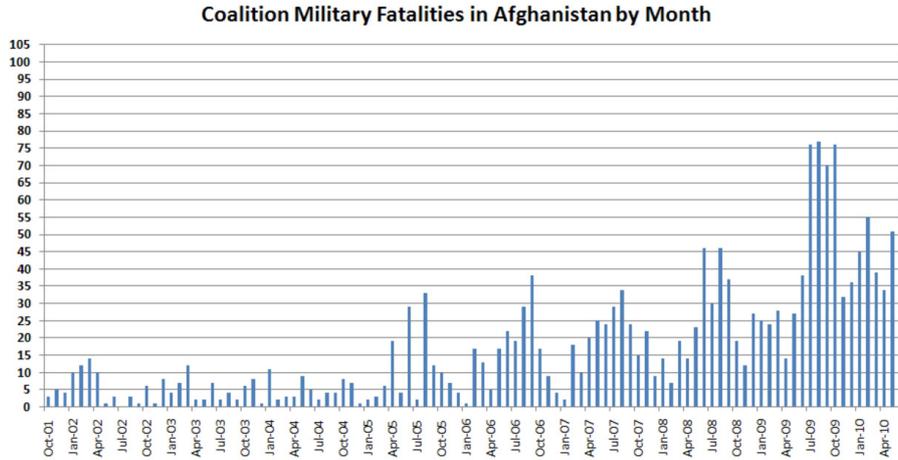


Figure 19: Coalition military casualties in Afghanistan by month. Oscillations are caused by seasonal effects. Source of diagram: <http://www.icasualties.org/oef/>

6 Paper 4: Passive Supporters of Terrorism and Phase Transitions

As one of the many applications of knowledge diffusion on networks, we consider radical political or ethical ideas as knowledge, which spread according to our model. Clearly, the notion of terrorism has to be handled with care, since ones terrorists are another ones freedom fighters and we do not want to take part in discussions about who is a terrorist and who is not. We assume terrorism to be steady acts of violence committed by so called terrorists and having an opponent side, the counter-terrorists, which tries to stop attacks by removing the originators from the network.

6.1 The Model

To represent the social network of a country, a district or a village, we chose a graph $G = (V, E)$ on a finite set of vertices V and a set of edges E . We assume that vertices can be in one of four states $\{0, 1, 2, 3\}$. State 0 represents the susceptible part of the population, any individual which has a neutral opinion on terrorists. State 1 represents the passive supporters of terrorism. This includes any individual that helps terrorists by acts of support, ranging from hiding terrorists, providing them with cash, information or other supplies, up to keeping still about terrorists activities. Note that individuals are not necessarily followers of the ideologies the terrorists stand and fight for. They may be suffering under violent methods like torture or they are blackmailed into passive support or they just get paid for it. State 2 denotes active terrorists, those individuals which are planning and sooner or later committing the acts of violence which we call terrorism. Finally, state 3 encodes all vertices which which are isolated from the network and do not interact anymore with other vertices.

Since radical political ideologies are a special case of knowledge, we used

the generalized epidemic process approach we already used to model knowledge diffusion, but with some minor changes. We assume that every vertex in one of the contagious states 1 or 2 has a probability ϵ of infecting susceptible vertices, i.e. of becoming a passive supporter. Also every vertex v has a threshold $\Delta(v)$. In contrast to the knowledge diffusion model, this threshold is not due to the dramatical infection probability of repetition effects displayed by learning. It should be viewed as a rioting threshold in the sense of Granovetter (see above), i.e. whenever there are enough passive supporters and active terrorists in an individuals environment the probability of joining in increases dramatically to a given high probability α . And even if individuals would like to withstand this social pressure, the passive supporters and active terrorists in their environment are not likely to tolerate neutral behavior.

Hence, for a vertex i with $N_i(t)$ infected neighbors at time t the probability P_{loc} of changing state χ_i from neutral 0 at time to passive supporter 1 by the local infection process is given by

$$P_{loc} \{ \chi_i(t+1) = 1 \mid \chi_i(t) = 0 \} = \begin{cases} \epsilon N_i(t), & N_i(t) < \Delta_i \\ \alpha, & N_i(t) \geq \Delta_i \end{cases}$$

where we used the standard estimation for $1 - (1 - \epsilon)^N \approx \epsilon N$ in case of small ϵ .

Also we included a mean-field process which represents the effects of collateral damage. Collateral damage (from lat. *collateralis*, on the side) is an euphemism for civilian casualties caused by a military action. We assume that due to action of external forces like the state or allied troops active terrorists get removed from the network at a rate ρ . Since such actions may involve civilian casualties, and friends, relatives, and otherwise emotionally involved are - in case of having been neutral up to that event - likely to change their attitude and start to sympathize with the troops opposing force, ending up as a passive supporter of terrorism. Even if there are no casualties, reports on tortured and humiliated prisoners in Television and on the Internet may have the same effect. Therefore we assume that for each terrorist removed from the network (which happens at some rate ρ) a fraction κ of the population turns to passive supporters.

Further on, we assume that active terrorist are recruited at rate k , and that there is a probability γ that terrorists and passive supporters change their mind and change back into the susceptible state 0.

6.2 Results and Conclusions

Since the local process is the same generalized epidemic process used in modeling knowledge, the critical initial density is given by formula (17). The alternative definition of the mean field process also yields a critical value, namely if

$$\frac{k\rho\kappa n}{(\gamma + k(1 - \gamma))(\gamma + \rho(1 - \gamma))} > 1,$$

then the process becomes over-critical.

Simulations with several choices of parameters were done, some of them are depicted in fig. 20 and fig. 21. In fig. 20 an Erdős-Renyi random graph $G(n, c/n)$ was chosen as underlying network, with $n = 50'000$ and $c = 4$.

The main observation is the existence of a phase transition in the number of passive supporters. This contradicts the common military strategy of "throw

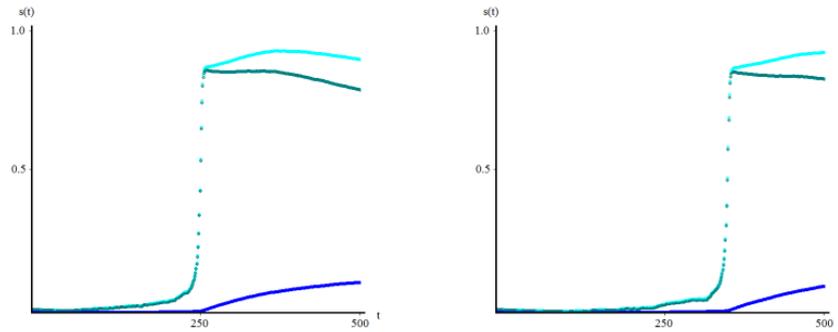


Figure 20: Densities of active terrorists and passive supporters on an Erdős-Renyi random graph $G(n, p), p = c/n$ with $n = 50'000$ and $c = 4$, threshold $\Delta = 2$, initial density of passive supporters $b_0 = 0.01$, mean field $\kappa = 0.001$, in a time span of $t = 500$ steps. Choosing $\rho = 0.0004$ (left) or $\rho = 0.00025$ (right) only alters the moment when the phase transition happens.

a bomb - make an enemy; donate a gift - make a friend” which is designed to keep the civilian population on the side of the counterterrorist forces in case of collateral damage. The network maintains a transient state in which some passive supporters change back to the susceptible state, some individuals turn from susceptible to passive supporter, until this rearranging process yields a configuration where the α -process creates a huge cascade of infections, changing the phase state of the network from almost all vertices are susceptible to almost all vertices are passive supporters.

Figure 21 shows the simulation run on the StudiVz network. There, it can be observed that already one or two cases of collateral damage are sufficient to yield the phase transition.

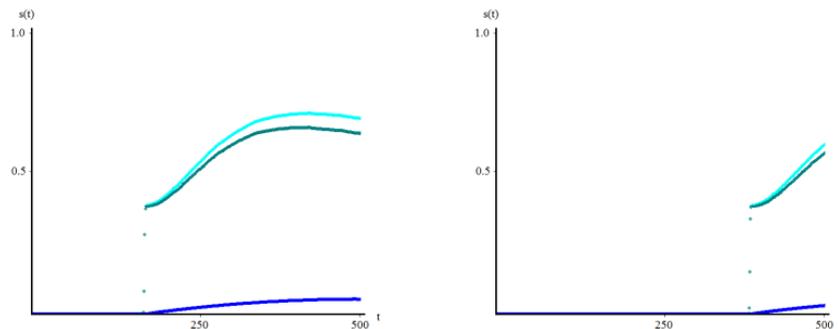


Figure 21: Densities of active terrorists and passive supporters on a sample of the StudiVz network, two single runs. The threshold is $\Delta = 2$, initial density of passive supporters $b_0 = 0.00005$, mean field $\kappa = 0,0001$, in a time span of $t = 500$ steps. The phase transition occurs as soon as one or two active terrorists get captured.

Passive Supporters of Terrorism and Phase Transitions

F. August, P. Blanchard, S. Delitzscher, G. Hiller, T. Krueger

NATO Science for Peace and Security Series - E: Human and Societal Dynamics, Volume 75, 2010, Complex Societal Dynamics - Security Challenges and Opportunities, Edited by K. Martínás, D. Matika, A. Srbljinovic, pp 190-198, arXiv:1010.1953v2 [physics.soc-ph]

Abstract: We discuss some social contagion processes to describe the formation and spread of radical opinions. We use threshold dynamics to describe the local spread of opinions, and mean field effects. We calculate and observe phase transitions in the dynamical variables resulting in a rapidly increasing number of passive supporters. This strongly indicates that military solutions are inappropriate.

Find Paper 4 attached to the framework text.

7 Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich bei der Verwirklichung dieses Projekts unterstützt haben.

Ich bedanke mich herzlich bei meinen Betreuern, die es mir ermöglicht haben, eine solche Arbeit anzufertigen und in diesem faszinierendem Themengebiet einen Beitrag zu leisten, Prof. Dr. Philippe Blanchard und Dr. Tyll Krueger, der mir im Laufe der Zeit Mentor und Freund geworden ist.

Ebenso bedanke ich mich bei Prof. Dr. Niloy Ganguly für die Möglichkeit an seiner Forschung teilzuhaben und ein aufregendes Land kennen lernen zu dürfen.

Ich danke der Universität Bielefeld für alle Unterstützung, besonders Hanne Litschewsky, deren Zuverlässigkeit und Aufopferungsbereitschaft für die ihr Anvertrauten beispiellos ist.

Besonderer Dank gilt der Fachschaftsinitiative der Fakultät für Physik der TU Berlin, allen Mitgliedern und Nahestehenden, für Geduld, Verständnis und Beistand.

Größter Dank gebührt meinen Eltern, Adeline und Hans, ohne deren Unterstützung und Vertrauen ich diese Arbeit niemals hätte vollenden können.

References

- [1] R. Albert, A.-L. Barabási, "Statistical mechanics of complex networks", *Rev. Mod Phys.* 74 (2002), 47-97.
- [2] R. Albert, A.-L. Barabási, "The Emergence of Scaling in Random Networks", 1999, *Science* **286**, 509
- [3] R. A. Berk, "A Gaming Approach to Crowd Behavior", *American Sociological Review*, 1974, Vol. 39, 355-73
- [4] P. Blanchard, T. Krueger, "The 'Cameo Principle' and the Origin of Scale-Free Graphs in Social Networks", *Journal of Statistical Physics*, Vol. 114, Numbers 5-6, 1399-1416
- [5] B. Bollobás, S. Janson, O. Riordan, "The Phase Transition In Inhomogeneous Random Graphs", *arXiv:math/0504589v3 [math.PR]*
- [6] B. Bollobás, "Random Graphs", 2. ed, Cambridge University Press, 2001
- [7] B. Bollobás, "Mathematical results on scale-free random graphs", *Handbook of Graphs and Networks: From the Genome to the Internet*, Wiley, 2003
- [8] E. Callen, D. Shapiro, "A theory of social imitation", *Physics Today*, 23-28 (July 1974)
- [9] D. Centola, M. Macy, "Complex Contagion and the Weakness of Long Ties", *AJS Volume 113 Number 3 (November 2007): 702-34*
- [10] J. Chandra, S. Delitzscher, N. Ganguly, A. Jhunjhunwala, T. Krueger, N. Sharma, "Optimizing Topology in Bit Torrent Based Networks", *Computer Communications Workshops (INFOCOM WKSHP)*, 2011 IEEE Conference on, Issue Date: 10-15 April 2011, 888 - 893
- [11] F. Chung, L. Lu, "Complex Graphs and Networks", Vol. 107 of *CBMS Regional Conference Series in Mathematics*
- [12] L. da Fontoura Costa, O. N. Oliveira Jr., G. Travieso, F. A. Rodrigues, P. Ribeiro Villas Boas, L. Antigueira, M. P. Viana, L. E. C. da Rocha, "Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications", *arXiv:0711.3199v3 [physics.soc-ph]*
- [13] M. Dehmer, "Information-theoretic Concepts for the Analysis of Complex Networks", *Applied Artificial Intelligence*, Volume 22 Issue 7-8, August 2008
- [14] M. Dehmer, F. Emmert-Streib, "Information theoretic measures of UHG graphs with low computational complexity", *Applied Mathematics and Computation*, 190 (2007), 2, S. 1783 - 1794.
- [15] S. Delitzscher, "Time-Ordered Information-Processing on the Binary Tree",
- [16] R. Durrett, "Probability: Theory and Examples", 1991.

- [17] R. Durrett, "Random Graph Dynamics", Cambridge University Press New York, NY, USA, 2006
- [18] P. Erdős, A. Renyi, "On Random Graphs, I", *Publicationes Mathematicae (Debrecen)*, Vol. 6 (1959), pp. 290-297
- [19] P. Erdős, A. Renyi, "On the evolution of random graphs", *Publ. Math. Inst. Hung. Acad. Sci.*, Vol. 5 (1960), pp. 17-61
- [20] S. Galam, "Sociophysics: A Personal Testimony", arXiv:physics/0403122v1 [physics.soc-ph]
- [21] S. Galam, Y. Gefen, Y. Shapir, "Sociophysics: A mean behavior model for the process of strike", *Journal of Mathematical Sociology* 9, 1-13 (1982)
- [22] M. Granovetter, "Threshold Models of Collective Behavior", *The American Journal of Sociology*, 1978, Vol. 83, No. 6, 1420-1443
- [23] M. Granovetter, "The Strength of Weak Ties", *American Journal of Sociology* 78:1360-80
- [24] M. Grodzins, "Metropolitan Segregation", *Scientific American*, Vol 197(4), 1957, 33-41
- [25] T. E. Harris, "Branching Processes", *AMS* 19, 474-494 (1948).
- [26] S. Janson, T. Luczak, T. Turova, T. Vallier, "Bootstrap percolation on the random graph $G(n,p)$ ", arXiv:1012.3535v1
- [27] C. Lee, T. Scherngell, M. J. Barber, "Real-world separation effects in an online social network", 2009, arXiv:0911.1229v1 [physics.soc-ph]
- [28] A. Mayer, R. Woods, "Change without Conflict. A Case Study of Neighborhood Transition in Detroit.", in N. Glazer (ed.) *Studies in Housing and Minority Groups*, University of California Press, 1960
- [29] J. Travers, S. Milgram, "An Experimental Study of the Small World Problem.", *Sociometry*, Vol. 32, No. 4, pp. 425-443, 1969
- [30] http://oxforddictionaries.com/view/entry/m_en_us1261368
- [31] B. Pourebrahimi, K.L.M. Bertels, S. Vassiliadis, "A Survey of Peer-to-Peer Networks", Proceedings of the 16th Annual Workshop on Circuits, Systems and Signal Processing, ProRisc 2005, November 2005.
- [32] O. Pikhurko, O. Verbitsky, "Logical complexity of graphs: a survey", arXiv:1003.4865v3
- [33] E. M. Rogers, "Categorizing the adopters of agricultural practices", *Rural Sociology* 23: 345-354
- [34] T. C. Schelling, "Dynamic Models of Segregation", *Journal of Mathematical Sociology*, 1971, Vol. 1, 143-186
- [35] T. C. Schelling, "On the Ecology of Micromotives", *Public Interest*, No. 25, 61-98

- [36] T. C. Schelling, "A Process of Residential Segregation: Neighborhood Tipping" *Racial Discrimination in Economic Life*, edited by A. Pascal, Lexington, Mass.: Heath
- [37] G. Simonyi, "Graph entropy: A survey", *Combinatorial Optimization*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 20, AMS, 399-441.
- [38] T. W. Valente, "Social network thresholds in the diffusion of innovations" *Social Networks* 18, 1996, 69-89
- [39] D. J. Watts, "A simple model of global cascades on random networks", *PNAS* April 30, 2002 vol. 99 no. 9 5766-5771
- [40] D. Watts, S Strogatz, "Collective Dynamics of 'Small World' Networks", *Nature* 393:440-42
- [41] W. Weidlich, "The statistical description of polarization phenomena in society", *Br. J. math. statist. Psychol.* 24, 251-266 (1971)

Optimizing Topology in Bit Torrent Based Networks

Joydeep Chandra⁺, Sascha Delitzscher*, Niloy Ganguly⁺,
Ashish Jhunjhunwala⁺, Tyll Krueger*, Naveen Sharma⁺

* Department of Physics, Bielefeld University, Germany.

⁺ Department of Computer Science & Engineering, Indian Institute of Technology,
Kharagpur, India.

Abstract

In this paper, we discuss the importance of the network connectivities of the peers in Bit Torrent based systems in determining the download performance of the peers. In this context, assuming that the fraction of the peers of each bandwidth are known, we derive optimal connectivities of the peers that help to improve the average latency of the peers. We represent the topology of a Bit Torrent based system as a weighted graph, where the average edge weight of the graph directly relates to the download latency of the peers. We formulate the average edge weight of the whole system as a linear function of the fraction of the edges that connect peers of different bandwidth and derive the topology that maximizes the average edge weight of the network. Simulation results based on the Bit Torrent protocol validates the fact that in the optimal topology, peers have 13% better download latency as compared to topologies formed in the normal Bit Torrent based systems. Further the obtained topology also improves the fairness of the system as compared to normal Bit Torrent significantly.

1 Introduction

The popularity of Bit Torrent as a file sharing protocol has grown immensely in the last few years, thus gaining huge research interest in the scientific community. A major research objective in the case of Bit Torrent systems is to improve the download performance of the peers by addressing several key issues like incentive mechanisms, piece and peer selection mechanisms, and fairness issues.

Another important aspect that determines the download performance of the peers is their own bandwidth as well as the bandwidth of their neighbors. The neighbors of a peer are randomly selected by an entity called *tracker*, which the peers contact while joining the network. However at any particular time, links to only a subset of these neighbors remain active (these links are formed based on a set of rules elaborated in section 2), as pieces are transferred through these links. We refer to these neighbors, the links to which are active at a particular time, as the *active neighbors* and the topology formed by these set of active neighbors as the *active topology*. To distinguish from the active neighbors, we refer to all the neighbors (inclusive of both active as well as non-active) as the *static neighbors* and the topology formed by these static neighbors as the *static topology* (ref. figure 1). Hence in this context, selecting suitable static

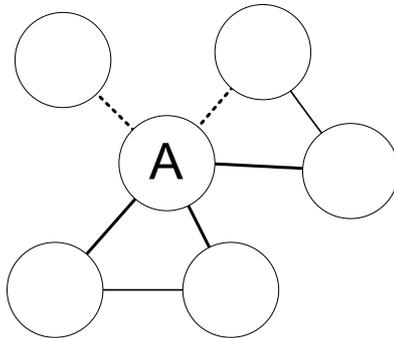


Figure 1: The static and active neighbors of peer A in a Bit Torrent system. The set of active neighbors (indicated by the solid lines) is a subset of the set of static neighbors.

neighbors based on the peers' bandwidths can be an effective technique to improve the download performance of the peers. However, researches have been mainly directed towards improving peer associations in the active topology through various means like developing better incentive mechanisms, piece and peer selection strategies etc. Improving the static topology has not been considered previously; the main reason being that previous experimental results [1] indicated that the download performance of the peers in a Bit Torrent system is dependent on the active topology which, researchers concluded, is independent of the underlying static topology that is actually formed. Our simulation of the assortativity coefficient of the peers, based on peer bandwidth, in the active and static topology (shown in figure 2(a)) also indicates that while the static topology is fairly random, the active topology is largely correlated and hence clusters of similar bandwidth peers will appear. However beyond this apparent independence if one looks in the entire spectrum of possible static networks, the observations are contrary. A major contribution of this paper is to report the observation that the nature of active topology *do* correlate with the static topology when its assortativity¹ is pushed (both in negative and positive direction) beyond a point. We experimentally validate this statement that we discuss next.

