ICTeCollective – Harnessing ICT-enabled collective social behaviour
Project no. 238597

Grant agreement: Small or medium-scale focused research project
Programme: FP7-ICT

Deliverable D2.2
[Report on structure and communication channel preferences of communities]

Submission date: 2011-11-05

Start date of project: 2009-10-01 **Duration: 36 months**

Organisation name of lead contractor for this deliverable: ISI

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	**X**

Document information

1.1 Author(s)

Author	Organisation	E-mail
S. Fortunato	ISI	fortunato@isi.it
J. Kertész	BME	kertesz@phy.bme.hu
G. Tibély	BME	tibelyg@gmail.com
K. Kaski	Aalto	kaski@gmail.com
J. Saramäki	Aalto	jari.saramaki@gmail.com
M. Karsai	Aalto	karsai.marton@gmail.com
R. Dunbar	Oxford	robin.dunbar@anthro.ox.ac.uk
L. Kovanen	Aalto	lauri.kovanen@gmail.com
F. Reed-Tsochas	Oxford	felix.reed-tsochas@sbs.ox.ac.uk

1.2 Other contributors

Name	Organisation	E-mail
R. Pan	Aalto	rajkp@cc.hut.fi
K. Zhao	Northeastern University	zhao.k@husky.neu.edu
G. Bianconi	Northeastern University	g.bianconi@neu.edu
M. Kivelä	Aalto	mtkivela@lce.hut.fi
E. Leicht	Oxford	Elizabeth.leicht@sbs.ox.ac.uk
E. Lopez	Oxford	Eduardo.lopez@sbs.ox.ac.uk
S. Roberts	Oxford, Univ. of Chester	Sam.roberts@chester.ac.uk

1.3 Document history

Version#	Date	Change
V0.1	15/10/2011	Starting version, template
V0.2	1/11/2011	First version for circulation
V1.0		Approved version to be submitted to EU

1.4 Document data

Keywords	Community structure, entropy
Editor address data	jari.saramaki@gmail.com
Delivery date	5/11/2011

1.5 Distribution list

Date	Issue	E-mail
01/11/2011	Consortium members	kaski@gmail.com
	Project officer	
	EC archive	

Contents

ICTeCollective Consortium

This document is part of a research project funded by the ICT Programme of the Commission of the European Communities as grant number ICT-2009-238597.

Aalto University (Coordinator)
School of Science
Department of Biomedical Engineering and Computational Science
FI-00076 AALTO, Espoo
Finland
Contact person: Prof. Kimmo Kaski
E-mail: kimmo.kaski@tkk.fi

Budapest University of Technology and Economics
Institute of Physics
Budapest, H-1111
Hungary
Contact person: Prof. János Kertész
E-mail: kertesz@phy.bme.hu

University of Oxford
Saïd Business School
The CABDyN Complexity Centre
Oxford, OX1 1HP
United Kingdom
Contact Person: Dr Felix Reed-Tsochas
E-mail: felix.reed-tsochas@sbs.ox.ac.uk

I.S.I Foundation
Torino 10133, Italy
Contact person: Dr. Santo Fortunato
E-mail: fortunato@isi.it

University of Warsaw
Faculty of Psychology
Warsaw 00927, Poland
Contact Person: Prof. Andrzej Nowak
E-mail: nowak@fau.edu

ICTeCollective introduction

ICTeCollective (*Harnessing ICT enabled collective social behaviour*) aims to develop systematic means of exploring, understanding and modelling2 systems where ICT is entangled with social structures. In particular, we will focus on behavioural patterns, dynamics and driving mechanisms of social structures whose interactions are ICT-mediated, from the level of individuals to the level of groups and large-scale social systems. Our unique approach is based on combined expertise in complex systems and the social sciences. By contrast with the majority of complexity studies that start from extremely simplified assumptions concerning social dynamics and concentrate on diagnosing structural features of social systems, we emphasize that ICT networks are dynamic systems of interacting humans and groups, and fully utilize the theories and methods of the social sciences are to be in ICTeCollective.