1.1 Validation of the Importance of Static Topology

We simulated the assortativity coefficient r [2] of the nodes, based on their bandwidths, for the static and calculated the coefficient of the active network with time, the results of which are shown in figure 3. The figures indicate a huge dependence of the active network on the static network topology. When the static network is assortative, peers tend to exchange more number of pieces with its similar bandwidth neighbors. This indicates that the active network is also assortative. Similarly, when the static network is anti-assortative, peers exchange more number of pieces with dissimilar bandwidth neighbors. When the active network is assortative, the high bandwidth peers benefit

¹Assortativity of the peers is defined as the probability of mixing of peers of similar types, like here peers of similar bandwidth

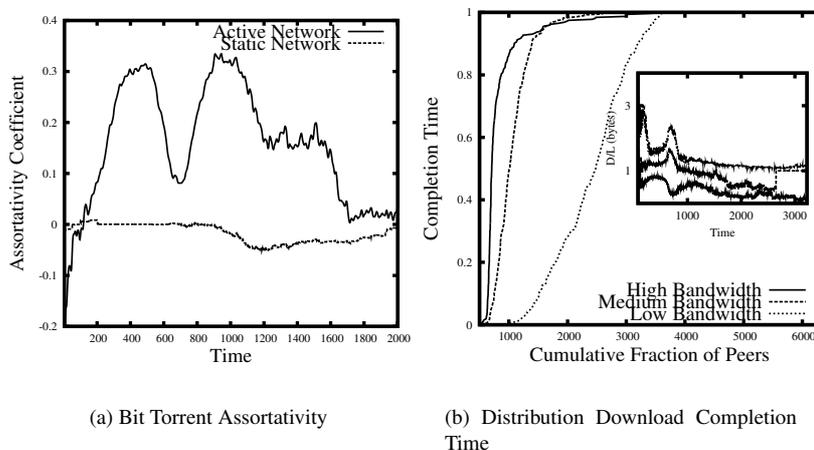


Figure 2: Figure 2(a) shows the assortativity of the active and static topologies in normal Bit Torrent protocol. Figure 2(b) shows the cumulative distribution of the download completion time of the high, medium and low bandwidth peers in normal Bit Torrent protocol. Figure in inset shows the ratio of bytes downloaded and uploaded over the time by the high, medium and low bandwidth peers.

more as compared to the low bandwidth peers in terms of download latency; the reverse situation occurs in case of anti-assortative networks, where the download latency of the low bandwidth peers improves much as compared to the high bandwidth ones. This dependence gives us the hope that optimizing the static graph will yield improvement in the average download performance of the peers. An optimized static/active graph essentially means that the flow of information, hence the download of pieces, is maximized. We next summarize our objectives.

1.2 Objectives

The static network of a Bit Torrent system is represented using a weighted graph, where the nodes of the graph represent the peers and the edges represent the links of the static network. The weights of these edges are determined by the bandwidth categories of the peers that are connected through these edges. Since the edge weights represent the volume of information flow possible through the links, the edge weight of a link is thus a representative of the download performance achieved by the corresponding peers of the link. The edge weight can be determined by the nature of the nodes (peers) joining it. For example, a high (low) bandwidth peer connected with a high (low) bandwidth peer results in a high (low) edge weight, while a high connected to a low bandwidth peer will yield an edge weight somewhere between high and low. Hence the first step of constructing optimum topology is to choose the categories of edges in such a fashion so that *the average edge weight, hence information flow* within the network is maximized. The next step would be actually to build up the static topology satisfying or largely

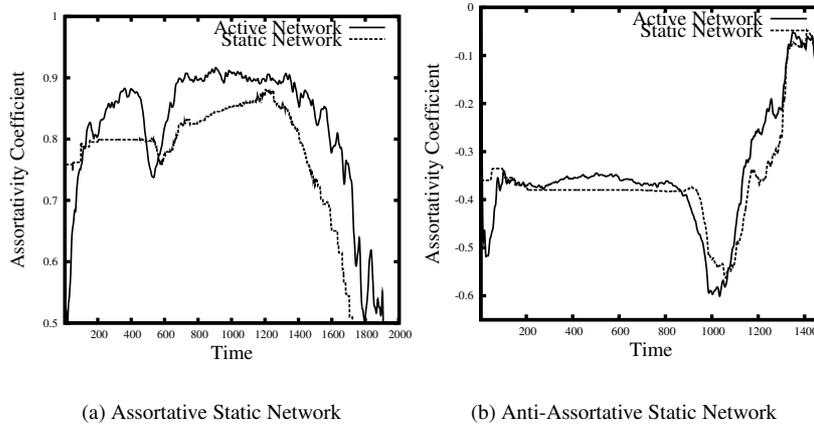


Figure 3: The correlation of the assortativity of the active network and static network for assortative [figure 3(a)] and anti-assortative [figure 3(b)] static topology.

conforming to the above constraints.

Rest of the paper is organized as follows. We next provide a brief overview of the Bit Torrent protocol performance. In section 3 we attempt to derive optimal topologies for a given set of bandwidth categories and fraction of nodes in each category. The simulation results are presented in section 4. Finally, we draw conclusions in section 5.

2 Bit Torrent Overview

We provide here a brief description of the Bit Torrent protocol [3].

Peers willing to obtain a file (say F) initially download a torrent file containing the meta-info of file F . The torrent file contains the address of the tracker as well as the information about the file pieces. The torrent file is opened using a Bit Torrent client software, which connects to the tracker that sends a list of around 50 remote peers, selected randomly from the existing peer set. The peers connect to these remote peers thus forming the static topology, which is largely regular. However, peers can generally download simultaneously from a subset of around 40 peers but can simultaneously upload to a more smaller subset of peers (~ 5) in its neighborhood. This subset of active links forms the active topology and changes after every time slot of 10 seconds (ref. figure 1). In contrast to the static topology, the active topology changes much frequently over time [4].

The file F is broken into smaller pieces which peers exchange among themselves. Peer selection for uploading pieces is done using a CHOKING/UNCHOKING mechanism [5]. Every peer sends an INTERESTED message to its neighbors if it has some missing piece to offer. After an interval of every 10 seconds, every peer selects four neighbor peers preferentially from whom it has recently obtained pieces with the high-

est bandwidth rate (the tit-for-tat principle). The selected neighbors are then said to be unchoked by the peer, which means that the peer will upload requested pieces to them if they are interested. Rest of the neighbors are said to be choked. After every 30 seconds the peer selects a random neighbor (which is not already unchoked) that has sent an INTERESTED message and unchoke it. This process, called optimistic unchoking is primarily aimed towards helping newly arriving peers that does not have any piece to exchange.

Bit Torrent Simulator To simulate the Bit Torrent protocol and to test the proposed optimizations, we have developed a discrete event simulator that follows the actual Bit Torrent official protocol [3], including the newly introduced modified seeder choking algorithm [1]. The bandwidth categories of the peers in the simulator can be tuned to any values; however in this paper we use 2 (high and low) or 3 (high, medium and low) bandwidth categories for the simulations. Further the arrival and departure of the peers, that generates the churn in the network has been modeled according to the recent empirical studies made in [4].

Simulation results of the normal Bit Torrent protocol with equal proportions of high, medium and low bandwidth peers indicate that the download performance is heavily biased towards the high bandwidth peers. The cumulative distribution of the total download time of the peers for low and high bandwidth in a normal Bit Torrent system (figure 2(b)), indicates a huge difference in the download latency of the high and low bandwidth peers (nearly 6 times), although the bandwidth of the high bandwidth peers is nearly 3 times higher than the low bandwidth ones and the ratio of the number of bytes downloaded and uploaded is comparable for the high and low bandwidth peers (inset figure 2(b)). Thus the Bit Torrent system is not fair with respect to the low bandwidth peers. This observation also adds to an extra motivation, whereby one of the objectives which can be set is to ensure fairness. Although we are driving for better performance, fairness of the performance also needs to be tested.

We next formalize the topology optimization problem and discuss solutions for the same.

3 Topology Optimization

We next formalize the topology optimization problem and derive optimal topologies for certain special cases.

3.1 Formalizing the Problem

We model the static network of a Bit Torrent system as a *regular graph* $G(V, E)$, where the set of vertices V represent the peers in the network and E , the set of edges connect them. We assume that the bandwidth of each peer belongs to any one of the categories x_1, x_2, \dots, x_n , where $x_1 < x_2 < \dots < x_n$. Further we assume that the fraction of peers of each bandwidth category is known and is given as $\rho_1, \rho_2, \dots, \rho_n$. Let q_{ij} denote the fraction of edges that connects nodes of bandwidth category x_i with that of x_j .

We assume that for an edge, when the two peers corresponding to the edge belong to the same bandwidth category, say x_i , the weight of the edge is assumed to be x_i . However, if one of the nodes is of bandwidth x_i and the other x_j , where $x_i < x_j$, then the weight of the edge is assumed to be x_j with probability $p_{ij}^{(x_j)}$ and x_i with probability $p_{ij}^{(x_i)} = 1 - p_{ij}^{(x_j)}$ respectively, depending upon whether the peer with bandwidth x_j transfers piece to peer with bandwidth x_i or vice versa. We later derive expressions for $p_{ij}^{(x_j)}$.

If $E(\gamma)$ represents the mean edge weight of the edges of the graph, then $E(\gamma)$, which is our objective function, can be represented as

$$E(\gamma) = \sum x_i q_{ii} + \sum_i \sum_{j:j>i} \left[x_j p_{ij}^{(x_j)} q_{ij} + x_i (1 - p_{ij}^{(x_j)}) q_{ij} \right] \quad (1)$$

We can eliminate the term $\sum x_i q_{ii}$ from the above expression of $E(\gamma)$ by establishing a relation between the fraction of the nodes in a category and its corresponding link fraction, which we state next using a theorem, the proof of which is avoided due to want of space.

Theorem 1 *Suppose in a network of peers of equal degrees (say d), where each peer has bandwidth $x_i \in \{x_1, x_2, \dots, x_n\}$ and the fraction of edges connecting peers with bandwidth x_i and x_j denoted as q_{ij} ($1 \leq i, j \leq n$) is known, then the fraction of peers of bandwidth i can be represented as,*

$$\rho_i = q_{ii} + \frac{1}{2} \sum_{j<i} q_{ij} + \frac{1}{2} \sum_{j>i} q_{ij}. \quad (2)$$

From equation 2 we find that

$$\begin{aligned} \sum_{i=1}^n x_i \rho_i &= \sum_{i=1}^n x_i q_{ii} + \frac{1}{2} \sum_{i=1}^n x_i \sum_{j=1}^{i-1} q_{ij} + \frac{1}{2} \sum_{i=1}^n x_i \sum_{j=i+1}^n q_{ij} \\ &\Rightarrow \sum_{i=1}^n x_i q_{ii} = E(x) - \frac{1}{2} \sum_i \sum_{j:j>i} x_j q_{ij} - \frac{1}{2} \sum_i \sum_{j:j>i} x_i q_{ij}, \end{aligned} \quad (3)$$

where $E(x) = \sum_{i=1}^n x_i \rho_i$ is the average bandwidth of the peers in the network. From equations 1 and 3, we derive the objective function as

$$E(\gamma) = E(x) + \sum_i \sum_{j:j>i} \left[\Delta_{ji} \left(p_{ij}^{(x_j)} - \frac{1}{2} \right) q_{ij} \right], \quad (4)$$

where $\Delta_{ji} = x_j - x_i$ for $i < j$. Thus our objective is to find optimum values of q_{ij} that maximizes $E(\gamma)$ and follows the constraint in equation 2. We next attempt to derive expressions for p_{ij} .

3.2 Deriving $p_{ij}^{(x_j)}$ and $p_{ij}^{(x_i)}$

To determine the value of $p_{ij}^{(x_j)}$, we assume that the link connecting peers with bandwidth x_i and x_j respectively is active under exactly one of the following conditions

1. A piece is being transferred in the direction $x_j \rightarrow x_i$ via regular unchoke mechanism, the probability of which is assumed to be ϕ_{ji} .
2. A piece is being transferred in the direction $x_i \rightarrow x_j$ via regular unchoke mechanism. Since the peers follow a *tit-for-tat* mechanism, then according to the previous case, if the probability that a piece is transferred in direction $x_j \rightarrow x_i$ is ϕ_{ji} , the probability that a piece will be transferred in the opposite direction, i.e. $x_i \rightarrow x_j$ is $\phi_{ij} = \frac{x_j}{x_i} \phi_{ji}$.
3. A piece is being transferred in the direction $x_i \rightarrow x_j$ via optimistic unchoke mechanism, the probability of which is suppose ξ_{ij} .
4. A piece is being transferred in the direction $x_j \rightarrow x_i$ via optimistic unchoke mechanism, the probability of which is ξ_{ji} . Since the peers for optimistic unchoke are selected randomly, we can assume that $\xi_{ij} = \xi_{ji} = \xi$.

Thus assuming that exactly one of the above four events must occur when the link is active, we have

$$\begin{aligned} \phi_{ji} + \xi + \frac{x_j}{x_i} \phi_{ji} + \xi &= 1 \\ \Rightarrow \phi_{ji} &= \frac{(1 - 2\xi)x_i}{x_i + x_j} \end{aligned} \quad (5)$$

Thus $p_{ij}^{(x_j)} = \phi_{ji} + \xi = \frac{(1-2\xi)x_i}{x_i+x_j} + \xi$.

Finding the values of q_{ij} from the expression in equation 4, for any values of n , is difficult. However for most practical cases, the bandwidth categories of the peers are restricted to 2 or 3; we next attempt to find solutions for networks with two bandwidth categories, high and low.

3.3 The Case of 2 Bandwidth Levels

For the case $n = 2$, we consider two bandwidths, high and low, denoted as x_h and x_l respectively and let ρ_h and ρ_l denote the fraction of high and low bandwidth peers. Similar to $p_{ij}^{(x_j)}$ in the n bandwidth category case, for $n = 2$, we denote the probability of a transfer from high to low bandwidth peer as $p^{(x_h)}$, i.e. a link connecting high and low bandwidth peer has weight x_h with probability p and x_l with probability $1 - p^{(x_h)}$. Further, let q_{hh} , q_{ll} and q_{hl} denote the fraction of edges connecting high-high, low-low and high-low bandwidth peers respectively. Then the average edge weight (similar to equation 4) can be derived as

$$E(\gamma) = E(x) + (x_h - x_l) \left(p^{(x_h)} - \frac{1}{2} \right) q_{hl} \quad (6)$$

subject to conditions

$$\rho_h = q_{hh} + \frac{q_{hl}}{2} \text{ and } \rho_l = q_{ll} + \frac{q_{hl}}{2} \quad (7)$$

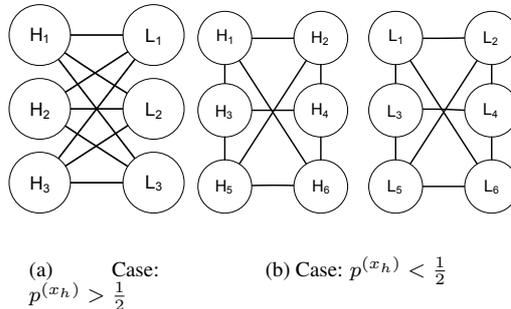


Figure 4: The connectivities of high and low bandwidth nodes, that maximizes the edge weight in a regular graphs, for two cases $p^{(x_h)} > \frac{1}{2}$ (figure 4(a)) and $p^{(x_h)} < \frac{1}{2}$ (figure 4(b)). For $p^{(x_h)} > \frac{1}{2}$, the graph forms a bipartite graph, where no peers of high or low bandwidth are directly connected to each other, on the other hand when $p^{(x_h)} < \frac{1}{2}$, the high and low bandwidth peers gets isolated from each other to maximize the edge weight.

Our objective is to maximize $E(\gamma)$. We find from equation 6 that since the term $E(x)$ is a constant and $x_h - x_l > 0$, hence for $p^{(x_h)} > \frac{1}{2}$, the term $(x_h - x_l) (p^{(x_h)} - \frac{1}{2}) q_{hl}$ becomes greater than zero and thus needs to be maximized for maximizing $E(\gamma)$. Similarly, when $p^{(x_h)} < \frac{1}{2}$, the corresponding term becomes negative and hence needs to be minimized. As can be observed, when $p^{(x_h)} > \frac{1}{2}$, $E(\gamma)$ is maximum when $q_{hl} = 1$ and when $p^{(x_h)} < \frac{1}{2}$, the corresponding value of $q_{hl} = 0$. These values imply that when $p^{(x_h)} < \frac{1}{2}$, to maximize the edge weight the high and low bandwidth peers should have no connections among themselves, thus indicating a total clustering of the peers based on their bandwidth. Correspondingly, for $p^{(x_h)} > \frac{1}{2}$, the edge weight is maximized when $q_{hl} = 1$, i.e. there is a maximum mixing between the peers of different bandwidth. Thus we find that there exists a critical value of $p^{(x_h)}$ for which the nature of the topology changes completely when we intend to maximize the average edge weight (ref. figure 4). However, for practical purpose, to maintain the scalability of the network the network needs to be connected. Hence to maintain a given connectivity of the network and yet maximize the edge weight, we need to do the following.

When $p^{(x_h)} < \frac{1}{2}$, we need to minimize the connections between the different bandwidth nodes. Thus the solution is to partition the graph following a min-cut algorithm, where the partition sizes are known apriori. However the graph partitioning problem where the fraction of nodes in each partition is given is a known NP-Complete problem[6], and efficient heuristics for the solution of the same exists [7].

Similarly when $p^{(x_h)} > \frac{1}{2}$, the problem of partitioning the graph to maximize the average edge weight requires to partition the graph into high and low bandwidth peer sets such that the total edge weight between the two sets is maximum. Thus this problem evolves as a max-cut problem which is also a known NP-complete problem [6] and similar heuristics for approximating the partition exists. For our simulations,

we use the Kernighan-Lin heuristic [7] to generate the min-cut partition of the network.

3.4 The Case of n Bandwidth Levels

On attempting to derive optimal values of q_{ij} for any generic value of n , we find from equation 4 that the average edge weight, $E(\gamma)$ depends on all the values of $p_{ij}^{(x_j)}$ for all possible values of i, j .

From the expression of $p_{ij}^{(x_j)}$ obtained from equation 5, we find that for all $x_i < x_j$, we have $p_{ij}^{(x_j)} > \frac{1}{2}$ if

$$\begin{aligned} \frac{(1-2\xi)x_i}{x_i+x_j} + \xi &> \frac{1}{2} \\ \Rightarrow \xi &> \frac{1}{2}. \end{aligned} \tag{8}$$

This essentially means that optimistic unchoking is performed always. This is an impossible situation, hence $p_{ij}^{(x_j)}$ is less than $\frac{1}{2}$ for any practical situation. Thus in this case, we always need to minimize $E(\gamma)$ (refer equation 4, where $(p_{ij}^{x_j} - \frac{1}{2})$ becomes negative for all values of i, j) and hence the optimal topology can be obtained by iteratively applying the min-cut algorithm for every pair of bandwidth categories.

In the next section, we show with the help of simulations that using our model and partitioning the network accordingly yields substantial improvement in the link utilization and download latency as compared to the random network generated using the tracker.

4 Simulation Results

We discuss the simulation results obtained to validate the proposed models. The Bit Torrent simulator developed by us was briefly described in section 2. In the simulator the static graph gets evolved over time. But the min-cut algorithm is essentially producing a complete static graph. Developing an evolving algorithm which encompasses all the dynamics of node churn is a non-trivial problem and not discussed here. However, for simulation purpose, we assume that this static graph is a representation of the stable underlay arising out of evolution and churn.

We simulated various network parameters like the average edge weight of the network and the average download latency of the peers in the system for various values of q_{ij} , where $i \neq j$, and also for various values of ρ_i , for both 2 and 3 bandwidth categories. The downloading file was broken into 300 pieces; we measured the average edge weight and the average download latency of a peer after it has downloaded 30 pieces. This artificially ensures the stability of the underlay which is required to fully understand the impact of static network. We initially show the correlation between the average download latency of the system and average edge weight of the peers and then show the variation of the average edge weight for various parameters stated above.

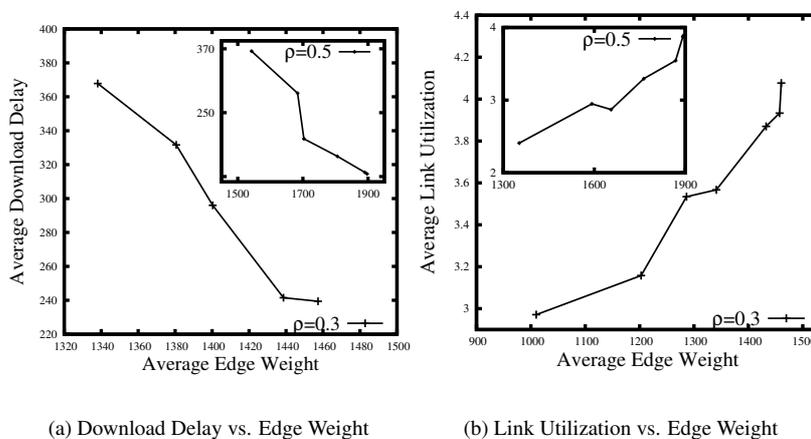


Figure 5: Figure 5(a) shows the relation of the average edge weight and the average download latency of the peers for $\rho = 0.3$ and the figure in inset shows the same for $\rho = 0.5$. Figure 5(b) shows the correlation between average edge weight and link utilization of the peers.

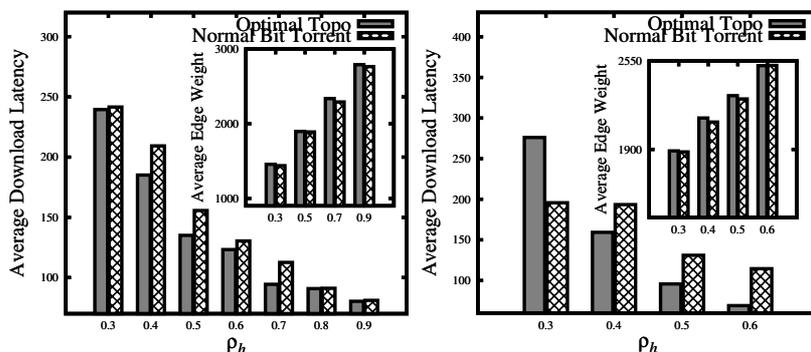
4.1 Download Performance vs. Average Edge Weight

We simulated the average download latency of the system and the average edge weight to establish the correlation between the two. The results, shown in figure 5(a), indicate that the average download latency of the peers decreases with increasing average edge weight. Hence improving the average edge weight will improve the download latency of the peers. Further, we consider another parameter, the link utilization, U , which we define as the average number of links of a peer that are active for download per time slot. A higher value of U indicates more number of parallel download occurring per time slot. We also establish a correlation of U and the average edge weight of the system. Simulation results, shown in figure 5(b) indicate a strong positive correlation between these parameters. The results validate our principal proposition that improving the edge weights in static topologies directly affects the performance of the system.

4.2 Performance of Min-cut Topology

In this section we compare the download performance of the peers in topologies obtained using the min-cut algorithm and the ones formed in normal Bit Torrent systems.

We calculated the average download latency of the peers for the min-cut topologies and normal Bit Torrent systems for various values of ρ_h , for 2 and 3-bandwidth levels. Simulation results for the 2-level case with nearly 2000 peers (shown in figure 6(a)) show that with increasing values of ρ_h , the average download latency of the peers for the min-cut topology steadily improves as compared to the normal Bit Torrent topology. However, for very low values of ρ_h , the average download latency is slightly better in case of normal Bit Torrent topology. This is because of the extreme high proportion of



(a) Average Download Latency Comparison with ρ_h (2 Bandwidth Levels)

(b) Average Download Latency Comparison with ρ_h (3 Bandwidth Levels)

Figure 6: Figure 6(a) shows the variation of average download time for the optimal topology and the normal Bit Torrent protocol with ρ_h and the inset figure shows the corresponding average weight of the peers. Figure 6(b) shows the same for 3 bandwidth levels.

low bandwidth peers that have a slightly better download latency than in the min-cut topology. Figure 6(b) shows very similar results for 3 bandwidth levels, high, medium and low. In the 3 bandwidth level case, for ease of representation, we fixed the fraction of medium bandwidth peers to $\rho_m = 0.3$ and the high and low bandwidth peers have been varied accordingly. The average download latency of the peers in the min-cut topology is nearly 13% lower as compared to the normal Bit Torrent topology for the 2-bandwidth level case, when $\rho_h = 0.5$ and nearly 18% lower in case of 3-bandwidth levels, when $\rho_h = 0.4$, thus indicating a huge improvement in download latency of the peers.

The min-cut algorithm forms a topology with the minimum possible value of cut-point q_{hl} for given values of ρ_h ; in the next section, we observe the average download latency and the average edge weight of the peers for the entire spectrum of the cut-points.

4.3 Effect of Cut-points

We observe the average download latency of the peers and the average edge weight for various value of q_{hl} . Simulation results for the 2-bandwidth level case, shown in figure 7(a), for $\rho_h = 0.3, 0.5$ and 0.7 , reveal that for all the three values of ρ_h , the average download latency of the system increases with increasing values of q_{hl} . For each of these three cases we observed the value of $p^{(x_h)}$ to be lesser than $\frac{1}{2}$, and hence as discussed in section 3.3 that when $p^{(x_h)} < \frac{1}{2}$, the edge weight decreases with increasing q_{hl} (ref. inset figure 7(a)).

The figure shows that the average download latency of the system increases very

slowly with q_{hl} , when the value of q_{hl} is small ($q_{hl} < 0.3$, not shown in figure) and then increases at a faster rate with further increase in q_{hl} . The average edge weight of the system also decreases accordingly with increasing q_{hl} . Thus our observation reveals that topologies similar to the min-cut topology have very similar download performance.

4.4 Effect on Fairness

In this section we discuss the fairness of the system in case of min-cut topology and compare it with the normal Bit Torrent topology. We also observe the change in fairness of the system with q_{hl} . To measure the fairness of a system, we introduce a term called fairness index, which we define as follows:

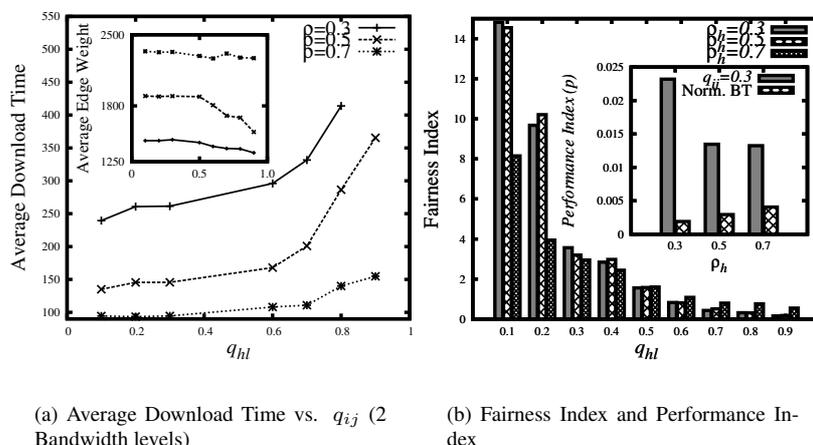
Definition 1 *If d_l and d_h represents the average download latency of the low and high bandwidth peers respectively, the fairness index f of the system in a 2-bandwidth level case is defined as the ratio, $f = \frac{d_h}{d_l}$. The system is considered to be fair if f is nearer to an optimal fairness value $f_o = \frac{b_h}{b_l}$, where b_h and b_l are the download bandwidth of the high and low bandwidth peers respectively.*

Figure 7(b) shows the variation of the fairness, f , of the system in a 2-bandwidth level case with q_{hl} , when the download latency of the high and low bandwidth peers are 3000 Kbps and 800 Kbps respectively, for 3 values of ρ_h (0.3, 0.5, 0.7). Thus the system will be considered as fair if the value of f is nearer to the optimal fairness value of $\frac{3000}{800} = 3.75$. As can be seen, f is far from the optimal value for very low values of q_{hl} as the download latency of the high bandwidth peers are much lower as compared to the low bandwidth ones. The value of f is also very low at higher values of q_{hl} indicating that low bandwidth peers gaining undue advantage over high bandwidth peers.

If we traverse the curve up from low values of q_{hl} , with slight increase the fairness improves very fast and reaches the optimal point. Hence, although the min-cut topology ($q_{hl} = 0.1$) obtained using the min-cut partitioning algorithm is not optimal in terms of fairness; however the fairness is maximum at $q_{hl} = 0.3$, the configuration of which is very similar to the min-cut value of $q_{hl} = 0.1$. Moreover, as seen from figure 7(a), the average download latency of the peers at $q_{hl} = 0.3$ is only slightly greater than in case of $q_{hl} = 0.1$, thus indicating that the topology, which is nearly optimal in terms of the average download latency of the peers as well as the fairness index is very similar to the min-cut topology that we have derived. We state such a topology as a near-optimal topology.

Hence in an effort to measure the effectiveness of a topology we introduce a measure called the *performance index* p , combining fairness and download performance. The performance index of a system is defined as follows:

Definition 2 *If f represents the fairness index of a system with the optimal fairness represented as f_o , and d_a represents the average download latency, then the performance index, p is represented as $\frac{1}{(|f-f_o|) \cdot d_a}$, when $f \neq f_o$, and is represented as $\frac{1}{d_a}$ when $f = f_o$. Note, $|f - f_o|$ indicates how far one is away from the optimal point.*

(a) Average Download Time vs. q_{ij} (2 Bandwidth levels)

(b) Fairness Index and Performance Index

Figure 7: Figure 7(a) shows the variation of average download time of the peers with q_{ij} ($i \neq j$) in a 2-level case for $\rho_h = 0.3, 0.5$ and 0.7 . The figure in inset shows the corresponding average weight of the peers. Figure 7(b) shows the variation of fairness index f with q_{hl} for the same values of ρ . The figure in inset compares the performance index p for the near-optimum topology at $q_{hl}=0.3$ and the normal BitTorrent topology for $\rho_h = 0.3, 0.5$ and 0.7

Thus a higher value of p indicates better performance of the system. Figure 7(b) (inset) compares the performance index of the normal BitTorrent topology and the near-optimum topology with $q_{hl} = 0.3$; as the figure indicates the near-optimum topology has much better fairness (nearly 12 times better for $\rho_h = 0.3$) as compared to the BitTorrent system.

5 Conclusion

The principal contribution of this work is realizing that static network can affect the performance of active networks. However, we did not stop at this realization, using analytical and algorithmic techniques we show that there are optimum topologies which can minimize download latency. But beyond performance maximization, there are fairness issues and we show that there are zones where high fairness is achieved without undermining the performance too much. The next work remains in developing a more realistic model so that the properties of the static graph get imbibed dynamically.

References

- [1] Arnaud Legout, Nikitas Liogkas, Eddie Kohler, and Lixia Zhang, "Clustering and sharing incentives in bittorrent systems," in *Proceedings of the 2007 ACM SIG-*

METRICS international conference on Measurement and modeling of computer systems, New York, NY, USA, 2007, SIGMETRICS '07, pp. 301–312, ACM.

- [2] M. E. J. Newman, “Assortative mixing in networks,” *Physical Review Letters*, vol. 89, no. 20, October 2002.
- [3] BitTorrent, “<http://en.wikipedia.org/wiki/bittorrent>,” .
- [4] Daniel Stutzbach and Reza Rejaie, “Understanding Churn in Peer-to-Peer Networks,” in *IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet Measurement*, New York, NY, USA, 2006, pp. 189–202, ACM.
- [5] Arnaud Legout, G. Urvoy-Keller, and P. Michiardi, “Rarest first and choke algorithms are enough,” in *IMC '06: Proceedings of the 6th ACM SIGCOMM on Internet measurement*, New York, NY, USA, 2006, pp. 203–216, ACM Press.
- [6] D. Johnson M. Garey and L. Stockmeyer, “Some simplified np-complete graph problems,” *Theoretical Computer Science*, vol. 1, pp. 237–267, 1976.
- [7] B. W. Kernighan and S. Lin, “An Efficient Heuristic Procedure for Partitioning Graphs,” *The Bell system technical journal*, vol. 49, no. 1, pp. 291–307, 1970.

Time-Ordered Information-Processing on the Binary Tree

Sascha Delitzscher

Department of Physics, Bielefeld University, Germany

Abstract

In this paper we study the effects on information processing on a graph that come in to play, when a certain time ordering is introduced on that graph. If the communication between vertices does not take place simultaneously but consecutively, the set of possible paths on which information may proceed becomes an interesting entity to investigate.

1 Introduction

In the theory of diffusion processes the most interesting questions are that of the existence and configuration of equilibrium states and how these depend on process parameters and network topology. A widely used approach is to model the object which diffuses through a medium (the network) as a contagious infection, i.e. there are some initially infected vertices which have a probability to carry the object over to their environment, following certain rules which depend on the process under consideration.

Usually there is a static network with some vertices (sometimes also called agents) infected at some discrete time t . Then, all infection probabilities get evaluated to determine which vertices are infected at time $t + 1$, and so on. Here, we want to discuss an aspect of such diffusion processes on networks which has not gathered much attention so far. We assume that dissemination does not affect all vertices at the same time. There is a rule that forces vertices to query its neighbors sequentially. The time resolution of these query processes is determined by the maximum degree in the graph. Then, already simple graph structures can lead to a different behavior of diffusion processes.

To illustrate this issue, suppose there are three people, say Alice, Bob, and Christine, who all know each other and want to meet in a face-to-face conversation, but not all three simultaneously. Since each knows exactly two people, they agree that two occasions to meet are sufficient, say in the morning and in the evening. Now one vertex, say Alice, wants to meet Bob in the morning who agrees to do so. Then Christine will meet none in the morning and one of Alice and Bob in the evening. The remaining one will meet none in the evening. If Bob meets Christine, then Alice will have no contact to Christine on this time step at all, although they are neighbors in the network.

The schedule which follows from resolving one time step t into two phases leads to a different infection behavior than the classical setup without time ordering. Assume Alice is contagious with probability p_A of infecting on contact. Then, the probability that Christine gets infected is p_A times the probability that Bob infects her. On the other hand, assume Christine to be contagious with probability p_C , then the probability that Alice gets infected at time t is zero.

As we will work out below, this phenomenon is due to the fact that there is no strict (degree bounded) schedule for the triangle at all.

The example shows that the process topology can change dramatically if scheduling is introduced and whether there is a reasonable schedule or not. In this paper we want to make a first step toward attacking this problem, focusing on graphs which have a degree bound schedule. In detail, we consider a single contagion process on the binary tree, with the root being initially infected and all infection probabilities are chosen to be 1. Thus, we will speak of information processing rather than contagion or infection. In this way only scheduling will influence the process, there are no interference effects that would come from a non trivial diffusion process or a more complex network than the binary tree.

The paper is organized as follows. At first we introduce some notation and the notions of schedules and admissible paths in general graphs. We then turn our view to the scheduled binary tree, deriving a recurrence relation that carries the structural information of the graph and the schedule. Solving this equation leads to formulas for the number of vertices that get infected at time t and other quantities deduced from that.

2 Definitions and Notation

In this section we introduce the notion of a schedule for graph and its effects on the process topology of information flowing through a graph.

Definition 2.1 (schedule) *Let $G = (V, E)$ be a graph, denote by $\Delta(G) = \max_{v \in V} \deg(v)$ the maximum degree of G . A vertex schedule for a vertex $v \in V$ is an enumeration of all edges $e \in E$ incident to v , i.e. an injective mapping $\iota_v : \{e \in E : e \text{ incident to } v\} \rightarrow \{1, 2, \dots, \Delta(G)\} \subset \mathbb{N}$.*

Schedules ι_v, ι_w of adjacent vertices $v, w \in V$ are said to be compatible if $\iota_v((v, w)) = \iota_w((v, w))$.

A graph schedule for a graph G is a family $\{\iota_v\}_{v \in V}$ of schedules, such that any pair of schedules of adjacent vertices is compatible.

Notice that a graph schedule does not need to exist, figure 1 shows an example. It may happen, that regardless of which schedule for a vertex is chosen, there are no compatible schedules on the other vertices.

The name schedule rises from the idea, that communication between adjacent vertices may follow a time ordering, i.e. the communication does not take place simultaneously but consecutively. A vertex v communicates first with the vertex adjacent to $\iota_v^{-1}(1)$, second with the vertex adjacent to $\iota_v^{-1}(2)$ and so on, Here we suppose that $\Delta(G)$ is the time unit in which communication on the graph takes place, and that each single communication between two vertices requires a fraction $1/\Delta(G)$ of the time unit.

We now want to investigate the processing of information on a graph equipped with a graph schedule. A vertex v communicates an information consecutively to its neighbors and according to its schedule. Notice that the other vertices are only able to spread the information *after* they communicated to v or to a vertex, which already has the information. Thus, the time ordering induced by the schedules prohibits some paths to participate in spreading the information. The admissible paths are the following:

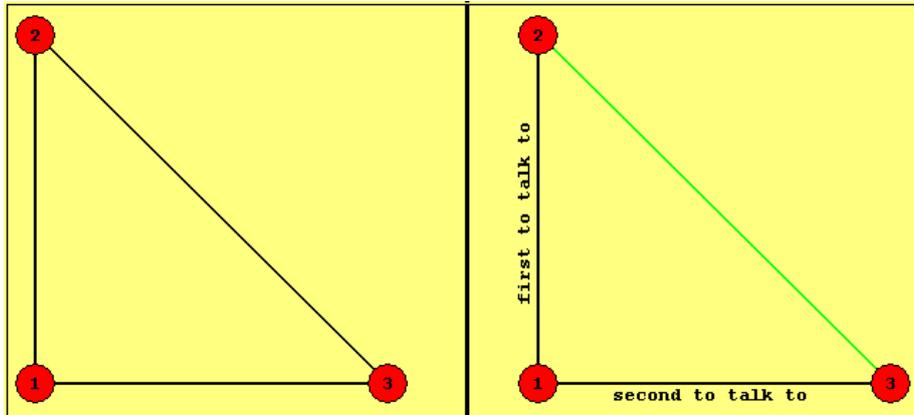


Figure 1: Example of a graph, which admits no graph schedule. If one chooses a schedule for vertex "1", the vertices "2" and "3" can not find a compatible schedule.

Definition 2.2 Let G be a graph equipped with a graph schedule. An admissible path $p = (v_1, v_2, \dots, v_k)$ is a tuple of vertices $v_i \in V, i = 1, \dots, k$ such that $\iota_{v_{j-1}}((v_{j-1}, v_j)) < \iota_{v_j}((v_j, v_{j+1}))$, for all $j = 2, \dots, k - 1$.

One may interpret the edge weights defined by the values of the scheduling functions as a communication potential $U(e)$ of the corresponding edge, and information can only proceed along paths (e_1, \dots, e_k) with positive potential difference at every step, i.e. $\Delta U = U(e_{i+1}) - U(e_i) > 0$. In that case, this difference describes a force affecting the possible paths of information processing.

We now introduce a simple schedule for all rooted regular trees and discuss, how much the choice of a particular schedule affects the process topology on such trees. The *enumeration schedule* is defined inductively by choosing the root to be the first vertex with a schedule and then choosing all other schedules in a way such that they are compatible with the other ones. To do so, let $T = (V, E)$ be a rooted d -regular tree and choose an enumeration for the vertices of T . Then the schedule for the root vertex v_0 is defined by $\iota_{v_0}((v_0, v_k)) = k, k = 1, \dots, d - 1$, since the root has $d - 1$ children and no parent. Schedules for all other vertices $v \in V$ must fulfill the compatibility constraint $\iota_v((v, w)) = \iota_w((w, v)) = l$, for some l , where w denotes the unique parental vertex of v . To all other edges we assign the lowest number in $\{1, \dots, d\} \setminus \{l\}$ to the edge connecting v with the child with the lowest number in the enumeration scheme of the tree, assign the second lowest number to the edge connecting v with the child with the second lowest number in the enumeration scheme, and so on.

The admissible paths are of the form $(v_{i_1}, v_{i_2}, \dots, v_{i_k})$, where $k \leq d + 1$, since no admissible path in T can consist of more than d edges. Figure 2 shows all admissible paths on the rooted binary tree, equipped with the enumeration schedule, the root is interpreted as the source of the information. To point out the influence of the scheduling on the information processing, admissible paths with the root as starting point are drawn red. This can be seen as the paths, through which information can flow in exactly one time step. Figure 3 shows the admissible paths on the second time step, where each vertex that got

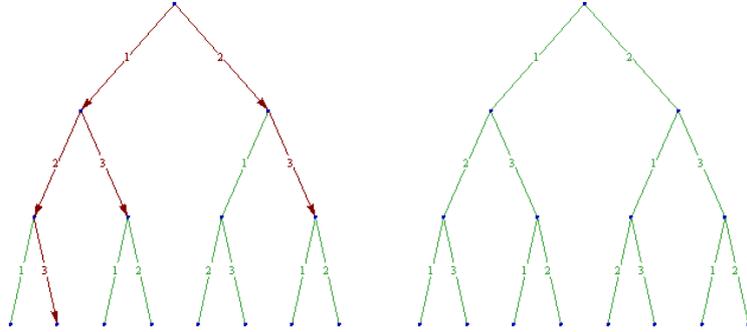


Figure 2: First Step, $t = 1$: The left tree (with the values of the schedules as edge weights) shows how scheduling of communication influences the processing of the communicated information. Only paths with ascending edge weights are admissible, here $(1, 2, 3)$, $(1, 3)$, $(2, 3)$. Of the possible 6 vertices on the first and the second generation of the tree, only 5 get the information spread by the root.

'infected' by the information at the preceding time step is considered as a source of information.

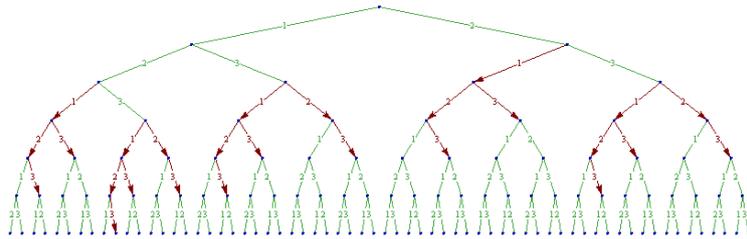


Figure 3: Second Step, $t = 2$: All vertices, which were infected by the root at $t = 1$ are now sources. Admissible paths, which infect new vertices at $t = 2$ are rendered red. Of the possible 62 vertices on the first 5 generations of the tree, only 31 get the information spread by the root at $t = 1$.

Since T is a regular tree, the particular choice of the schedule does not affect the process topology. Every other choice for a schedule leads to the same number of infected vertices on a certain generation g at a time t .

3 Results

In this section we investigate the topology of the information transfer process on a binary tree.