We will study and relate high quality datasets on ICT mediated social interactions and groups that have already been acquired, and also create new sets of data by conducting experiments with human subjects to examine the properties of social interactions mediated by technological means. The first source of data, electronic records of interactions, is a by-product of how ICT mediated communities operate. In particular, we will use some of the most extensive ICT datasets available at present, such as time-stamped data sets on mobile telephone communications between millions of users, the editing history of Wikipedia documents, and the popularity of Facebook applications. Secondly, entirely new data will be generated and released into the public domain by conducting laboratory experiments on ICT-mediated human interactions.

This project addresses the goals of the FP7 FET-OPEN call by trying to build an integrated picture of ICT-mediated social systems focussing on some aspects that are

i) critical to social interaction,
ii) can be easily tracked in large datasets and confirmed in experiments, and
iii) have a considerable

chance of improving our understanding and usage of ICT, with the possibility of leading to new and exciting technologies that can shape the future of ICT. The particular aspects that we focus on are *activity patterns*, *social influence*, and *group dynamics*. This choice helps us to address a large number of practical issues such as the driving mechanisms of social interactions mediated by ICT, and how these mechanisms then shape groups and society. All of these are critical to the goals of ICTeCollective.

We define the above terms as follows: *Activity patterns* are temporal sequences of social interaction and communication events, measurable in electronic communication records and representing the "atoms" of social interaction processes. *Social influence* refers to all processes where individuals affect each others' beliefs, behaviour, activities, and representations of reality. *Group dynamics* comprises processes such as emergence, growth, merging, and splitting of groups, and associated behavioural patterns of individuals.

Executive Summary

This deliverable summarizes research activities concerning community structure in our datasets, specifically the data on mobile phone communications (DS1). The contributions are both of theoretical and of empirical nature. The goals of the task were essentially two: 1) pointing out the variety of communities based on the different communication channels, specifying properties both at the level of the groups and at the level of the individual nodes and edges; 2) analysing the community structure of DS1, and 3) deepening our understanding of the community structure from the perspective of individuals and their communication patterns; this was achieved using the longitudinal data set DS2 that combines call records with survey results..

The first goal was addressed mostly by the work of Zhao, Karsai and Bianconi [1]. In it, the authors manage to find an important difference in the patterns of mobile phone communications as opposed to face-to-face interactions, as revealed by the different distributions of contact durations in the two cases. Also, they use a fairly recent entropy measure to estimate the information content of networks of mobile communications and to estimate the predictability of future interactions. The analysis shows that the predictability depends on the circadian rythms.

As for the second goal, the consortium has produced three papers. The paper by Tibély [2], mostly theoretical, presents a new definition of community, that accounts for the two main features that one would expect to be common of most clusters in real systems: separation and cohesion. The main point of the paper is that most methods proposed so far are based on the concept of separation, without considering the fact that clusters are supposed to be cohesive as well. This leads to a method that tries to find an ideal tradeoff between separation and cohesion, which could lead to a class of better-performing methods.

In the paper by Tibély et al. [3], one moves from the analysis of partitions found by methods in artificial benchmark graphs to results in real systems, which ultimately are the systems one wishes to explore. The authors consider three popular methods of community detection [Louvain, Infomap and Clique Percolation (CP)], they apply it to DS1 and compare the properties of the detected clusters. The main finding is that, while the Louvain method and Infomap tend to find clusters, which may not be very cohesive, especially when they are small, the internal link density of the clusters found by CP is larger, seemingly more consistent with social communities. The CP, in turn, may also find structures that do not match the intuitive properties that real clusters should have. Nevertheless, by comparing the partitions obtained by different methods, it turns out that the clusters detected by one method can be tiled by those found by the other methods, which typically are subsets or supersets of the former. So, the take home message of the work is that one could get meaningful results by using any method, but that adopting several techniques for the same system give a more insightful perspective and more information on the community structure of the network.

In the paper by Pan et al. [4], the authors have used DS1, along with a network of scientific collaborations, to check whether a recent percolation model proposed by Achlioptas et al., called explosive percolation and mostly studied at the theoretical level, has any relationship with real systems. The model consists in a progressive addition of links to the system, such to slow down the growth of the clusters of nodes joined by the links, until there is a sudden coalescence of many of them to form a giant percolation cluster, with a transition that is continuous, despite early allegations, but very special as for its scaling behaviour. The authors check what happens when the links of the social networks analysed are added according to the similar rules as in the Achlioptas process. They find that there is a close relationship between the explosive percolation transition

and the networks' community structure, and that the transition helps to distinguish quantitatively the two different types of social interactions underlying the two datasets.