Theorem 3.1 *The number $n(g, t)$ of vertices on generation g that get infected*

at time t is given by $n(g, t) = \frac{1}{g!} \frac{d^g}{dx^g} \Phi(x, t) |_{x=0}$, where $\Phi(x, t)$ is the solution of

$$\begin{aligned}\Phi(x, t) &= (3+x)x^2\Phi(x, t-1) + x^3\Phi(x, t-2) \\ \Phi(x, 1) &= 2x + 3x^2 + x^3 \\ \Phi(x, 2) &= x^2 + 7x^3 + 11x^4 + 6x^5 + x^6.\end{aligned}\tag{1}$$

Proof. The number of vertices which get infected at time t on generation g can be derived with the following consideration: Each vertex can be infected only by another infected vertex on the parental generation over an edge with weight "1", "2" or "3". Counting the frequencies of these edge weights will give us the number of newly infected vertices on generation. Denote by $n_i(g, t)$ the number of active edges at time t with weight " i ", $i = 1, 2, 3$, connecting vertices on generation $g-1$ and g . Thus $n(g, t) = n_1(g, t) + n_2(g, t) + n_3(g, t)$, where the $n_i(g, t)$ fulfill the following recurrence relations

$$\begin{aligned}n_1(g, t) &= n_2(g-1, t-1) + n_3(g-1, t-1) \\ n_2(g, t) &= n_1(g-1, t) + n_3(g-1, t-1) \\ n_3(g, t) &= n_1(g-1, t) + n_2(g-1, t),\end{aligned}$$

with initial conditions

$$\begin{aligned}n_1(1, 1) &= 1 \\ n_1(g > 1, 1) &= 0 \\ n_2(1, 1) &= n_2(2, 1) = 1 \\ n_2(g > 2, 1) &= 0 \\ n_3(1, 1) &= n_3(g > 3, 1) = 0 \\ n_3(2, 1) &= 2 \\ n_3(3, 1) &= 1.\end{aligned}$$

Multiplying with x^g , summing up from 1 to ∞ and setting $\Phi_i(x, t) = \sum_{g=1}^{\infty} n_i(g, t)x^g$ yields the system of coupled difference equations

$$\begin{aligned}\Phi_1(x, t) &= x\Phi_2(x, t-1) + x\Phi_3(x, t-1) \\ \Phi_2(x, t) &= x\Phi_1(x, t) + x\Phi_3(x, t-1) \\ \Phi_3(x, t) &= x\Phi_1(x, t) + x\Phi_2(x, t),\end{aligned}$$

with initial conditions

$$\begin{aligned}\Phi_1(x, 1) &= x \\ \Phi_1(x, 2) &= x^2 + 3x^3 + x^4 \\ \Phi_2(x, 1) &= x + x^2 \\ \Phi_2(x, 2) &= 3x^3 + 4x^4 + x^5 \\ \Phi_3(x, 1) &= 2x^2 + x^3 \\ \Phi_3(x, 2) &= x^3 + 6x^4 + 5x^5 + x^6.\end{aligned}$$

Uncoupling the difference equations for Φ_1 , Φ_2 and Φ_3 yields

$$\Phi_i(x, t) = (3+x)x^2\Phi_i(x, t-1) + x^3\Phi_i(x, t-2),$$

thus the Φ_i all fulfill the same difference equation. So does $\Phi = \Phi_1 + \Phi_2 + \Phi_3$, with initial conditions $\Phi(x, t_0) = \Phi_1(x, t_0) + \Phi_2(x, t_0) + \Phi_3(x, t_0)$.

□

The particular form of the right hand side of the recurrence equation (1) carries the information about the structure of the underlying graph, while the initial conditions characterize the influence of the graph schedule.

Figure 4 shows the behavior of the normalized solution $\rho(g, t) = (1/2^g)n(g, t)$ of equation (1) for different times t , with variable g .

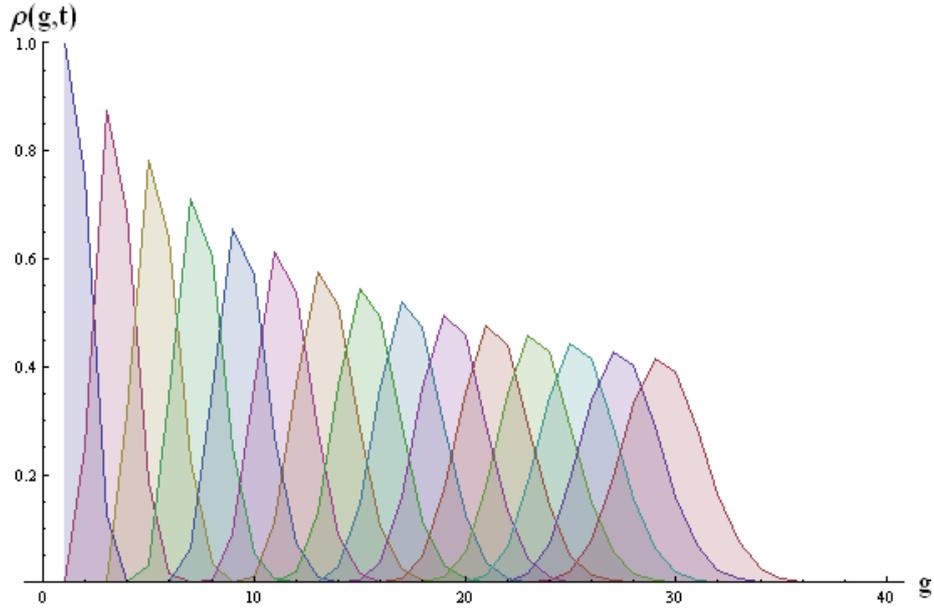


Figure 4: Densities $\rho(g, t)$ of newly infected vertices at time t , depending on generation, in time-ascending order from left ($t = 1$) to right ($t = 15$).

Observe, that with increasing time, the curves become wider but their maximum decreases, indicating that vertices on a growing number of generations are infected at the cost of a smaller efficiency of infecting vertices on a certain generation.

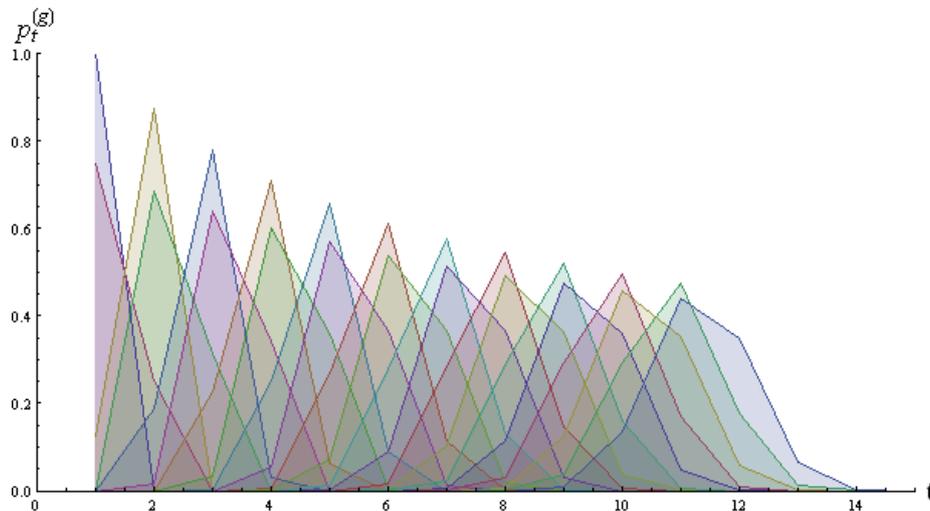


Figure 5: Densities $\rho(g, t)$ of newly infected vertices on generation g , depending on time, in generation-ascending order from left ($g = 1$) to right ($g = 22$).

Figure 5 shows the same function $\rho(g, t) = (1/2^g)n(g, t)$ as figure 4, but now with variable t and one curve for each generation. Observe that the curves behave similar to a dissolving wave packet. Fix, for instance, a generation g and consider the function $p_t^{(g)} = \frac{1}{2^g}n(g, t)$. There is a time $t_*(g)$ at which each vertex on g is infected, so $n(g, t > t_*(g)) = 0$. Thus, $\sum_{t=1}^{\infty} p_t^{(g)} = \sum_{t=1}^{t_*(g)} p_t^{(g)} = 1$. The $p_t^{(g)}$ can be interpreted as the probability that a randomly chosen vertex on generation g gets infected at time t . Computing the entropies $H(\{p_t^{(g)}\}_t)$ of the $p_t^{(g)}$ shows a monotonically increasing behavior of the entropy function as g increases from $g = 5$ to infinity (fig.6).

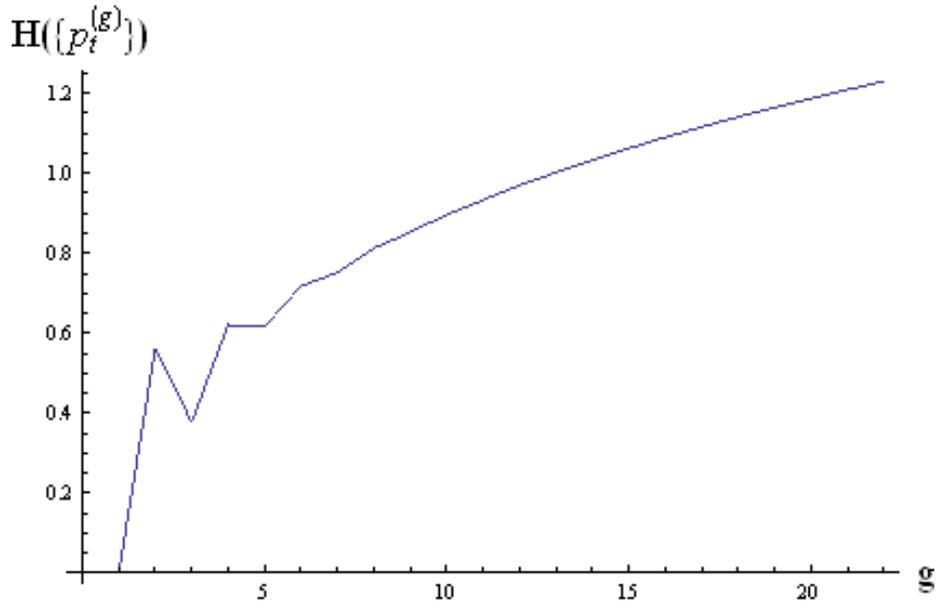


Figure 6: Entropies of the infection-probability distributions of generation g .

We now wish to compute the time $t_*(g)$ at which the whole generation g is infected, i.e. the minimal value for t at which $\sum_{\tau=1}^t n(g, \tau) = 2^g$. Setting $\varphi(x, t) = \sum_{\tau=1}^t \Phi(x, \tau)$ yields the following lemma

Lemma 3.2 *The minimal time $t_*(g)$ at which the whole generation g is infected is given by the minimal solution of*

$$\frac{d^g}{dx^g} \varphi(x, t) \Big|_{x=0} = g!2^g.$$

Proof. The number of vertices on generation g , which get infected at time τ is given by $\frac{1}{g!} \frac{d^g}{dx^g} \Phi(x, \tau) \Big|_{x=0}$. Thus, there must be a minimal time $t_*(g)$ at which the whole generation is infected, i.e. $\sum_{\tau=1}^{t_*(g)} \frac{1}{g!} \frac{d^g}{dx^g} \Phi(x, \tau) \Big|_{x=0} = 2^g$. It follows that $\frac{d^g}{dx^g} \varphi(x, t_*(g)) \Big|_{x=0} = g!2^g$, proving the lemma.

□

Figure 7 shows the behavior of $\rho(g, t) = (1/2^g)n(g, t)$ for some sample curves.

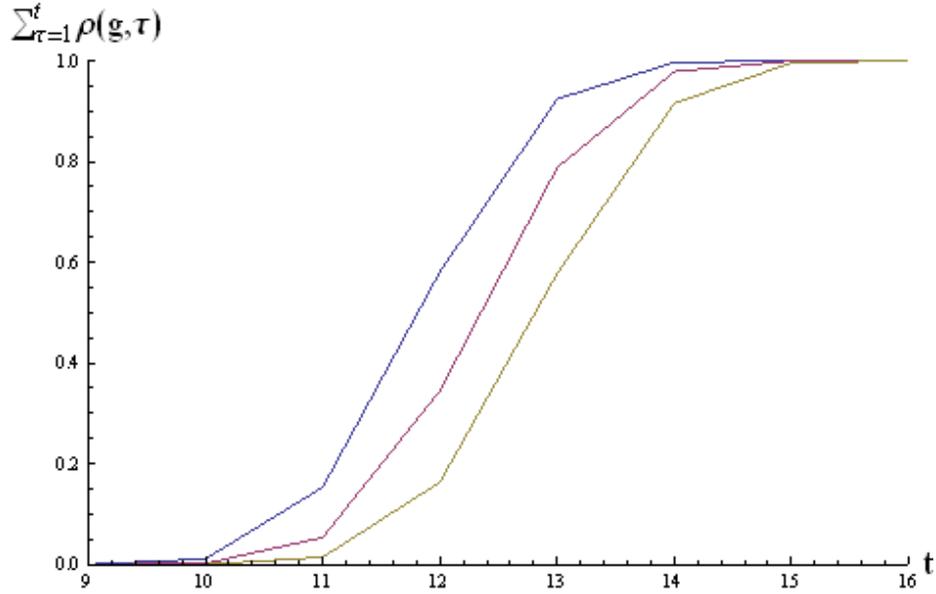


Figure 7: Densities $\rho(g, t)$ of infected vertices on generation $g = 24$ (blue), $g = 25$ (red) and $g = 26$, depending on time.

Lemma 3.3 *The total number $n(t)$ of vertices getting infected at time t is given by*

$$n(t) = \left(1 - \frac{1}{\sqrt{5}}\right) (2 - \sqrt{5})^t + \left(1 + \frac{1}{\sqrt{5}}\right) (2 + \sqrt{5})^t. \quad (2)$$

Proof. The number $n(g, t)$ of newly infected vertices on generation g at time t is given by $n(g, t) = n_1(g, t) + n_2(g, t) + n_3(g, t)$. Thus, the total number $n(t)$ of vertices infected at time t is $n(t) = \sum_{g=1}^{\infty} n(g, t) = \Phi(1, t)$. Using the difference equation (1) yields a recurrence relation for $\Phi(1, t) = n(t)$:

$$n(t) = 4n(t-1) + n(t-2),$$

with initial conditions

$$n(1) = 6, \quad n(2) = 26,$$

and with solution given by equation (2).

□

Summing $n(\tau)$ over all times τ from 1 to a particular time t yields the overall number of infected vertices at time t :

Lemma 3.4 *The overall number $\rho(t)$ of infected vertices at time t , is independent of the scheduling and is given by*

$$\rho(t) = -2 + \left(1 - \frac{2}{\sqrt{5}}\right) (2 - \sqrt{5})^t + \left(1 + \frac{2}{\sqrt{5}}\right) (2 + \sqrt{5})^t.$$

Proof.

$$\rho(t) = \sum_{n=1}^t \left(\left(1 - \frac{1}{\sqrt{5}}\right) (2 - \sqrt{5})^n + \left(1 + \frac{1}{\sqrt{5}}\right) (2 + \sqrt{5})^n \right).$$

Using the geometric sum expression $\sum_{n=0}^t x^n = \frac{1-x^{t+1}}{1-x}$ and simplifying yields the desired result.

□

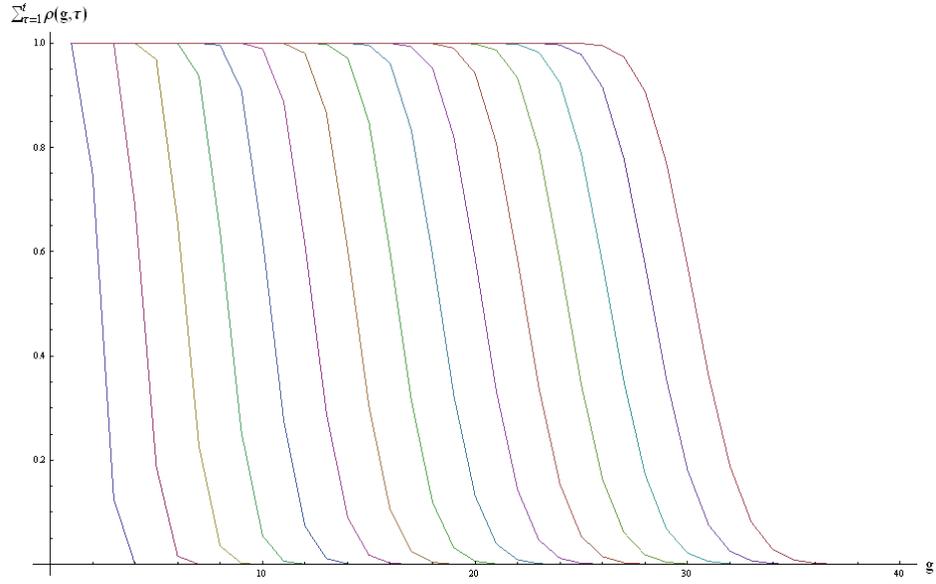


Figure 8: Densities of infected vertices, in time-ascending order from left ($t = 1$) to right ($t = 15$), depending on generation.

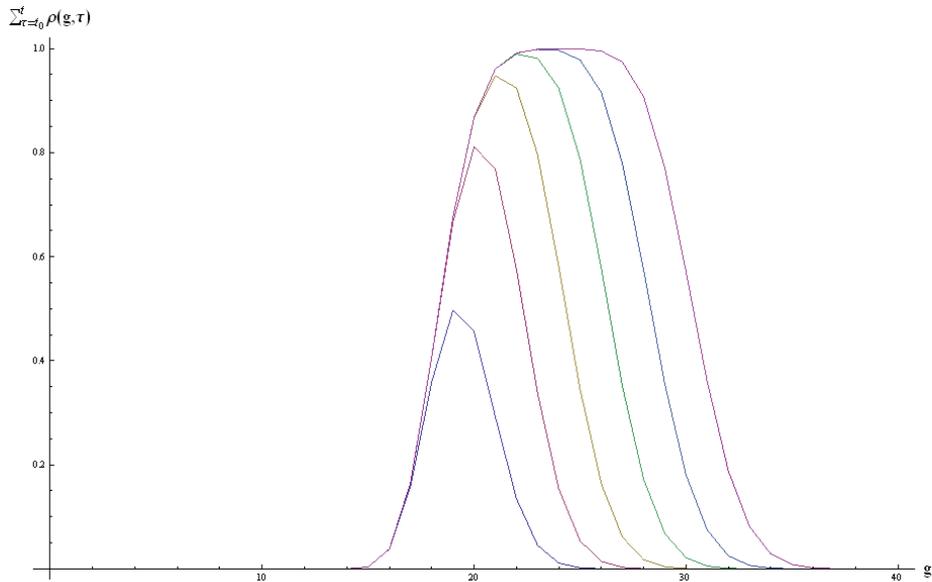


Figure 9: Densities of vertices getting infected between $t = 10$ and $t = 15$, depending on generation.

4 Summary and Perspectives

We have introduced scheduling on graphs and discussed some properties of a scheduled diffusion process on the binary tree. We found a recurrence relation carrying the structural information of the underlying graph, with initial conditions which represent the particular choice of the schedule. These equations translate to general d -ary trees, as well as the qualitative behavior of all graphs shown.

Also, an analysis of arbitrary trees should yield similar results, since trees always allow scheduling, but numerical treatment may be necessary to solve the resulting recursion relations. Nevertheless, on more complex networks including circles or clusters the behavior of scheduled diffusion processes is not clear. The resulting redundancies in communication may lead to an even slower dissemination or no at all.

Finally, it would be interesting to see if graphs which do not allow a degree bound schedule can be embedded in some manner in a more general graph which allows scheduling and so can be treated with the methods presented here.

Generalized Epidemic Processes and Threshold Percolation with Application to Knowledge Diffusion

Phillipe Blanchard, Sascha Delitzscher, Gerald Hiller, Tyll
Krueger, Rainer Siegmund-Schultze

Department of Physics, Bielefeld University, Germany

CITEC Center of Excellence Cognitive Interaction Technology, Bielefeld
University, Universitätsstrae 2123, 33615 Bielefeld Germany

Abstract

In this paper we use a generalized epidemic process to model knowledge diffusion in social networks. We find that, using results from threshold percolation theory, there can happen a phase transition in the number of vertices with a particular knowledge if the number of initially knowing vertices is larger than some critical value. In sociological literature this value is often referred to as critical mass, which now can be explained with the existence of thresholds for each vertex in the network.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 2 |
| 2 | Generalized Epidemic Processes | 3 |
| 2.1 | The Local Process | 4 |
| 2.2 | The Mean Field Process | 4 |
| 2.3 | Other Transitions | 5 |
| 3 | Phase Transitions and Threshold Percolation | 5 |
| 3.1 | The α - Process on $G(n, p)$ | 6 |
| 3.2 | The pure mean field infection process and forgetting | 11 |
| 3.2.1 | The Quadratic Case | 11 |
| 3.2.2 | The Linear Case | 14 |
| 3.3 | The Pure ϵ -process and Forgetting | 15 |
| 4 | Application to Knowledge Diffusion in Complex Networks | 17 |
| 4.1 | Examples of Knowledge Diffusion | 17 |
| 4.2 | Learning, Understanding and Forgetting Knowledge | 18 |
| 5 | Simulation Results | 19 |
| 6 | Summary | 20 |

1 Introduction

In this paper we model knowledge diffusion within a group of individuals or a society. To describe and understand such dynamics is a difficult and challenging task. It is a specific example of a complex system and therefore mathematical models are an appropriate tool to uncover some of the hidden dynamical properties. The society is represented by a graph, where each vertex is an individual and edges between vertices stand for some specific social relationship, such as friendship, consanguinity, sexual intercourse, etc. The choice of the underlying network depends on the choice of the problem under consideration.