Finally, for a more detailed view on the networks of individuals and the structure of their social ties, we have studied the data set DS2 on 30 students who are in transition from school to university and are thus experiencing a period of flux in their networks. We have found from call and text message records augmented with survey results that although the time allocation patterns of communication show similar features, such as an unexpectedly large fraction of communication going to a few top-ranked alters, there is still individual variation; however, the patterns of individuals remain surprisingly persistent and similar in time, although the networks themselves change in composition.

1. Entropy of dynamical social networks

This section is based on Kun Zhao, Márton Karsai and Ginestra Bianconi. *Entropy of dynamical social networks* (submitted).

1.1 Introduction

Human dynamical social networks encode information and are highly adaptive. To characterize the information encoded in the fast dynamics of social interactions, we introduce the entropy of dynamical social networks. By analysing a large dataset of phone-call interactions we show evidence that the dynamical social network has an entropy that depends on the time of the day in a typical weekday [1]. Moreover we show evidence for adaptability of human social behavior showing data on duration of phone-call interactions that significantly deviates from the statistics of duration of face-to-face interactions. This adaptability of behavior corresponds to a different information content of the dynamics of social human interactions. We quantify this information by the use of the entropy of dynamical networks on realistic models of social interactions.

Social networks are characterized by complex organizational structures revealed by network community and degree correlations. These structures are sometimes correlated with annotated features of the nodes or of the links such as age, gender, and other annotated features of the links such as shared interests, family ties or common work locations. In a recent work it has been shown by studying social, technological and biological networks that the network entropy measures can assess how significant are the annotated features for the network structure. Moreover social networks evolve on many different time-scales and relevant information is encoded in their dynamics. In fact social networks are highly adaptive. Indeed social ties can appear or disappear depending on the dynamical process occurring on the networks such as epidemic spreading or opinion dynamics. Several models for adaptive social evolution have been proposed showing phase transitions in different universality classes. Social ties have in addition to that a microscopic structure constituted by fast social interactions of the duration of a phone call or of a face-to-face interaction. Dynamical social networks characterize the social interaction at this fast time scale.

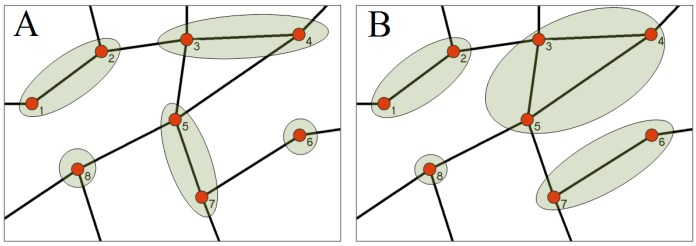


Fig. 1. The dynamical social networks are composed by different dynamically changing groups of interacting agents. In panel (A) we allow only for groups of size one or two as it typically happens in mobile phone communication. In panel (B) we allow for groups of any size as in face-to-face interactions.

Thanks to the availability of new extensive data of dynamical social networks, it has been recently recognized that many human activities are bursty and not Poissonian [5]. New data on social dynamical networks start to be collected with new technologies such as of Radio frequency Identification Devices and Bluetooth. These technologies are able to record the duration of social interactions and report evidence for a bursty nature of social interaction characterized by a fat tail distribution of the duration of face-to-face interactions. This bursty behavior of social networks is coexisting with modulations coming from periodic daily (circadian rhythms) or weakly patterns. Thus these newly available data let us to take the question: How much can humans intentionally change the statistics of social interactions and the level of information encoded in the dynamics of their social networks, when they are interfacing with a new technology? In this work we tried to find an answer by defining the entropy for dynamical networks, which is a sensible measure to study differences between communication channels.

1.2 Entropy of dynamical social networks

In order to define entropy for dynamical social networks we assume to have a quenched social network G of friendships, collaborations or acquaintances formed by N agents and we allow a dynamics of social interactions on this network. If two agents i, j are linked in the network they can meet and interact at each given time giving rise to the dynamical social network under study in this paper. If a set of agents of size N is connected through the social network G the agents i_1, i_2, . . . i_n can interact in a group of size n. Therefore at any given time the static network G will be partitioned in connected components or groups of interacting agents as shown in Fig 1.

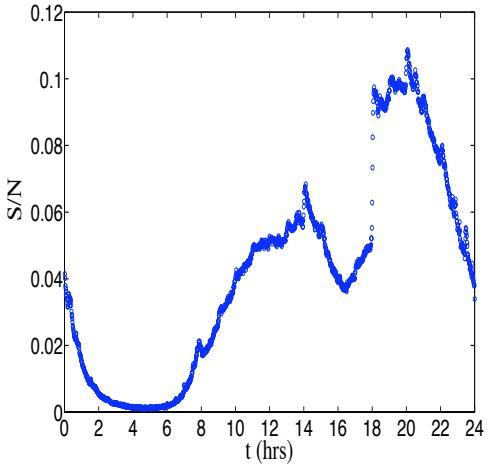


Fig. 2. Mean-field evaluation of the entropy of the dynamical social networks of phone calls communication in a typical week-day. In the nights the social dynamical network is more predictable.

In order to indicate that a social interaction is occurring at time t in the group of agents $i_1, i_2, \ldots i_n$ and that these agents are not interacting with other agents, we write $g_{i1, i2, \ldots in}(t)=1$ otherwise we put $g_{i1, i2, \ldots in}(t)=0$. Therefore each agentis interacting with one group of size n > 1 or non-interacting (interacting with a group of size n = 1). The entropy S characterizes the logarithm of the typical number of different group configurations that can be expected in the dynamical network model at time t. According to the information theory results, if the entropy is vanishing, i.e. S = 0 the network dynamics is regular and perfectly predictable, if the entropy is larger the number of future possible configurations is growing and the system is less predictable. If we model face-to-face interactions we have to allow the possible formation of groups of any size, on the contrary, if we model the mobile phone communication, we need to allow only for pairwise interactions. Therefore, if we define the adjacency matrix of the network G as the matrix a_{ij}, the entropy can be written as:

$$S = -\sum_i p(g_i(t)=1|\mathcal{S}_t)\log p(g_i(t)=1|\mathcal{S}_t) - \sum_{ij} a_{ij} p(g_{ij}(t)=1|\mathcal{S}_t)\log p(g_{ij}(t)=1|\mathcal{S}_t).$$

where:

$$g_i(t) + \sum_j a_{ij} g_{ij}(t) = 1$$

1.3 Social dynamics and entropy of phone call interactions

We have analyzed the call sequence of subscribers of a major European mobile service provider. We considered calls between users who at least once called each other during the examined 6 months period in order to examine calls only reflecting trusted social interactions. The resulted event list consists of 633,986,311 calls between 6, 243, 322 users. For the entropy calculation we selected 562,337 users who executed at least one call per a day during a week period. First of all we have studied how the entropy of this dynamical network is affected by circadian rhythms. We assign to each agent i=1,2 a number n_i=1,2 indicating the size of the group where he/she belongs. If an agent i has

coordination number n_i=1 he/she is isolated, and if n_i=2 he/she is interacting with a group of n=2 agents. We also assign to each agent i the variable t_i indicating the last time at which the coordination number n_i has changed. If we neglect the feature of the nodes, the most simple transition probabilities that includes for some memory effects present in the data, is given by a probability p_n=p_n(,t) for an agent in state n at time t to change his/her state given that he has been in his/her current state for a duration = t − t_i.