By knowledge we mean the class of all facts, information and certainties, one can have a trusted belief in. We assume that representatives of this class, e.g. knowledge about a scientific theory, a childrens book or a TV-Show, spread over the network similar to a disease. There are some initially infected vertices (the ones that already have a certain knowledge), which at every time step have a small chance to infect their neighbors with the knowledge. A remarkable property of knowledge is, that, given a particular knowledge, once an individual has heard of it sufficiently often, the probability of getting infected jumps to a much higher value due to curiosity or persuasion. Each vertex has a so called *threshold* in the number of infected neighbors, which, as soon as it gets reached, assures that the vertex gets almost surely infected, too.

In mathematics, special cases of such processes like threshold percolation have gained much attention lately, for instance Janson et al. discussed threshold percolation on the Erdős-Renyi random graph, [17]. They found out that there is a phase transition in the density of initially infected vertices and described this effect in great detail. In this paper, besides modeling knowledge diffusion, we use another, more intuitive but less elaborate approach to obtain the same result.

Other work on bootstrap percolation has been done on the grid $[n]^d$, see [1], [2] (for $d = 3$), [18] ($d = 2$), [20], and on \mathbb{Z}^2 , see [9], [10], and on trees, see [3], [5], [6], [13], [14]. Bootstrap percolation on complex networks with given degree sequence has been analyzed in [4], with arbitrary degree distributions in [15], and clique percolation on $G(n, p)$ in [7].

Since we want to apply the model to knowledge diffusion on complex networks, we add the possibility of forgetting knowledge, i.e. there is a probability of a transition from an infected state to the susceptible ground state. Bootstrap percolation with the possibility of vacating sites has been treated on \mathbb{Z}^2 in [11] and on \mathbb{Z}^d in [12].

The notion of a threshold is well known in social sciences. Granovetter, [16], used thresholds to discuss decision making on the complete graph, where he adopted ideas from Schelling's model of residential segregation, [22].

In addition to face-to-face transmission we include the transmission of knowledge via public sources like mass media, books or journals. Since the likeliness of something to appear in such public sources is often proportional to the prevalence of the knowing individuals in the society we model this type of infection path using an infection probability mean field term depending on the infection density but not on the network structure.

It turns out that for both types of stochastic processes there are phase transitions of mainly two types. The first relates to parameters describing the transmission probabilities as well as to parameters describing the network structure

(primarily the edge density). Below some critical values any initial infection will not be able to spread over the network at all. This phase transition resembles the ones known from percolation theory. The second type of phase transition is of a more dynamical nature and unknown from classical epidemic processes. In this case even in the supercritical parameter domain where epidemic growth is possible, one still needs to start above a critical initial density of infected individuals (depending of course on the parameters chosen) to reach high prevalence.

Since knowledge transfer and exchange has so many faces and each of these different aspects would perhaps require a specific model of its own, we want to focus in this paper mainly on the local threshold dynamics, the mean field dynamics and the interplay of distinct levels of depth or understanding of a certain type of knowledge, which we classify as passive and active knowledge here.

Unfortunately, good data about knowledge transmission are notoriously difficult to obtain and even for simple "fact - knowledge" essentially not available, we will concentrate primarily on questions of qualitative nature. The various simulation results discussed in section 5 should also be interpreted in this spirit. The main purpose of the simulations is the illustration of the possible qualitative scenarios, with reasonably chosen simulation parameters.

Similar to thresholds, the existence of a critical value for the initial infection density is a well known phenomenon in social sciences, where it is often referred to as critical mass. In [21], the author defines critical mass as "*the point at which enough individuals have adopted an innovation so that the innovations further rate of adoption becomes self-sustaining [...] Until a critical mass occurs at a relatively early stage in the diffusion process, the rate of adoption is slow. After a critical mass is achieved, the rate of adoption accelerates.*" In conclusion, if we interpret this acceleration as the phase transition, this strongly indicates the existence of thresholds, yielding critical values for the initial infection density.

The paper is organized as follows. We begin by introducing the dynamical processes via their transition probabilities, the following chapter 3 is devoted to threshold percolation, where we calculate the phase transitions of the pure threshold process, discuss percolation for the mean field process with forgetting, and find a criticality condition for the classical infection process with forgetting. Then we turn our view to knowledge diffusion in chapter 4 and discuss in chapter 5 some simulations to illustrate how the model works. With the help of the simulation, we are able to find all k -cores of any network, exemplarily done for a real world network, the social network of all students of Bielefeld.

2 Generalized Epidemic Processes

In the following we introduce a certain type of generalized epidemic process on complex networks, where each node can be either susceptible or in one of two distinct infection states. The probabilities of infection are determined by a local threshold process, describing the correlations of linked vertices, and a mean field process, describing the influence of infected vertices on the network independently of the underlying network structure.

Let $G = (V, E)$ be a network with $|V| = n$ vertices. Each $v \in V$ has a number $s_t(v) \in \{0, 1, 2\}$, which is called the state of v at time t . These may

be different levels of evolutionary states of one entity, e.g. knowledge, where a higher level indicates higher knowledge or capabilities of using knowledge. Since we use epidemic processes and its terminology, vertices in the zero state are called *susceptible*, while vertices in non-zero states are called *infected*.

First we discuss the probabilities $P_{0 \rightarrow 1}(v)$ of a vertex v to switch from state 0 to 1. Since there are three distinct states, there are six transitions possible, but not all of them must be considered to be physically meaningful. To keep the interpretation of the states to be subsequent evolutionary states, the transition "0 \rightarrow 2" for instance makes no sense as one can not simply skip one state. Thus we set $P_{0 \rightarrow 2} = 0$. Further on we assume that the transition "1 \rightarrow 2" takes place at a fixed rate ρ .

2.1 The Local Process

The local process is designed as a classical epidemic process with a threshold. Each vertex v has a threshold $\Delta_v \geq 0$. One reasonable choice for the distribution of Δ_v is a Poisson distribution with mean λ , which will differ from network to network. Of course, in some cases it may be useful to choose a completely different distribution for Δ_v , but we will not consider such distributions here.

Denote the number of neighbors of v in state i at time t by $N_{i,t}(v)$, and let $N_t(v) = N_{1,t}(v) + N_{2,t}(v)$ be the number of infected neighbors. We need to consider two cases, infection below the threshold and infection above. To discuss the first case, let $N_t(v) < \Delta_v$. Assuming that for each vertex in state "i", $i = 1, 2$, the probability of infecting a neighbor is ϵ_i and that both infection processes are independent of each other yields a probability of switching state from "0" to "1" of

$$P_{0 \rightarrow 1}^\epsilon(v) = 1 - (1 - \epsilon_1)^{N_{1,t}(v)}(1 - \epsilon_2)^{N_{2,t}(v)}.$$

If $N_t(v) \geq \Delta(v)$, i.e. if the number of infected neighbors exceeds the threshold, the contribution of single probabilities gets replaced by a probability $\alpha > \epsilon$,

$$P_{0 \rightarrow 1}^\alpha(v) = \alpha.$$

The idea behind that is, that a vertex above the threshold is willing to switch its state, due to curiosity or persuasion but still has some issues, for instance switching costs too much money, too much time, something depending on the specific knowledge diffusion under investigation, and any more contact with knowing vertices will not change its probability to switch anymore. The value of α is then just a measure for these issues. Notice that for most applications it will be sufficient simply to choose $\alpha = 1$, since smaller values for α only result in rescaling of time.

2.2 The Mean Field Process

In addition to local infection processes, we add the possibility of knowledge to be transmitted via processes independent of the network structure. The infection probabilities defined by such a process will depend on the density of infected vertices, since these vertices are the ones, which draw attention. The influence of mass media and other global effects is modeled by a mean-field process with parameters β_i and a function $f : [0, 1] \rightarrow [0, 1]$, which takes the prevalence

densities of vertices in state "i" at time t as arguments. Denote by x_t the density of vertices in state "1", y_t the density of vertices in state "2". We set

$$\begin{aligned} P_{0 \rightarrow 1}^{\beta_1}(v) &= \beta_1 f(x_t) \\ P_{0 \rightarrow 1}^{\beta_2}(v) &= \beta_2 f(y_t). \end{aligned}$$

Again we assume both processes to be independent, thus

$$P_{0 \rightarrow 1}^{\beta}(v) = \min \{ \beta_1 f(x_t) + \beta_2 f(y_t), 1 \}.$$

Denote by $\Theta(x)$ the Heaviside function. From the considerations above it follows that the total probability $P(v)$ (we drop subscripts for the ease of reading) of a vertex v with threshold Δ and number N of infected neighbors to advance from "0" to "1" is

$$\begin{aligned} P(v) &= 1 - (1 - P^\epsilon(v)\Theta(\Delta - N))(1 - P^\alpha(v)\Theta(N - \Delta))(1 - P^\beta(v)) \\ &\approx P^\Delta(v) + P^\beta(v), \end{aligned}$$

where we set $P^\Delta(v) = P^\epsilon(v)\Theta(\Delta - N) + P^\alpha(v)\Theta(N - \Delta)$ and neglected all products of probabilities.

2.3 Other Transitions

Any infected vertex has a certain probability to switch back to the zero level, being susceptible again. We assume that this happens with a constant rate, thus

$$P_{i \rightarrow 0}(v) = \gamma_i, \quad i = 1, 2,$$

for all v .

One could choose a more complex process here, for instance add an explicit dependence on the number of infected neighbors, but since the local infection process is depending on the prevalence, there is an implicit dependence via the reinfection process.

The transition "1 \rightarrow 2" corresponding to the higher infection level is assumed to take place at a fixed rate ρ , so

$$P_{1 \rightarrow 2}(v) = \rho,$$

for all v .

3 Phase Transitions and Threshold Percolation

In this section we develop a new approach to investigate threshold percolation of the processes involved in our model. At first we study percolation of the pure α -process on the Erdős-Renyi random graph $G(n, p)$, $p = c/n$, finding a critical density for the phase transition in the density of the initially infected vertices a_0 as a function of the Erdős-Renyi parameter c .

Afterwards we discuss two cases of the pure mean-field process with forgetting. One where the mean-field term couples quadratically to the total prevalence, the other where the coupling is linear, finding a phase transition in a_0 only in the first case.

Finally we calculate the branching number of the pure ϵ -process with forgetting on a supercritical branching tree.

3.1 The α - Process on $G(n, p)$

In the following we want to give some analytic estimations for the spread of the pure α - process on random graphs of Erdős-Renyi type and the involved phase transitions. We deal with uniform constant local threshold Δ . Below the threshold we assume the transmission probability ϵ to be zero. Furthermore we look for a limiting case of SI-epidemics namely that the transmission probability above the critical threshold Δ equals 1.

We will deal here only with the recursion for conditioned expected values. Except for the behavior near the phase transitions the iterated expectation value is for large n in good agreement with the numerical simulations (due to the small variance of the involved stochastic variables in that case). The graph model we use is $G(n, p)$ with $p = \frac{c}{n}$ being the probability of an edge between any two vertices. Furthermore all edges are independent.

As usual $E(X)$ denotes the expectation of the random variable X .

Theorem 1 ([17]) *Let $G = G(n, p)$ be an Erdős-Renyi random graph on n vertices with $p = c/n$. Then, for the expected density $s_i = S_i/n$ of infected vertices of G at time i holds*

$$\begin{aligned} s_{i+1} &= 1 - (1 - a_0)e^{-cs_i} \sum_{k=0}^{\Delta-1} \frac{(cs_i)^k}{k!} \\ &= a_0 + (1 - a_0)e^{-cs_i} \sum_{k=\Delta}^{\infty} \frac{(cs_i)^k}{k!}, \end{aligned}$$

for $n \rightarrow \infty$.

We do the explicit computations here only in the case $\Delta = 3$ since this represents already the general case.

If M is a subset of the vertex set with cardinality $|M|$, denote by $N(x, M)$ the number of edges a vertex x has into M . In the Poisson limit $n \rightarrow \infty$ we have $\Pr(N(x, M) = k) = e^{-p|M|} p^k |M|^k / k!$.

Let \mathfrak{A}_t be the set of newly infected vertices at time t , define $A_t := E(|\mathfrak{A}_t|)$ and $a_t := A_t/n$. Further, define $\Sigma_t := \bigcup_{\tau=0}^t \mathfrak{A}_\tau$, the set of infected vertices at time t , with $S_t := E(|\Sigma_t|)$ and $s_t = S_t/n$.

Let $\Xi_{t,l}$ be the set of susceptible vertices at time t which have exactly l edges into Σ_{t-1} , and the set of susceptible vertices at time t be $\Xi_t := \Xi_{t,0} + \Xi_{t,1} + \Xi_{t,2}$, since $\Delta = 3$.

Then, for the expected number X_t of vertices in Ξ_t , conditioned on A_{t-1} , we can find and solve a recursion relation in the following way. At first, we consider the $X_{t,l}, l = 0, 1, 2$. The expected number $X_{t+1,0}$ of susceptible vertices at time $t+1$ with no edges into the set Σ_t of infected vertices at time t is equal to X_t

minus the average number of vertices in $X_{t,0}$, which have had edges into the set of newly infected vertices at time t :

$$\begin{aligned} X_{t+1,0} &= X_{t,0} - X_{t,0} \Pr(N(x, \mathfrak{A}_t) \geq 1) \\ &= X_{t,0} \Pr(N(x, \mathfrak{A}_t) = 0) \\ &= X_{t,0} e^{-ca_t} \end{aligned}$$

In the same manner, recursion relations for $X_{t,1}$ and $X_{t,2}$ can be obtained:

$$\begin{aligned} X_{t+1,1} &= X_{t,1} - X_{t,1} \Pr(N(x, \mathfrak{A}_t) \geq 1) + X_{t,0} \Pr(N(x, \mathfrak{A}_t) = 1) \\ &= X_{t,1} e^{-ca_t} + X_{t,0} ca_t e^{-ca_t}, \end{aligned}$$

$$\begin{aligned} X_{t+1,2} &= X_{t,2} - X_{t,2} \Pr(N(x, \mathfrak{A}_t) \geq 1) \\ &\quad + X_{t,1} \Pr(N(x, \mathfrak{A}_t) = 1) \\ &\quad + X_{t,0} \Pr(N(x, \mathfrak{A}_t) = 2) \\ &= X_{t,2} e^{-ca_t} + X_{t,1} ca_t e^{-ca_t} + X_{t,0} c^2 a_t^2 e^{-ca_t} \end{aligned}$$

Using the definition of X_t , inserting the recursions and writing $X_t^* = X_{t,0} + X_{t,1}$, yields

$$\begin{aligned} X_{t+1} &= X_{t+1,0} + X_{t+1,1} + X_{t+1,2} \\ &= X_t e^{-ca_t} + X_t^* ca_t e^{-ca_t} + \frac{1}{2} X_{t,0} c^2 a_t^2 e^{-ca_t} \end{aligned} \quad (1)$$

Inserting the recursion for X_t again yields

$$\begin{aligned} X_{t+1} &= X_{t-1} e^{-c(a_t + a_{t-1})} + X_{t-1}^* c(a_t + a_{t-1}) e^{-c(a_t + a_{t-1})} \\ &\quad + \frac{1}{2} X_{t-1,0} c^2 (a_t + a_{t-1})^2 e^{-c(a_t + a_{t-1})} \end{aligned} \quad (2)$$

From (1) and (2) we conclude

$$\begin{aligned} X_{t+1} &= X_0 e^{-c \sum_{\tau=0}^t a_\tau} + X_0^* c \left(\sum_{\tau=0}^t a_\tau \right) e^{-c \sum_{\tau=0}^t a_\tau} \\ &\quad + X_{0,0} c^2 \left(\sum_{\tau=0}^t a_\tau \right)^2 e^{-c \sum_{\tau=0}^t a_\tau}. \end{aligned}$$

Using $s_t = \sum_{\tau=0}^t a_\tau$ and initial conditions $X_0^* = X_{0,0} = X_0 = n - A_0$ gives

$$X_{t+1} = (n - A_0)(1 + cs_t + c^2 s_t^2) e^{-cs_t},$$

which implies

$$1 - s_{t+1} = (1 - a_0)(1 + cs_t + c^2 s_t^2)e^{-cs_t}.$$

The formula has an intuitive interpretation (although it is not so clear how to make this intuition into a short rigorous proof). Namely $n - S_{i+1}$ (the non infected) are the vertices from $n - A_0$ (the initially non infected set) which have less than Δ edges into s_i .

Finally we give the formula for the half infected at time i (that is the vertices which have at least 1 but less than Δ edges into s_{i-i}).

In general we have for the different types of half-infected

$$\frac{1}{n}X_{i+1,k} = (1 - a_0)e^{-cs_i} \frac{(cs_i)^k}{k!},$$

which can be derived in the following way. Let $X_i = X_{i,0} + X_{i,1} + \dots + X_{i,\Delta-1}$. The set of half infected is $H_i := X_i^* - X_{i,0}$. From the above derivation it follows that

$$\begin{aligned} \frac{1}{n}H_{i+1} &:= \frac{1}{n}(X_{i+1} - X_{i+1,0}) \\ &= (1 - a_0)e^{-cs_i} \sum_{l=0}^{\Delta-1} \frac{(cs_i)^l}{l!} - (1 - a_0)e^{-cs_i} \\ &= (1 - a_0)e^{-cs_i} \sum_{l=1}^{\Delta-1} \frac{(cs_i)^l}{l!}. \end{aligned}$$

Note that the above formulas remain valid also in the case $\Delta = 1$ where the fixed point equation for $A_0 = o(n)$ just provides the classical result for the size of the giant component in the Erdős-Renyi model.

We turn now to the discussion of the case $\Delta = 2$, the simplest nontrivial case. From the above formulas we have the following fixed point equation

$$s = f(s) = 1 - (1 - a_0)e^{-cs}(1 + cs). \quad (3)$$

Note that there can be several solutions. One has to take the dynamically stable one (closest to a_0). A closer examination shows that, depending on the value of a_0 and c , there are either 3 fixed points or just one in the domain $[0, 1]$, see fig. 1

Since the iteration mapping has no critical points in this interval the smallest of those fixed point is also the attractor for the orbit starting with s_0 . An example for the values: $a_0 = 0.08$, $c = 3.2$ is shown in figure 2.

The phase transition happens when the first two fixed points (in case there are three) join together and form an indifferent (slope one) fixed point. Near that value one would observe a long transient behavior till the orbit reaches its final destination the high prevalence fixed point.

Next we compute the critical density a_0^c for the phase transition in a_0 as a function of c .

Theorem 2 *Let $G = G(n, p)$ be an Erdős-Renyi random graph on n vertices with $p = c/n$. Then, the critical density is given by*

$$a_0^c = 1 - \frac{2e^{(-\frac{1}{2} + \frac{1}{2}c - \frac{1}{2}\sqrt{c^2 - 3 - 2c})}}{c(-1 + c - \sqrt{c^2 - 3 - 2c})}$$

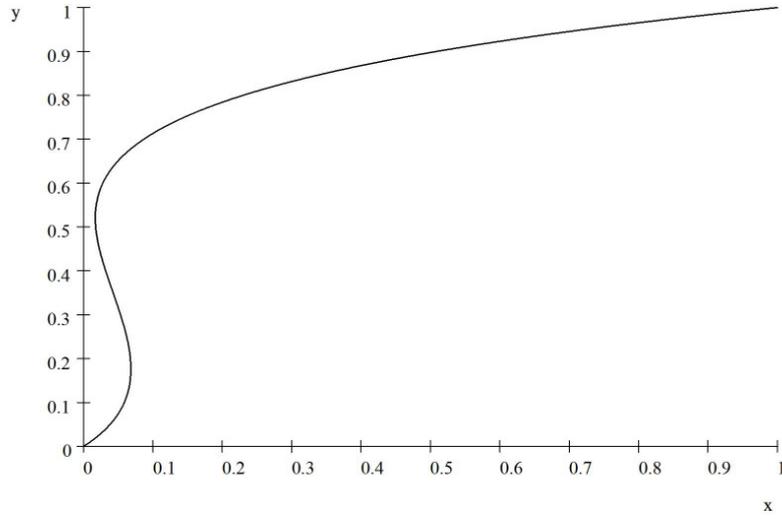


Figure 1: Plot of the solutions of equation 3, depending on $x = a_0$, for $c = 3.3$.