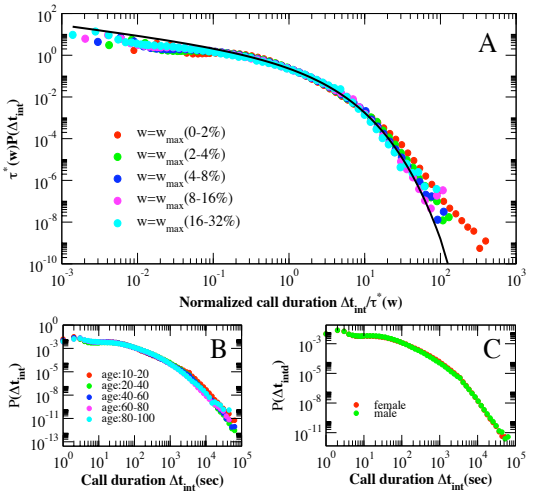


Fig. 3. (A) Probability distribution of duration of phone-calls between two given persons connected by a link of weight w. The data depend on the typical scale $\tau^{\star}(w)$ of duration of the phone-call. (B) Probability distribution of duration of phone calls for people of different age. (C) Probability distribution of duration of phone-calls for people of different gender. The distributions shown in the panel (B) and (C) do not significantly depend on the attributes of the nodes.

We have estimated the probability p_n(,t) in a typical week-day. Using the data on the probabilities p_n(, t) we have calculated the entropy, estimated by a mean-field evaluation of the dynamical network as a function of time in a typical week-day. The entropy of the dynamical social network is reported in Fig.2. It significantly changes during the day describing the fact that the predictability of the phone-call networks change as a function of time. In fact, as if the entropy of the dynamical network is smaller the network is in a more predictable state.

1.4 Adaptive dynamics face-to face interactions and phone call durations

In this section we report evidence of adaptive human behavior by showing that the duration of phone calls, a binary social interactions mediated by technology, show different statistical features respect to face-to-face interactions. The distributions of the times describing human activities are typically broad and are closer to power- laws, which lack a characteristic time scale, than to exponentials. In particular in [6] data on Radio Frequency Identification devices were reported, with temporal resolution of 20s, showing that both distribution duration of face-to-face contacts and inter-contact periods is fat tailed during conference venues.

Here we analysed the above defined mobile-call event sequence performing the measurements on all the users for the entire 6 months time period. The distribution of phone-call durations strongly deviates from a fat-tail distribution. In Fig 3 we report this distributions and show that these distributions depend on the strength w of the interactions (total duration of contacts in the observed period) but do not depend on the age, gender or type of contract in a significant way.

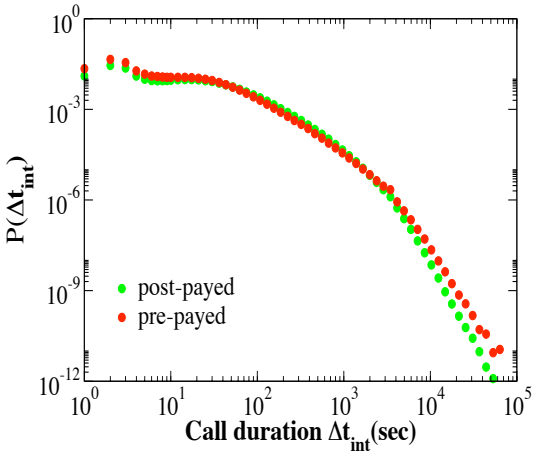


Fig. 4. Probability distribution of duration of phone-calls for people with different types of contract. No significant change is observed that modifies the functional form of the distribution.

The distribution $P^w(\Delta t_{in})$ of duration of contacts within agents with strenght w is well fitted by a Weibull distribution. The origin of this significant change in behavior of humans interactions could be due to the consideration of the cost of the interactions (although we are not in the position to draw these conclusions (See Fig. 4 in which we compare distribution of duration of calls for people with different type of contract) or might depend on the different nature of the communication.

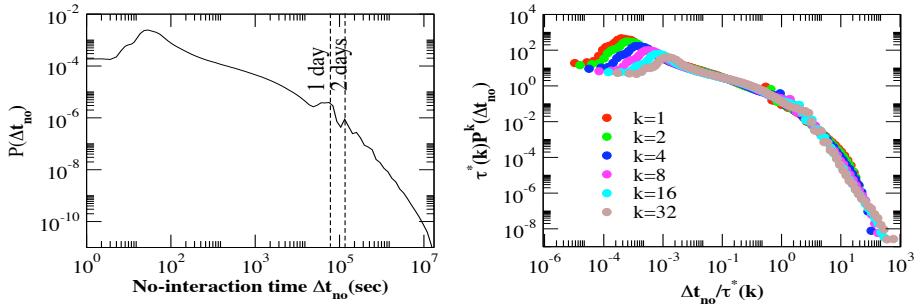


Fig. 5. Distribution of non-interaction times in the phone-call data. The distribution strongly depends on circadian rhythms. The distribution of rescaled time depends strongly on the connectivity of each node. Nodes with higher connectivity k are typically non-interacting for a shorter typical time scale $\tau^*(k)$.