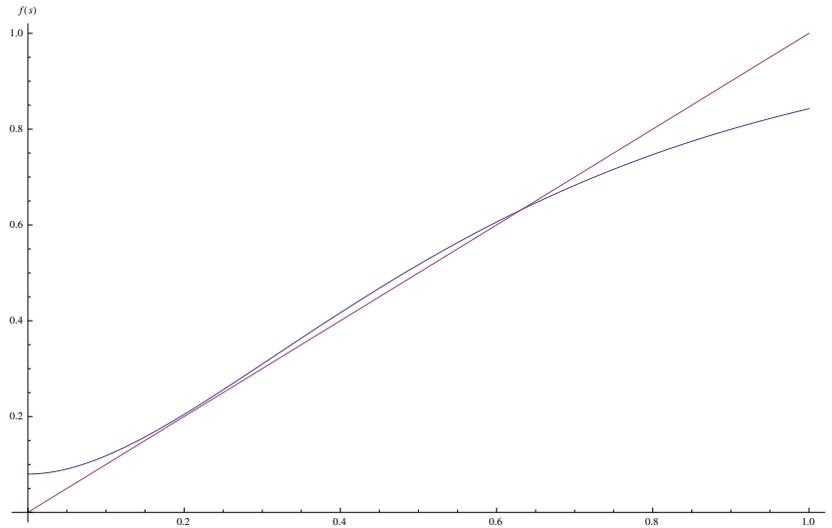


Figure 2: Example of fixed point equation with $a_0 = 0.08$, $c = 3.2$

Proof.

Define $a := a_0$. The condition for criticality is that $f'(x) = 1$ at the fixed point x , where

$$f(x) = 1 - (1 - a)(1 + cx)e^{-cx}.$$

We have

$$\frac{d(1 - (1 - a)(1 + cx)e^{-cx})}{dx} = c^2xe^{-cx} - ac^2xe^{-cx}$$

and

hence $c^2xe^{-cx} - ac^2xe^{-cx} = 1$. The solution is

$$\begin{aligned} a &= -\frac{1}{c^2xe^{-cx}}(-c^2xe^{-cx} + 1) \\ &= 1 - \frac{1}{c^2x}e^{cx}. \end{aligned}$$

Inserting into the fixed point equation gives

$$x = 1 - \left(1 - \left(-\frac{1}{c^2xe^{-cx}}(-c^2xe^{-cx} + 1)\right)\right)(1 + cx)e^{-cx},$$

which simplifies to

$$x = 1 - \left(1 - \left(-\frac{1}{c^2xe^{-cx}}(-c^2xe^{-cx} + 1)\right)\right)(1 + cx)e^{-cx} = \frac{1}{c^2x}(c^2x - cx - 1)$$

and solving

$$x = \frac{1}{c^2x}(c^2x - cx - 1)$$

for x gives

$$x = \frac{1}{c^2} \left(-\frac{1}{2}c + \frac{1}{2}c^2 - \frac{1}{2}\sqrt{-3c^2 - 2c^3 + c^4} \right)$$

for the fixed point. Inserting this back into the formula for a gives:

$$\begin{aligned} a &= 1 - \frac{1}{c^2x}e^{cx} \\ &= 1 - \frac{e^{c\frac{1}{c^2}\left(-\frac{1}{2}c + \frac{1}{2}c^2 - \frac{1}{2}\sqrt{-3c^2 - 2c^3 + c^4}\right)}}{c^2\frac{1}{c^2}\left(-\frac{1}{2}c + \frac{1}{2}c^2 - \frac{1}{2}\sqrt{-3c^2 - 2c^3 + c^4}\right)} \\ &= 1 - \frac{e^{\frac{1}{c}\left(-\frac{1}{2}c + \frac{1}{2}c^2 - \frac{1}{2}\sqrt{-3c^2 - 2c^3 + c^4}\right)}}{\left(-\frac{1}{2}c + \frac{1}{2}c^2 - \frac{1}{2}\sqrt{-3c^2 - 2c^3 + c^4}\right)} \\ &= 1 - \frac{2e^{\left(-\frac{1}{2} + \frac{1}{2}c - \frac{1}{2}\sqrt{c^2 - 3 - 2c}\right)}}{c(-1 + c - \sqrt{c^2 - 3 - 2c})} \end{aligned}$$

■

Since there is no real solution for the fixed point x for $c < 3$ one needs $c \geq 3$ to get a phase transition in a_0 . Figure 3 shows a plot of the critical value a_0^c as a function of c .

A few notes are in order. The phenomenon of a phase transition in the initial density is not specific for $G(n, p)$. It holds in all random graph spaces where the typical graphs have an almost-tree structure in the local neighborhood of a vertex. Usually this is the case even up to neighbors at distance $h \dim(G)$, where $h < \frac{1}{2}$ is a graph dependent constant. The same is still true for graphs which arise as projections of bipartite networks, for which again the local tree property holds (see also the next section).

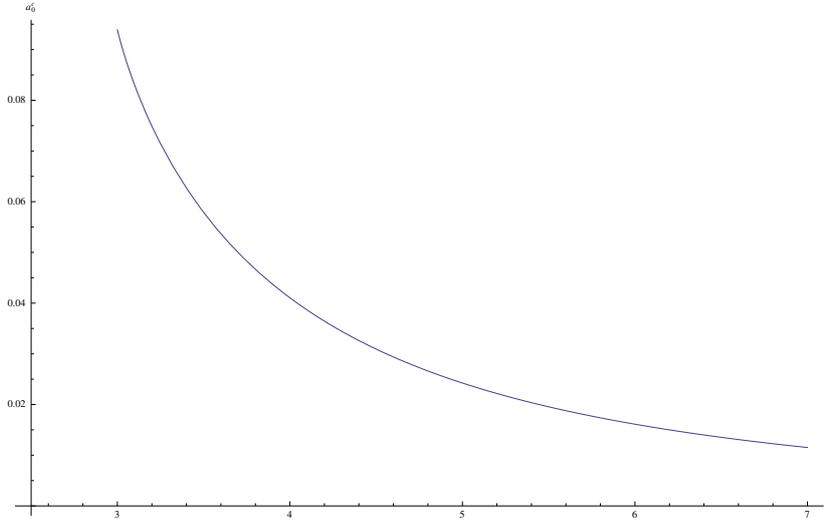


Figure 3: a_0^c as a function of c

3.2 The pure mean field infection process and forgetting

In the following we study the pure mean field process. Again we will observe that there are phase transitions with respect to the initial density a_0 . Let $f(x)$ be in the following either a linear or a quadratic function. The involved recursion for the mean field process including forgetting is very simple. Note that in each time step we first do the infection process and than the forgetting.

3.2.1 The Quadratic Case

We start the discussion of the recursion relations with the case $f(x) = x^2$.

$$\begin{aligned}
 s_{i+1} &= (1 - \gamma)(s_i + (1 - s_i)\beta f(s_i)) \\
 &= (1 - \gamma)(s_i + (1 - s_i)\beta s_i^2) \\
 &= (1 - \gamma)(s_i + \beta s_i^2 - \beta s_i^3).
 \end{aligned}$$

This gives the fixed point equation

$$\begin{aligned}
 s &= (1 - \gamma)(s + \beta s^2 - \beta s^3) \\
 \implies 1 &= (1 - \gamma)(1 + \beta s - \beta s^2), \text{ for } s \neq 0
 \end{aligned}$$

Define $a := \frac{1}{1-\gamma}$ and $b := \frac{a-1}{\beta} = \frac{\frac{1}{1-\gamma}-1}{\beta} = \frac{\gamma}{(1-\gamma)\beta}$ we have as fixed point equation $s^2 - s + b = 0$ with the solutions

$$s^{(1)} = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4b}$$

and

$$s^{(2)} = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4b}.$$

Note that for $b > \frac{1}{4}$ there is no solution except the value $s = 0$. Since the derivative at zero equals $1 - \gamma < 1$ we have zero as a local attractive fixed point. For $b < \frac{1}{4}$ the fixed point at $s^{(1)}$ is a repeller and is also the boundary for the two basins of attraction, with zero as the one attractor and $s^{(2)}$ as the other attractor. In the case of $b = \frac{1}{4}$ one has again that zero is a global attractor and $s = 1/2$ is an unstable fixed point.

We consider now the situation where the local α -process induces a set of infected individuals, which do not become extinct even in the presence of a small forgetting rate γ . They are immune to forgetting. Let therefore a_0 be the density of an immune infected set of vertices. The above equations change in the following way:

$$\begin{aligned} s_{i+1} &= a_0 + (1 - \gamma)((s_i - a_0) + (1 - s_i)\beta f(s_i)) \\ &= a_0 + (1 - \gamma)((s_i - a_0) + (1 - s_i)\beta s_i^2) \\ &= (1 - \gamma)(s_i + \beta s_i^2 - \beta s_i^3) + \gamma a_0, \end{aligned}$$

and for the fixed point equation one gets (with $d = ba_0$)

$$\begin{aligned} s &= (1 - \gamma)(s + \beta s^2 - \beta s^3) + \gamma a_0 \\ \Rightarrow \gamma s &= \beta(1 - \gamma)(s^2 - s^3) + \gamma a_0 \\ \Rightarrow bs &= s^2 - s^3 + ba_0 \Rightarrow \\ s^3 - s^2 + bs - d &= 0, d < b. \end{aligned}$$

Since $s_i \geq a_0$ for $\beta > 0$ we have $s_{i+1} < s_i$ and for $s_i \geq 1$ we have $s_{i+1} = (1 - \gamma)s_i + \gamma a_0 - \epsilon$ for some $\epsilon > 0$ (and consequently $s_{i+1} = s_i - \gamma(s_i - a_0) - \epsilon < s_i$) it follows that either 1) there is one stable fixed point in the interval $[a_0, 1]$ or 2) there are three fixed points, where the middle one is unstable and the boundary of the basins of attraction for the two other ones or 3) there are two fixed points where the smaller one is indifferent (derivative = 1) and the larger one is stable. The critical case (one indifferent fixed point is given by the condition "fixed point + derivative = 1") which gives

$$\begin{aligned} 2s\beta - \gamma - 2s\beta\gamma - 3s^2\beta + 3s^2\beta\gamma + 1 &= 1 \\ \Rightarrow 2s\beta - \gamma - 2s\beta\gamma - 3s^2\beta + 3s^2\beta\gamma &= 0 \\ \Rightarrow 2s\beta(1 - \gamma) - 3s^2\beta(1 - \gamma) &= \gamma \\ \Rightarrow 2s - 3s^2 &= b \\ \Rightarrow s^3 + \frac{b}{3}s - \frac{2}{3}s^2 &= 0 \end{aligned}$$

and the fixed point equation

$$s^3 - s^2 + bs - d = 0$$

subtraction gives (the derivative is given by $2s\beta(1 - \gamma) - 3s^2\beta(1 - \gamma) + 1 - \gamma$, which implies $\beta(1 - \gamma)(2s - 3s^2 + \frac{1}{\beta}) > 0$ for $s \in [0, 1]$)

$$\begin{aligned} \frac{1}{3}s^2 - \frac{2sb}{3} + d &= 0 \\ \Rightarrow s^2 - 2bs + d &= 0, \end{aligned}$$

$s^2 - 2bs + d = 0$, Solution is: $b - \sqrt{b^2 - d}, b + \sqrt{b^2 - d}$. It follows that there is no solution for $d > b^2 \Rightarrow a_0 > b = \frac{\gamma}{(1-\gamma)\beta}$ (there is only one, large attractor). Furthermore we have by the first equation $2s - 3s^2 = b$, with the solutions $\frac{1}{3} - \frac{1}{3}\sqrt{1-3b}, \frac{1}{3}\sqrt{1-3b} + \frac{1}{3}$ and hence for $b > \frac{1}{3}$ there is no derivative, and hence only one fixed point. By topological reasons the criterion that the upper one of the fixed points becomes the attractor is given by the following equation:

$$\begin{aligned} b - \sqrt{b^2 - d} &= -\frac{1}{3}\sqrt{1-3b} + \frac{1}{3} \\ \Rightarrow \frac{1}{3}\sqrt{1-3b} + b &= \sqrt{b^2 - d} + \frac{1}{3} \\ \Rightarrow \sqrt{1-3b} &= 3\sqrt{b^2 - d} + 1 - 3b \\ \Rightarrow b + 3d - 6b^2 &= 2(1-3b)\sqrt{b^2 - d} \\ \Rightarrow (b(1-6b) + 3d)^2 &= 4(1-3b)^2(b^2 - d). \end{aligned}$$

Solutions are $b - \frac{2}{9}\sqrt{27b^2 - 9b - 27b^3 + 1} - \frac{2}{9}, b + \frac{2}{9}\sqrt{27b^2 - 9b - 27b^3 + 1} - \frac{2}{9}$ and by topological reasons the second one is the relevant solution. Figure 4 shows a plot of the critical density a_0^c as a function of $b = \frac{\gamma}{(1-\gamma)\beta}$

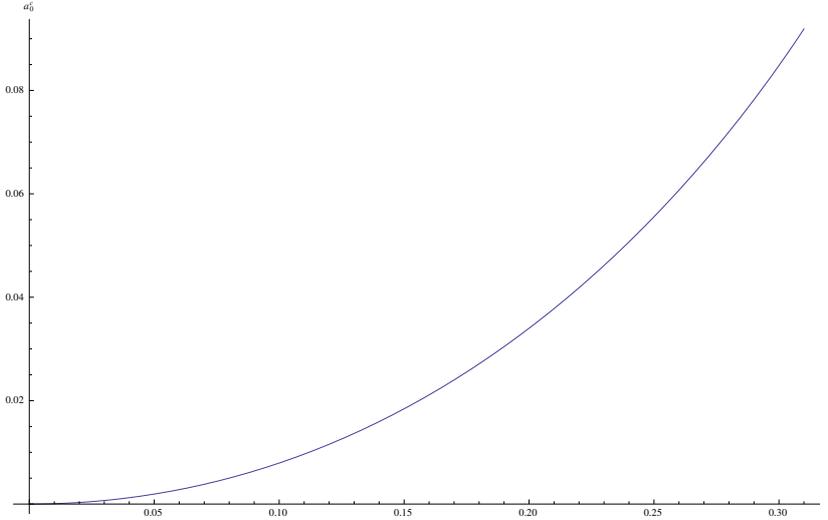


Figure 4: The critical density a_0^c as a function of $b = \frac{\gamma}{(1-\gamma)\beta}$

Figure 5 shows a plot of the critical density a_0^c as a function of β (the lower curve) and a plot of the critical density to be in the upper basin of attraction in case there is no immune (to forgetting) set. In both plots the value $\gamma = 0.01$ was used

Figure (5) illustrates how the presence of an immune set lowers the threshold value. In combination with the α -process this means that starting with an a_0 below the critical values for both (the pure α -process and the pure mean field process) may yield an over-critical process.

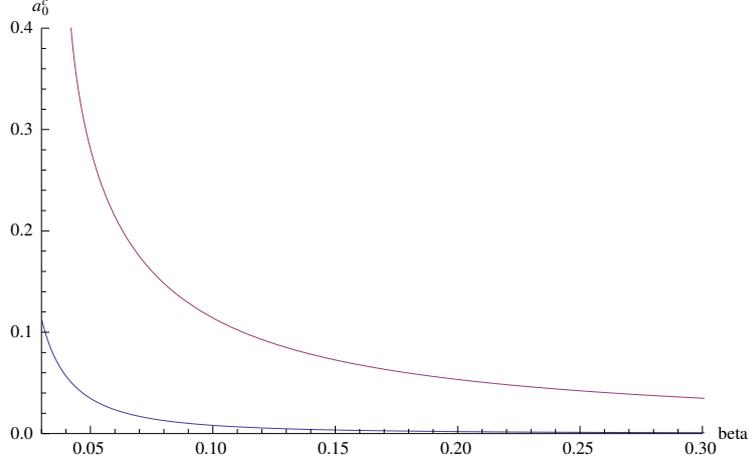


Figure 5: Critical density in case of an immune-to-forgetting set (lower curve) and without an immune set (upper curve) as a function of β for $\gamma = 0.01$

3.2.2 The Linear Case

We now discuss the linear case $f(x) = x$:

$$\begin{aligned} s_{i+1} &= (1 - \gamma)(s_i + (1 - s_i)\beta s_i) \\ &= (1 - \gamma)(s_i(1 + \beta) - \beta s_i^2) \end{aligned}$$

which gives for the fixed point equation

$$s = (1 - \gamma)(s(1 + \beta) - \beta s^2)$$

and the solutions $s = 0$ and $s = 1 - b$. Evaluating the derivate

$$\frac{d((1 - \gamma)(s(1 + \beta) - \beta s^2))}{ds} = \beta - \gamma - 2s\beta - \beta\gamma + 2s\beta\gamma + 1$$

at zero yields $1 + \beta(1 - \gamma) - \gamma = 1 - \gamma(1 + \beta) + \beta$. The critical case is therefore $\beta(1 - \gamma) - \gamma = 0 \Rightarrow 1 = \frac{\gamma}{\beta(1 - \gamma)} = b$ and for $\beta(1 - \gamma) - \gamma > 0 \Leftrightarrow 1 > b$, the fixed point at $1 - b$ becomes a global attractor and for $b \geq 1$ the zero solution becomes globally attractive. Note that in contrast to the quadratic case for f there is only one attractor and therefore no phase transition in the initial density.

We discuss now the case when there is a repeller of density a_0 . The corresponding formulas are

$$\begin{aligned} s_{i+1} &= a_0 + (1 - \gamma)((s_i - a_0) + (1 - s_i)\beta f(s_i)) \\ &= a_0 + (1 - \gamma)((s_i - a_0) + (1 - s_i)\beta s_i) \\ &= (1 - \gamma)(s_i + \beta s_i - \beta s_i^2 - a_0) + a_0 \\ &= (1 - \gamma)(s_i + \beta s_i - \beta s_i^2) + \gamma a_0, \end{aligned}$$

and for the fixed points (with $d = \beta a_0$ and $b = \frac{\gamma}{\beta(1 - \gamma)}$)

$$\begin{aligned}
s &= (1 - \gamma)(s + \beta s - \beta s^2) + \gamma a_0 \Rightarrow \\
\gamma s &= \beta(1 - \gamma)(s - s^2) + \gamma a_0 \Rightarrow \\
bs &= s - s^2 + ba_0 \Rightarrow \\
s^2 - s + bs - d &= 0, d < b,
\end{aligned}$$

which has as solutions:
 $\frac{1-b}{2} \pm \frac{1}{2}\sqrt{4d + (1-b)^2}$. Only the positive solution $\frac{1-b}{2} + \frac{1}{2}\sqrt{4d + (1-b)^2}$ is relevant.

Figure 6 shows a plot for the fixed point in case $\gamma = 0.01$ and $a_0 = 0.05$ in comparison with the upper fixed point in the case without an immune infected set

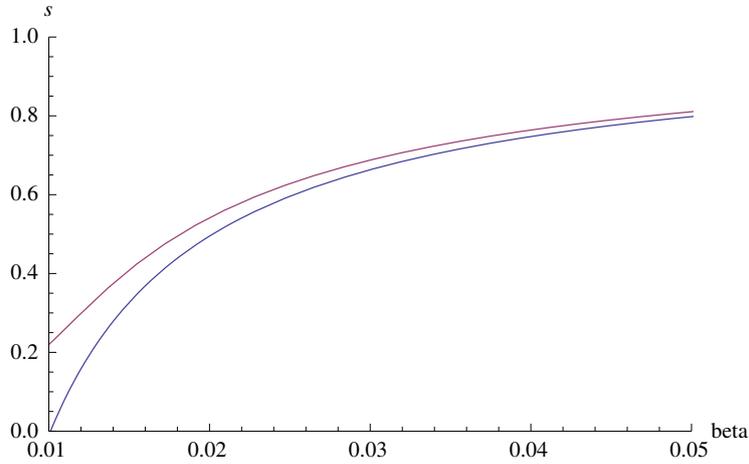


Figure 6: Comparison of fixed point in case of an immune set (upper curve) and without immune set (lower curve).

Note that in the case without an immune set, for $\beta < \frac{\gamma}{1-\gamma}$ the zero solution is a global attractor. In summary one sees that the effect of an immune set in the linear case of f is less dramatic than in the quadratic case, except for very small β - values.

3.3 The Pure ϵ -process and Forgetting

We now want to discuss 2-type threshold percolation of the pure ϵ -process, i.e. each vertex v can be in three states $s(v) \in \{0, 1, 2\}$ and we set $\Delta_v \geq n$ for each $v \in V$, so no vertex can succeed the threshold. We assume that vertices in state "i", $i = 1, 2$ have a probability of ϵ_i of infecting susceptible neighbors, i.e. to force a neighbor switch its state from "0" to "1".

Infected vertices have a probability of γ_i , $i = 1, 2$, to switch back to the susceptible state "0". Also we assume that vertices in state "1" switch to state "2" at a fixed rate ρ .

In the following we consider the limit supercritical branching tree model with branching number c , conditioned to the branching process not dying out. We

consider a given infection dynamics starting with one infected particle as root of the tree.

Theorem 3 *The expected number of vertices E , which an initially infected vertex $v \in V$ infects, is given by*

$$E = c \left(1 - \frac{\gamma_1}{q + (1-q)\epsilon_1} - \frac{\gamma_2}{\gamma_2 + (1-\gamma_2)\epsilon_2} \right),$$

where $q = \gamma_1 + (1-\gamma_1)\rho$ is the probability of leaving state "1".

Proof. Denote with t_1 and t_2 the time an infected vertex spends in state "1" or "2", respectively. The probability $P(k, l_1, t_1)$ that a vertex in state "1" for a time of t_1 steps infects l_1 descendants out of k is a binomial distribution $P(k, l_1, p(t_1)) = \binom{k}{l_1} p(t_1)^{l_1} (1-p(t_1))^{k-l_1}$ with $p(t_1) = 1 - (1-\epsilon_1)^{t_1}$ and therefore given by

$$P(k, t_1, l_1) = \binom{k}{l_1} (1 - (1-\epsilon_1)^{t_1})^{l_1} (1 - \epsilon_1)^{t_1(k-l_1)}.$$

The probability that an infected vertex stays in state "1" for exactly t_1 time steps is given by

$$[(1-\gamma_1)(1-\rho)]^{t_1} q,$$

where we set $q = \gamma_1 + (1-\gamma_1)\rho$. In a $G(n, p)$ with $p = c/n$, for large n the number of possible descendants is Poisson distributed with mean c . Thus, the expected number E_1 of infected descendants a vertex generates while in state "1" is given by

$$E_1 = \sum_{k=0}^{\infty} \frac{c^k}{k!} e^{-c} \sum_{t_1=0}^{\infty} [(1-\gamma_1)(1-\rho)]^{t_1} q \sum_{l_1=0}^k l_1 \binom{k}{l_1} (1 - (1-\epsilon_1)^{t_1})^{l_1} (1-\epsilon_1)^{t_1(k-l_1)}. \quad (4)$$

Using that $\langle l \rangle = \sum_l l P(k, l, p(t_1)) = k p(t_1)$ for binomial distributions, and the geometric series expression $\sum_t x^t = \frac{1}{1-x}$ yields

$$\begin{aligned} E_1 &= \sum_{k=0}^{\infty} \frac{c^k}{k!} e^{-c} k \left(1 - \frac{q}{q + (1-q)\epsilon_1} \right) \\ &= c \left(1 - \frac{q}{q + (1-q)\epsilon_1} \right). \end{aligned}$$

If $\rho = 0$ then $E_2 = 0$, otherwise we get for E_2 the following

$$\begin{aligned} E_2 &= \sum_{k=0}^{\infty} \frac{c^k}{k!} e^{-c} \sum_{t_1=0}^{\infty} [(1-\gamma_1)(1-\rho)]^{t_1} (\rho - \gamma_1 \rho) \sum_{l_1=0}^k \binom{k}{l_1} (1 - (1-\epsilon_1)^{t_1})^{l_1} (1-\epsilon_1)^{t_1(k-l_1)} \times \\ &\times \sum_{t_2} (1-\gamma_2)^{t_2} \gamma_2 \sum_{l_2=0}^{k-l_1} l_2 \binom{k-l_1}{l_2} (1 - (1-\epsilon_2)^{t_2})^{l_2} (1-\epsilon_2)^{t_2(k-l_1-l_2)} \\ &= \frac{(q-\gamma_1)c}{q + (1-q)\epsilon_1} \left(1 - \frac{\gamma_2}{\gamma_2 + (1-\gamma_2)\epsilon_2} \right) \end{aligned}$$

Thus,

$$E = c \left(1 - \frac{\gamma_1}{q + (1-q)\epsilon_1} - \frac{\gamma_2}{\gamma_2 + (1-\gamma_2)\epsilon_2} \right).$$

■

Since we did not consider reinfection processes in this calculation, the actual expected number of infected vertices in our model is larger than E . Thus, the condition $E > 1$ ensures over-criticality of the infection process.

4 Application to Knowledge Diffusion in Complex Networks

In this section we want to apply the generalized epidemic process to knowledge diffusion in complex networks. Since we are only interested in how knowledge disseminates through a network, we do not need to specify what knowledge exactly is or how it emerges. Here we rely on the fact, that there is a common sense about what knowledge is and stick to this.

To discuss knowledge diffusion, we interpret the three states "0", "1", "2" as distinct levels of the capability of a vertex to use knowledge. We say that a vertex in state "0" has no knowledge, it is not aware of the knowledge but as a susceptible vertex there is a probability of switching to state "1". We then say that this vertex has *passive knowledge*, what may be for instance being aware of the knowledge and having a rather crude understanding. Vertices in state "2" are said to have *active knowledge*, what may be a deep understanding of the knowledge and the capability to develop it further and take part in further production of new knowledge.

To illustrate the meaning of passive and active knowledge we give some examples first.

4.1 Examples of Knowledge Diffusion

For each example we want to specify, what is meant by passive knowledge and by active knowledge.

- **The spread of popular media titles**

An example of knowledge diffusion on complex networks is the spread of popular media titles, like books, movies or music. The underlying network is the social network of people using the media. A person has passive knowledge, if it is aware of the existence of the media, i.e. it knows that a certain book or movie exists (e.g. Harry Potter) and is curious about it, but did not read or watch it for several reasons, yet. Active knowledge then is the consumption of that media and talking about it in public.

- **Trends: Nordic walking**

Nordic Walking is a popular example for a trend on a social network, what can be described by the knowledge diffusion model as well. Having active knowledge means performing Nordic Walking in public, having passive knowledge means that one has recognized Nordic Walking and has a positive attitude about it. The positive attitude is important, since this is a

necessary condition to switch to active knowledge. So, a person with passive knowledge is willing to perform Nordic Walking but has some issues, like a lack of time or equipment.

- **The usage of certain hardware**

The knowledge diffusion model can describe the spread of certain hardware like MacBooks or iPhones on a network. Active knowledge means using that hardware and showing and telling it to friends. Passive knowledge is, similar to the Nordic Walking trend, the recognition and positive attitude about the hardware.

- **The usage of certain software**

As in the case of hardware, one can describe the spreading of certain software products, like Internet browsers or different media players, in a network. Having passive knowledge means having the software installed and used at least once and keeping it on the computer, having active knowledge means using that certain software as standard (browser or player), using other competing software only occasionally.

- **The usage of scientific products**

With scientific products we mean special science related products like a new alternative model for an observed effect or scientific software such as SKIN, which is an agent based simulation software. Having passive knowledge means knowing the product and using it occasionally, active knowledge means using the product and perhaps developing it further.

Many more examples may be found, and with slight changes in the infection rules one may also describe the spread of corruption, terroristic attitudes or trust.

4.2 Learning, Understanding and Forgetting Knowledge

Acquiring passive knowledge is what we call learning here. Assuming $\epsilon_1 = \epsilon_2$ yields an infection probability of the local process of

$$P_{0 \rightarrow 1}^\epsilon(v) = 1 - (1 - \epsilon)^{N_t(v)},$$

with $N_t(v) = N_{1,t} + N_{2,t}$, the number of knowing neighbors. Each knowing neighbor of v contributes an amount of ϵ to the probability of infection, by talking about it.

The mean-field process, as defined in chapter 2, is given by

$$P_{0 \rightarrow 1}^{\beta_1}(v) = \beta_1 f(b_{1,t}),$$

where $b_{1,t}$ is the density of vertices with passive knowledge, and by

$$P_{0 \rightarrow 1}^{\beta_2}(v) = \beta_2 f(b_{2,t}),$$

where $b_{2,t}$ is the density of vertices with active knowledge.

The differentiation between active and passive knowledge and their influence on the mass media reflects the different abilities of knowing vertices to spread knowledge via mass media. For instance, if the knowledge under consideration is a physical theory, actively knowing vertices, namely scientists, publish papers

much more and with a higher impact than passively knowing vertices (students or other non-professionals).

Acquiring active knowledge is what we call understanding here. The transition takes place at a fixed rate, $P_{1 \rightarrow 2}(v) = r$, so at every time step a fraction of r vertices with passive knowledge will gain active knowledge.

Also forgetting, i.e. the transition "1,2 \rightarrow 0" back to zero, happens at a fixed rate, $P_{1,2 \rightarrow 0}(v) = \gamma$, so at each time step a fraction of γ vertices with either active or passive knowledge will return back in the state of unawareness.

5 Simulation Results

All simulations were done on a $G(n, p)$, with $p = c/n$. After specifying the average initial infection s_0 , initially infected vertices were chosen at random, with a probability of s_0 each. The rate γ of forgetting was $\gamma = 0.01$ in all simulations.

We simulated the pure α -process, as an example for knowledge diffusion, where knowledge gets transmitted mainly by local processes (examples are jokes or Tupperware), and without active knowledge. We compare the results with a setup where vertices may gain active knowledge and turn the mean-field on for these.

Also we used the model to find the k -cores of a real world network with k as threshold, then infecting all vertices and performing threshold percolation on the network with forgetting turned on.

To investigate the interplay of the various infection processes, we simulated an supercritical pure α -process (with $\alpha = 1$), i.e. setting $\epsilon = 0$, $\beta = 0$, and with no active knowledge. Figure 7 shows that including vertices with active knowledge and activating the mean-field process for vertices with active knowledge (i.e. $\beta_1 = 0, \beta_2 > 0$) may lift a sub-critical α -process to an overcritical process in total.

Figure 7(a) shows the behavior of the number of vertices with (passive) knowledge for the pure α -process on a $G(n, c/n)$ with $n = 100'000$ and $c = 3.5$. One sees that the initial infection of $s_0 = 0.06$ suffices to infect many new vertices until a saturation is reached. Choosing $s_0 = 0.05$ yields a sub-critical pure α -process, the density of infected vertices does not change at all, but including active knowledge and a mean-field process for vertices with active knowledge helps the process to become overcritical.

Further on, we used the simulation program to find the size of the k -core of the StudiVz network of students of Bielefeld, see [19] on more details about this real world network. To find the k -cores, we chose an initial infection $s_0 = 1$, i.e. all vertices are infected, and ran the simulation only with the α -process ($\alpha = 1$) and forgetting turned on. Setting the threshold $\Delta = k$ for all vertices results in a final state of the network, in which all vertices in the k -core remain infected due to reinfection via the α -process, while all other vertices forget.

Other simulation results include the observation that Poisson distributed thresholds, i.e. $P(\Delta(v) = k) = e^{-\lambda} \lambda^k / k!$, accelerate the infection process in contrast to uniformly distributed thresholds with $\Delta(v) = \lambda$ for all v . This is due to the fact that on average a fraction $nP(\Delta(v) < \lambda)$ of all n vertices has a threshold below λ , which get infected very soon, helping to raise the number of infected vertices above the critical density.

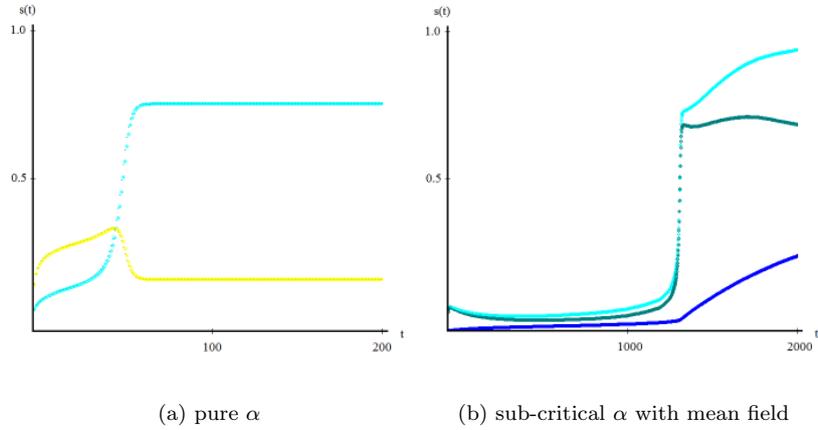


Figure 7: The α -process on $G(n, c/n)$ with $n = 100'000$, $c = 3.5$ and $\Delta = 2$ for all vertices. 7(a) shows the pure α -process for an initial infection $s_0 = 0.06$ ($a_0^c \approx 0.058$). The turquoise colored line is the total density of infected vertices, yellow shows the density of semi-infected vertices. 7(b) shows a sub-critical α -process, $s_0 = 0.05$, with an activated mean-field process for active vertices, $\beta_2 = 0.2$. The rate ρ of production of vertices with active knowledge is $\rho = 0.001$. The turquoise line is the total density of infected vertices, blue the density of vertices with passive knowledge, green the density of vertices with active knowledge.

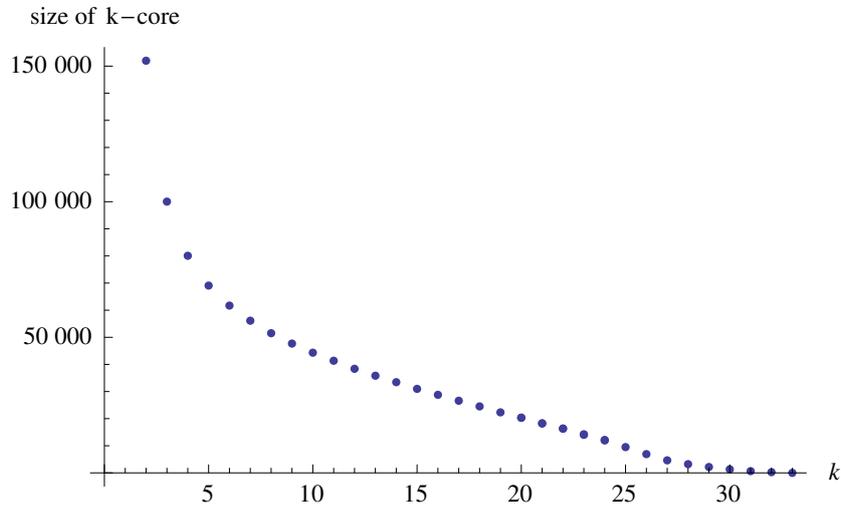


Figure 8: Size of the k -cores of the StudiVz network

6 Summary

We used threshold percolation to investigate a generalized epidemic process. This process, defined as the independent sum of a local threshold infection

process, a mean field process, and forgetting, shows two types of phase transition. One related to parameters describing the transmission probabilities and to parameters describing the network structure, the other one (unknown from classical epidemics) related to the initial infection density. The results about threshold percolation may be generalized to hold for inhomogeneous random graphs, discussed in depth in [8].

We applied the model to knowledge diffusion on complex networks, using two distinct levels of infection to distinguish capabilities of using knowledge, namely active and passive knowledge. We ran simulations to investigate the interplay between vertices with active and passive knowledge. After the phase transition in the number of vertices with passive knowledge, the number of vertices with active knowledge increases rapidly as well (showing no phase transition but a non differentiability in one point.)

Finally we used the simulation program to find the k -cores of a real world network, the StudiVz network (see [19]) by giving all vertices a threshold $\Delta = k$, setting the infection probability above the threshold to one, i.e. $\alpha = 1$, and infecting the whole network initially. Forgetting then yields that all vertices but the ones in the k -core forget.

References

- [1] J. Balogh, B. Bollobas, and R. Morris, *Bootstrap Percolation in High Dimensions*, arXiv:0907.3097v3 [math.PR], 2010
- [2] J. Balogh, B. Bollobas, and R. Morris, *Bootstrap percolation in three dimensions*, Annals of Probability, 2009, Vol. 37, No. 4, 1329-1380
- [3] J. Balogh, Y. Peres, G. Pete, *Bootstrap Percolation on Infinite Trees and Non-Amenable Groups*, Combinatorics, Probability and Computing, Vol 15, 715-730, 2005
- [4] G.J. Baxter, S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes, *Bootstrap Percolation on Complex Networks*, arXiv:1003.5583v2 [cond-mat.stat-mech], 2010
- [5] M. Biskup, R. H. Schonmann, *Metastable Behavior for Bootstrap Percolation on Regular Trees*, Journal of Statistical Physics (2009) 136:667-676, September 10, 2009
- [6] Ph. Blanchard, A. Krueger, T. Krueger, P. Martin, *The Epidemics of Corruption*, arXiv:physics/0505031v1 [physics.soc-ph], 2005
- [7] B. Bollobas, O. Riordan, *Clique percolation*, Random Structures and Algorithms Volume 35 Issue 3, 294 - 322, 2008
- [8] B. Bollobas, S. Janson, O. Riordan, *The phase transition in inhomogeneous random graphs*, 2006, arXiv:math/0504589v3 [math.PR]
- [9] K. Bringmann, K. Mahlborg, *Improved Bounds on Metastability Thresholds and Probabilities for Generalized Bootstrap Percolation*, arXiv:1001.1977v2 [math.PR], 2010

- [10] F. Camia, *Scaling Limit and Critical Exponents for Two-Dimensional Bootstrap Percolation*, Journal of Statistical Physics, Volume 118, Numbers 1-2 / January 2005, 85-101
- [11] H. Chen, *On the Stability of a Population Growth Model with Sexual Reproduction on Z^2* , Ann. Probab. Volume 20, Number 1 (1992), 232-285
- [12] H. Chen, "On the Stability of a Population Growth Model with Sexual Reproduction on $Z^d, d \geq 2$ ", Ann. Probab. Volume 22, Number 3 (1994), 1195-1226.
- [13] L.R.G. Fontes, R.H. Schonmann, "Bootstrap Percolation on Homogeneous Trees has 2 Phase Transitions", Journal of Statistical Physics, Volume 132, Number 5 / September 2008, 839-861
- [14] L.R.G. Fontes, R.H. Schonmann, "Threshold $\theta \geq 2$ -contact processes on homogeneous trees", arXiv:math/0603109v1 [math.PR], 2006
- [15] A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes, *k-core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects*, Phys. Rev. E 73, 056101 (2006)
- [16] M. Granovetter, "Threshold Models of Collective Behavior", The American Journal of Sociology, Vol. 83, No. 6 (May 1978), 1420-1443
- [17] S. Janson, T. Luczak, T. Turova, T. Vallier, "Bootstrap percolation on the random graph $G(n,p)$ ", arXiv:1012.3535v1
- [18] J. Gravner, A. E. Holroyd, "Slow convergence in bootstrap percolation", Ann. Appl. Probab. Volume 18, Number 3 (2008), 909-928
- [19] C. Lee, T. Scherngell, M. J. Barber, "Real-world separation effects in an online social network", 2009, arXiv:0911.1229v1 [physics.soc-ph]
- [20] T. Mountford, R. H. Schonmann, *The survival of large dimensional threshold contact processes*, Ann. Probab. Volume 37, Number 4 (2009), 1483-1501.
- [21] E. M. Rogers, "Diffusion of Innovation", Free Press, ISBN 0612628434, (1962)
- [22] T. C. Schelling, "Dynamic models of segregation", The Journal of Mathematical Sociology, Volume 1, Issue 2, 143-186, 1971

Passive Supporters of Terrorism and Phase Transitions

Friedrich August, Philippe Blanchard, Sascha Delitzscher, Gerald Hiller, Tyll Krüger
Universität Bielefeld, D-33615 Bielefeld, Germany

Abstract. We discuss some social contagion processes to describe the formation and spread of radical opinions. We use threshold dynamics to describe the local spread of opinions, and mean field effects. We calculate and observe phase transitions in the dynamical variables resulting in a rapidly increasing number of passive supporters. This strongly indicates that military solutions are inappropriate.

Keywords. Terrorism, Passive Supporters, Phase Transitions

1. Introduction

In this article we discuss some social contagion processes which may play an important role in the dynamics of radical opinion formation. The prime applications in mind are conflict situations as they are met at the time of writing in Afghanistan, Iraq or Palestine where a highly armed alliance of foreign troops fights against a part of the local population which has been radicalized in a way such that western social classification dubs them terrorists. In the following we will use the word "terrorist" solely to refer to a certain subpopulation (whose interaction features will be described below) in our model environment and do not intend to enter the difficult debate of what constitutes the social essence of terrorism. In terms of our model interaction and shortly speaking one could say that terrorists are those individuals on which counter terrorist throw bombs on.

Besides the radical terrorist groups there is a much larger grey-area of supporters of these terrorists. Support has many faces, ranging from just keeping still about what one knows about terrorists locations or movements up to supporting terrorists by providing various forms of logistic infrastructure. Again, we will avoid specifying exactly what is meant with this notion but use it to describe the potential, predecessor states of an opinion state from which radical groups may (by whatever means and tools) recruit new members.

In this paper we discuss some aspects of the dynamics of passive supporters of terrorist activities in virtual social networks. Our main interest lies in the study of phase transitions in the number of passive supporters induced by what is euphemistically called collateral damage as is common as a consequence of counter terrorist attacks on terrorists moving around in populated places. Phase transitions in the opinion of large parts of a population are particularly important since they violate the classical "linear" action-reaction view common among military leaders and politicians.

We are not concerned with real terrorist networks and their dynamics, for further information on this topic see [1]. Our model is based on one paradigm. Counter terrorist

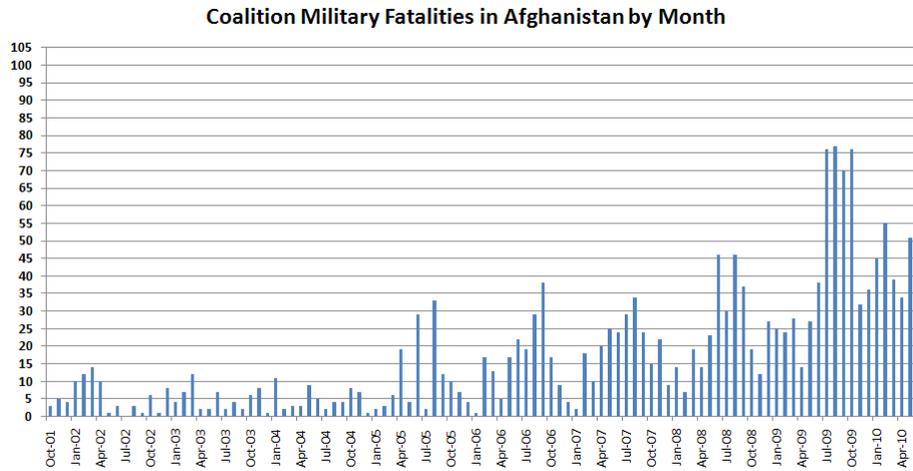


Figure 1. Coalition military casualties in Afghanistan by month: Neglecting seasonal effects, the sudden increase in the number of casualties is very similar to the density of active terrorists in figures 3, 4 and 5, caused by the phase transition in the density of passive supporters. Source of diagram: <http://www.icasualties.org/oef/>

strikes lead to collateral damage. In many cases terrorists use civilians as human shields, and civilian casualties in turn are likely to cause an increase in the number of passive supporters¹ and increase the willingness of civilians to become members of radical groups. As discussed by Galam in [2], the number of passive supporters is correlated to the physical mobility of active terrorists.