The duration of a phone call is quite short and is not affected significantly by the circadian rhythms of the population. On the contrary the duration of no-interaction periods is strongly affected by periodic daily of weekly rhythms. The distribution of no-interaction periods can be fitted by a double power-law but also a single Weibull distribution can give a first approximation to describe $P(\Delta t_{no})$. In Fig. 5 we report the distribution of duration of no-interaction periods in the day periods between 7AM and 2AM next day.

1.5 Conclusions

In the last ten years it has been recognized that the vast majority of complex systems can be described as networks of interacting units. Network theory has made tremendous progresses in this period and we have gained important insight into the microscopic properties of complex networks. Key statistical properties have been found to occur universally in the networks, such as the small world properties and broad degree

distributions. Moreover the local structure of networks has been characterized by degree correlations, clustering coefficient, loop structure, cliques, motifs and communities. The level of information present in these characteristics of the network can be now studied with the tools of information theory. An additional fundamental aspect of social networks is their dynamics. This dynamics encode for information and can be modulated by adaptive human behavior. In this work [1] we have introduced the entropy of social dynamical networks and we have evaluated the information present in dynamical data of phone-call communication. By analysing the phone-call interaction networks we have shown that the entropy of the network depends on the circadian rhythms. Moreover we have shown that social networks are extremely adaptive and are modified by the use of technologies. The statistics of duration of phone-call indeed is described by a Weibull distribution that strongly differ from the distribution of face-to-face inter- actions in a conference

2. Mesoscopic structures in large networks

This section is based on G. Tibély, *Criterions for locally dense subgraph*s, in press (Physica A, 2011).

We have continued our studies of intermediate, mesoscopic structures in large, ICT related networks. The identification of such structures is one of the big challenges of this field. The question raises: Are the viewpoints applied so far sufficient for a proper definition of communities? According to our suggestion the answer to this question is no. We introduced the new concept of cohesion of communities, which characterizes the homogeneity of the modules. A method, which includes this concept, has lead to encouraging results on medium size networks [2]
The fact that the number of methods is of the order of hundred already indicates the difficulties. One of the questions is: How to select from this overwhelming supply of methods? The effort of constructing helpful benchmarks has lead to remarkable results [7, 8], however, for huge complex networks, as we meet them in the ICT related context, there is need for direct comparison on real data. We carried out such a study on a large mobile call graph using three most popular community detection methods [3].
The mesoscopic structure has significant effect on several properties of the networks. We studied the recently introduced model of explosive percolation [9] on real networks, including ICT ones. We showed that using this method an interesting analysis of the network structure can be performed [4].

2.1. Separation and cohesion

So far most methods of community detection have concentrated on the ratio of the within community to outgoing links. In other words, the focus was on how well the module is separated from the rest of the network.
We have pointed out that for a precise definition of the communities, we should add the viewpoint of homogeneity of the community. A module cannot be considered as a proper one, if it is easy to split it, in spite of the fact that it is well separated from the rest. We have introduced the concept of coherence to capture this aspect. Fig. 6 shows two equally well separated subgraphs which cohesion differs radically.

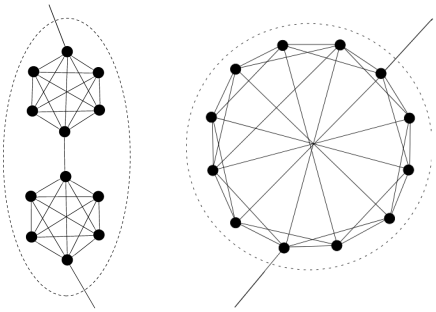


Figure 6. Two subgraphs where the number of corresponding in and out links are equal (same separation). The left figure has much lower coherence than the right one.