Cause and effect are not linearly coupled, there is a phase transition instead. It was widely believed among western observers that in Afghanistan a large part of the population was supporting the Allied forces at the beginning of the operation, in spite of many casualties caused by bombings and other military strikes. Then a sudden change in the public opinion occurred, the atmosphere inclined to the disadvantage of the allied forces in a short period of time, a phase transition from Allied-friendly to Taliban-friendly took place, causing a boost in the number of passive supporters.

Passive supporters usually do not reveal their nature towards outsiders and this phase transition happens hidden from Allied forces. The number of passive supporters can only be measured indirectly via the degree of cooperation of the civil population. After the phase transition the counter terrorists gain virtually no help from civilians and the recruiting pool for radical organizations becomes nearly inexhaustible, as it has been popularized in News reports about the Gaza Strip. Such a situation may result in an absurd and tragic solution to secure public safety and to avoid a downward spiral of violence like the segregation of people from each other with a fence.

2. Description of the Model

In the following we will specify the structure of our model which is inspired by generalized epidemic processes, (for more general information about this class of processes

¹at least in the local family and friendship network of a civilian victim this is highly plausible

and some other applications see [3]). Let $G = (V, E)$ be a graph on a finite set of vertices V and a set of edges E between vertices. We start with the description of the state space. To keep things simple we distinguish only four states $\{0, 1, 2, 3\}$. State 0 encodes the susceptible population which is more or less neutral in their opinion about terrorists. State 1 individuals correspond to the passive supporters of terrorism and state 2 encodes active terrorists. State 3 vertices correspond to vertices, which are isolated from the network and do not interact anymore with other vertices. Although it would be natural, to consider more refined states we can restrict the discussion of phase transitions to this rather simple setting by arguing that the type of phase transition described here is robust to many refinements of the model.

There is usually an immune subpopulation which is resistant to any radical opinion influence. Including this group will reduce the size of the susceptible population but does not influence the existence of phase transitions. Also there usually exist dynamics within the terrorist groups in complex hierarchical networks which in turn may have a strong influence on terrorists activities (see [1] for a recent model concerning this question). These intrinsic network dynamics within the relatively small subpopulation of terrorists have only marginal influence on the 0 – 1 transition process (the process of becoming a passive supporter via a social contagion process).

The dynamics will be defined via transition probabilities P_{loc} and P_{mean} which mainly depend on the state distribution in the neighborhood of the network and a mean field term depending only on the overall prevalence of the state variables at a given time t . Although mesoscopic dependencies are in principle possible, we think they are of secondary importance and do not change the dynamical picture in general.

With $\chi_i(t)$ denoting the state variable of individual i at time t , for the probability of switching the state from a to b we formally have:

$$P\{\chi_i(t+1) = b \mid \chi_i(t) = a\} = P_{loc}(\chi_{B_1(i)}(t)) \oplus P_{mean}(s(t)); a, b \in \{0, 1, 2\},$$

where the sum sign stands for independent superposition² of the local and global infection events. The subscript $B_1(i)$ denotes the neighborhood of i and $s(t)$ the total number (or density) of vertices not in a susceptible or immune state at time t .

The local dynamics we use here is inspired by generalized epidemic processes [3]. In contrast to classical epidemics, where the probability of infection is an independent superposition of the single infection events caused by each infected neighbor (for small infection probabilities this means that the probability to get infected is proportional to the number of infected neighbors), every vertex i admits a threshold Δ_i of infected neighbors, at which the probability of infection suddenly jumps to a high probability. More precisely, let $N_i(t)$ denote the number of neighbors of a susceptible individual i at time t which support or accomplish terrorists. Hence $N_i(t)$ is the sum of state 1 and state 2 neighbors of i at time t . Then, the probability that individual i will become a passive supporter at time $t + 1$ is given by:

$$P_{loc}\{\chi_i(t+1) = 1 \mid \chi_i(t) = 0\} = \begin{cases} \epsilon N_i(t), & N_i(t) < \Delta_i \\ \alpha, & N_i(t) \geq \Delta_i \end{cases}$$

²Independent superposition of two probabilities A and B is just $A + B - AB$

Examples of this type of local dynamics are the spread of rumors, prejudices, knowledge or beliefs. So, whenever one vertex has contact with too many (here Δ or more) passive supporters or terrorists, they will sooner or later become passive supporters of the terrorists, either because they get convinced by the ideology or because they get exploited, unintentionally or even unnoted by themselves. Passive supporters of terrorism do not necessarily support terrorist ideologies, they may have been forced to support or may have been paid for. They may become uncooperative or useless to the terrorists, so they discontinue support. We assume that for every vertex there is a chance of switching back from 1 or 2 to 0 with a given probability γ .

Also there is a fixed rate k at which terrorists recruit from passive supporters. Due to the action of external forces like the state or allied troops, or by death or by migration, terrorists can be neutralized and removed from the network with probability ρ . Emotionally involved relatives and friends of victims, who where in state 0 until then, will potentially decide to join the terrorists side to take revenge. Civilians watching the News may be upset and enraged when they observe counter terrorists taking out their fellow citizens and destroying their buildings to capture or to eliminate terrorists.

We assume a fixed rate κ at which a neutralized terrorist generates new passive supporters. Thus, for a susceptible vertex in the network, the probability of becoming a passive supporter depends (besides the above describe local infection way) on the number of neutralized terrorist per time step and therefore on the number of active terrorists in the network. In this way, κ constitutes the mean field dynamics of the infection process. If $C(t)$ denotes the number of captured terrorists at time t we have formally:

$$P_{mean} \{ \chi_i(t+1) = 1 \mid \chi_i(t) = 0 \} = 1 - (1 - \kappa)^{C(t)} \quad (1)$$

$$\sim \kappa C(t); \text{ for small } \kappa. \quad (2)$$

Note that if one ignores mass media effects the total number of new mean field induced passive supporters should stay bounded irrespective of the size of the population. We therefore scale the κ value with the population size appropriately.

Table 1 summarizes all parameters and their interpretations.

Table 1. Summary of all parameters

| parameter | transition | affiliation |
|------------|----------------------|------------------------------------|
| ϵ | $0 \rightarrow 1$ | local process below the threshold |
| α | $0 \rightarrow 1$ | local process above the threshold |
| κ | $0 \rightarrow 1$ | mean field |
| γ | $1, 2 \rightarrow 0$ | changing mind |
| k | $1 \rightarrow 2$ | procreation rate of new terrorists |
| ρ | / | neutralization rate of terrorists |

3. Phase Transitions in the Local Infection Process

The usage of local threshold dynamics yields additional phase transitions that are different from those in classical epidemics, which are phase transitions in the parameters.

Here, in addition to these phase transitions in the parameters there are such in the dynamical variables like the number of passive supporters. In the following we state some analytic results for these phase transitions for the case of the classical Erdős&Renyi random graph $G(n, p)$, the graph with n vertices and independent edge probability p between each pair of vertices.

To investigate the average infection density b_t on a $G(n, p)$ with $p = c/n$ we assume for simplicity $\epsilon = 0, \alpha = 1$ and that the thresholds Δ_i are distributed uniformly, i.e. $\Delta_i = \Delta$ for all vertices i . It is easy to see that the case of very small ϵ and rather large α gives similar results although the formulas become more complicated. For a detailed mathematical treatment see [3]. There is a recursion for b_t , namely

$$b_{t+1} = 1 - (1 - b_0)e^{-cb_t} \sum_{k=0}^{\Delta-1} \frac{(cb_t)^k}{k!}.$$

Thus, for $\Delta = 2$ we get for the fixed point equation

$$b^* = 1 - (1 - b_0) e^{-cb^*} (1 + cb^*). \quad (3)$$

Note that there can be several solutions one has to take the dynamically stable one (closest to b_0). A closer examination shows that, depending on the value of b_0 and c , there are either 3 fixed points or just one in the domain $[0, 1]$. Since the iteration mapping has no critical points in this interval the smallest of these fixed point is also the attractor for the orbit starting with b_0 . Three examples are shown below. The phase transition happens when the first two fixed points (in case there are three) join together and form an indifferent (slope one) fixed point, see fig. 3. That means that for $b_0 < b_0^c$, i.e. if the initial prevalence is smaller than the critical value, the final infection density b^* will be not much larger than b_0 . If $b_0 > b_0^c$ then b^* will be close to the number of vertices in the k -core³ of the graph. The phase transitions are closely connected with the k -cores of a network. If $k \geq \Delta$ then every vertex in the k -core succeeds the threshold and gets infected sooner or later.

The critical density b_0^c , depending on the edge density c of the graph, is given by

$$b_0^c = 1 - \frac{2 \exp(-\frac{1}{2}(1 - c + \sqrt{c^2 - 2c - 3}))}{c(-1 + c - \sqrt{c^2 - 2c - 3})}$$

Figure 2 shows the function of the right hand side of eq. 3.

³A subgraph G' of a graph G is called k -core of G if every vertex in G' has at least k neighbors.

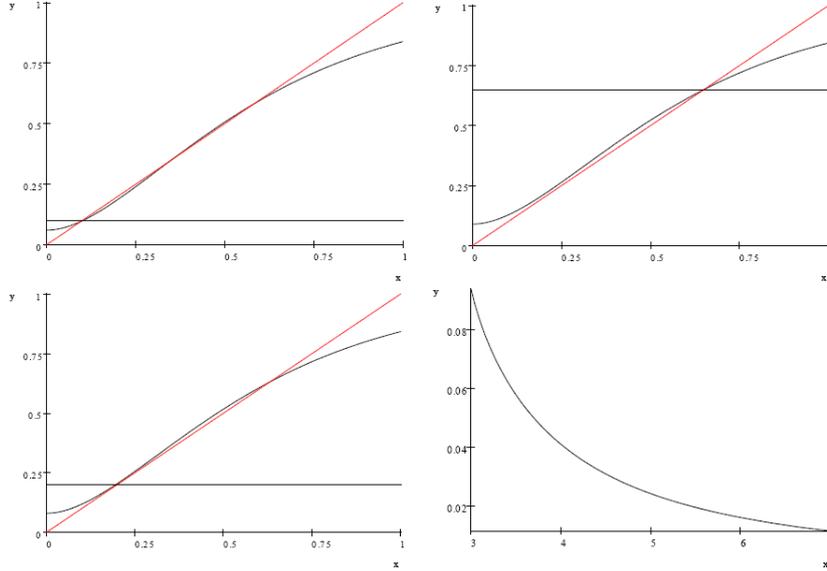


Figure 2. Right hand side of eq 3 (top left, top right, lower left). The black lines point to the stable fixed point. The graph in the lower right shows b_0^c as a function of b_0 .

The mean field dynamics described in section 2 induces now an effective slow increase of the initial density for the local spreading process until the critical density for the local dynamics is reached. Above the phase transition there remains then a nearly inexhaustible pool for recruiting new terrorists and any attempt to resolve the situation by military means has to fail. We close this section with a short calculation for the pure mean field dynamics. Let k be the probability that a passive supporter becomes a terrorist (for the simulations this is taken to be 0.1 percent). Assume that n is large and $\kappa n =: \varkappa$. Then the branching process approximation to the pure mean field process (which is valid for small numbers of initially infected) becomes overcritical if

$$\frac{k\rho\varkappa}{(\gamma + k(1 - \gamma))(\gamma + \rho(1 - \gamma))} > 1. \quad (4)$$

Note that in case the critical threshold for the local infection process is very small even in the subcritical case the mean field dynamics can due to stochastic fluctuations trigger the dynamics above to the phase transition.

4. Simulation Results

The simulations are run on an Erdős&Renyi random graph $G(n, p)$. Throughout all simulations the parameters of the local infection process are chosen to be $\alpha = 1$, $\epsilon = 0$. On average, one percent of the passive supporters change back to being susceptible, $\gamma = 0.01$. The probability of becoming an active terrorist is zero for all susceptible vertices

and 0.1 percent for all passive supporters. Thus, on average 1000 passive supporters generate one single active terrorist.

On each picture the abscissa is time, where we interpret one time step $\Delta t = 1$ to be one week. There are other associations possible of course, but in the context of social networks in regions, which are struck by terror, associating one time step with one week, giving people the opportunity to act and communicate, seems to be a natural choice.

The y-axis is the density of vertices in state 1 or 2, thus having values running from 0 to 1. There are three graphs shown, the dark blue one being the number of active terrorists, the dark green one being the number of passive supporters and the turquoise one being the sum of these two.

Figure 3 shows the case of an Erdős-Renyi random graph $G(n, p), p = c/n$ with $n = 200'000$ and average degree of $c = 4$. The threshold for the local dynamics is set to $\Delta = 2$, i.e. as soon as one vertex has two passive supporters in its neighborhood it turns to a passive supporter as well. The initial density of passive supporters is set to $b_0 = 0.04$, i.e. there are 8000 passive supporters from the start. The mean field parameter κ , which is the rate at which one captured active terrorist generates passive supporters, is set to $\kappa = 0,00001$, i.e. one captured active terrorist generates on average 2 passive supporters.

The process is run for a span of $t = 500$ steps, corresponding to about 10 years of time. Choosing $\rho = 0.01$ (left) yields a phase transition in the number of passive supporters and a sudden growth of the number of active terrorists, while choosing $\rho = 0.001$ (right) yields no phase transition. While in both cases the number of passive supporters increases directly from the start to a level, which is slightly decreasing for one third of the time, the higher value of the capturing rate triggers the mean field process to produce enough passive supporters to lift the local process to an overcritical level, where the calamitous phase transition occurs, resulting in an irreversible increase of active terrorists.

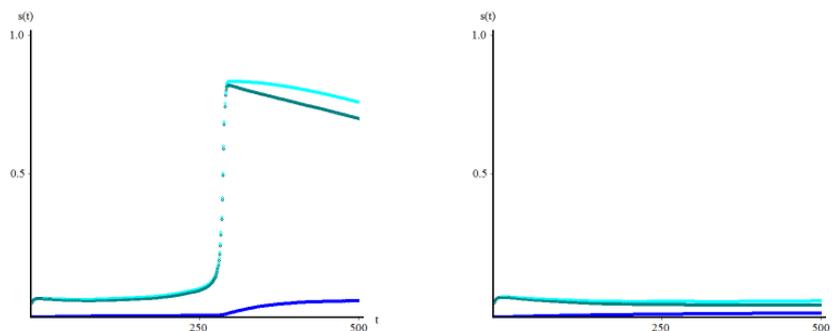


Figure 3. Densities of active terrorists and passive supporters on an Erdős-Renyi random graph $G(n, p), p = c/n$ with $n = 200'000$ and $c = 4$, threshold $\Delta = 2$, initial density of passive supporters $b_0 = 0.04$, mean field $\kappa = 0.00001$, in a time span of $t = 500$ steps. Choosing $\rho = 0.01$ (left) yields a phase transition in the number of passive supporters and a sudden growth of the number of active terrorists, while choosing $\rho = 0.001$ (right) yields no phase transition or a phase transition outside of the time scale under consideration.

The impact of changing the capturing parameter ρ is shown in Figure 4, on an Erdős&Renyi random graph $G(n, p), p = c/n$ with $n = 50'000$ vertices and average degree $c = 4$, threshold $\Delta = 2$, initial density of passive supporters $b_0 = 0.01$, i.e.

500 passive supporters from the start. The mean field parameter is $\kappa = 0,001$, i.e. one captured active terrorist generates 50 passive supporters here. Choosing $\rho = 0.004$ (left) or $\rho = 0.0025$ (right) only alters the moment when the phase transition happens.

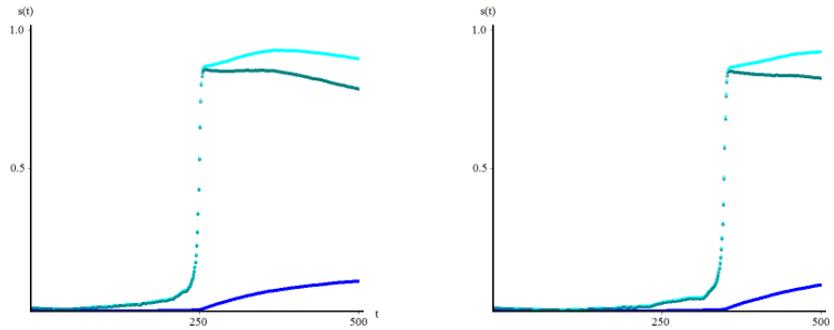


Figure 4. Densities of active terrorists and passive supporters on an Erdős&Renyi random graph $G(n, p)$, $p = c/n$ with $n = 50'000$ and $c = 4$, threshold $\Delta = 2$, initial density of passive supporters $b_0 = 0.01$, mean field $\kappa = 0.001$, in a time span of $t = 500$ steps. Choosing $\rho = 0.0004$ (left) or $\rho = 0.00025$ (right) only alters the moment when the phase transition happens.

Fig 5 shows the simulation run on a sample of the anonymized StudiVz network, including members associated with scholars of Bielefeld and their friends. The network has a size of approximately 400'000 vertices with an average degree of about 7.2, for more details see [4]. With an initial density of passive supporters of $b_0 = 0.00005$, corresponding to an average of 20 passive supporters from the start, and with capturing rate $\kappa = 0.0001$, corresponding to an average of 40 citizens getting converted to passive supporters, we observe the phase transition occurring as soon as one or two active terrorists get caught.

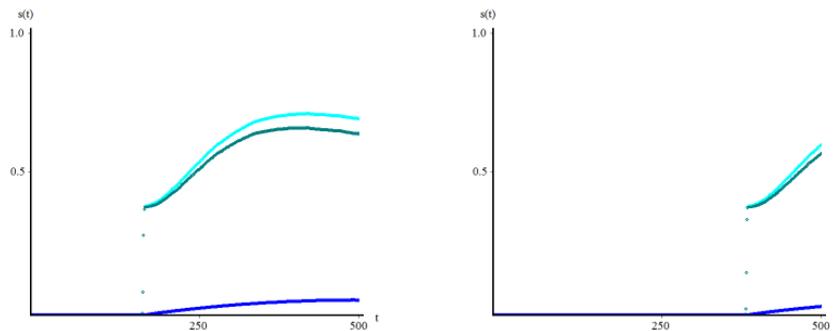


Figure 5. Densities of active terrorists and passive supporters on a sample of the StudiVz network, two single runs. The threshold is $\Delta = 2$, initial density of passive supporters $b_0 = 0.00005$, mean field $\kappa = 0,0001$, in a time span of $t = 500$ steps. The phase transition occurs as soon as one or two active terrorists get captured.

5. Conclusions and Perspectives

The main observation is the existence of a phase transition in the number of passive supporters of terroristic activities. Whenever counter terrorist activities lead to collateral

damages, the likelihood of outraging civilians rises. A high number of passive supporters provides a steady pool to recruit active terrorists, so the number of active terrorists and their attacks increases, as fig. 1 suggests and as resembled by the graph for the number of active terrorists in figures 3, 4 and 5.

Lowering the rate ρ of removal of active terrorists to avoid the phase transition is not what our results suggest. The interplay of the mean field term κ , which is the rate at which removed active terrorists generate passive supporters, and ρ has to be taken into account. Avoidable failures resulting in casualties, high collateral damage, pictures and videos of humiliated inmates in Allied prisons, are factors which increase the probability that the civil population will join the terrorist side instead of fighting against it.

If the Allied forces want to avoid the phase transition in the number of passive supporters to not gain a stable number of active terrorist, capturing or removing active terrorists from the network would make sense therefore only if this happened practically without casualties, fatalities, applying torture or committing terroristic acts against the local population.

If this is not possible - and evidence is pointing towards this - our results strongly indicate that there is no military solution to fight terrorism, so only political solutions are available.

A refinement of the model may use weighted graphs, where the vertices have properties like credibility or importance, to include intra-organization dynamics, e.g. to model the emergence of terror cells. In the same fashion the influence of local warlords and tribal conflicts may be described.

References

- [1] Juan Camilo Bohorquez, Sean Gourley, Alexander R. Dixon, Michael Spagat, Neil F. Johnson, "Common ecology quantifies human insurgency", Nature Vol 462|17 December 2009, 911-914
- [2] Serge Galam, "Global Physics: From Percolation to Terrorism, Guerilla Warfare and Clandestine Activities", 2004, arXiv:cond-mat/0404265v1 [cond-mat.dis-nn]
- [3] P. Blanchard, S. Delitzscher, A. Krüger, T. Krüger, R. Siegmund-Schultze, "Generalized Epidemic Processes and Threshold Percolation with Application to Knowledge Diffusion", submitted
- [4] C. Lee, T. Scherngell, M. J. Barber, "Real-world separation effects in an online social network", 2009, arXiv:0911.1229v1 [physics.soc-ph]