As for the mathematical formulation of coherence we suggested the second larges eigenvalue of the Laplacian matrix as constructed from the adjacency matrix.
Using the two criterions (separation and coherence), we managed to build a new community detection method, which takes into account both of them. For measuring the separation of a subgraph, the ratio of inside edges and all edges is appropriate, while for measuring the cohesion, the second eigenvalue of the Laplacian matrix of the subgraph can be utilized. Then, a fitness function containing both measures can be defined; the local optima of which correspond to the communities. Although the current realization of the method is quite slow, it is able to produce good results even in the presence of densely overlapping and hierarchically embedded clusters. For an illustration, the communities in a word association network around the word "bright" are shown on Fig. 7.

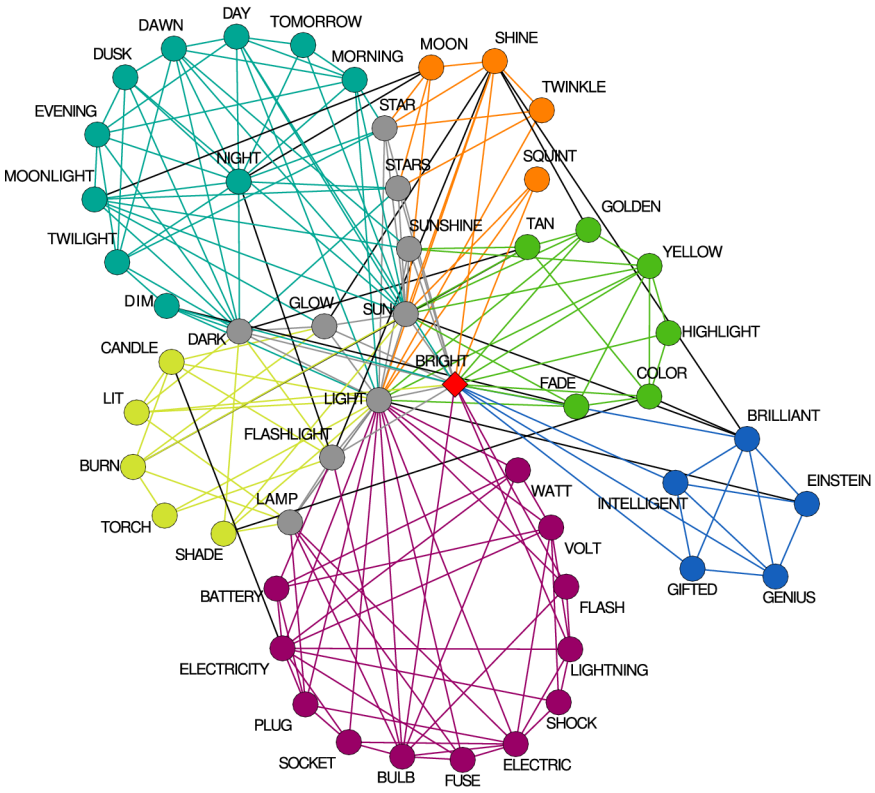


Figure 7. Communities of the word "bright" (at the center, red) in a word association network. Coloring indicates the communities. Gray nodes and edges are overlaps between different communities. Black edges are outside of communities.

2.2. Communities and beyond: mesoscopic analysis of a large social network with complementary methods

This section is based on G. Tibély, L. Kovanen, M. Karsai, K. Kaski, J. Kertész, J. Saramäki: *Communities and beyond: mesoscopic analysis of a large social network with complementary methods*, Phys. Rev. E 83, 056125 (2011).

Although several community detection methods have appeared recently, applications to large empirical networks are very scarce. Here, we compare three community detection methods - Infomap (IM) [10], Louvain (LV) [11] and Clique Percolation (CP) [12] on a network of phone calls from a single mobile phone provider, containing 4.9 million of anonimous users, aggregated over calls of 126 days. The edges of the networks are placed between users who mutually called each other. Using data of 126 days allows assigning weights to the edges. Therefore, we analyzed also the weighted version of each method (denoted by wIM, wLV, wCP, correspondingly).

All distributions are broad, as suggested by previous results on empirical community structures [12, 13, 14]. The tail of the size distributions appears power-law-like (Fig.8), with exponents around 3 for the unweighted and 5.7 for the weighted case (except CP,

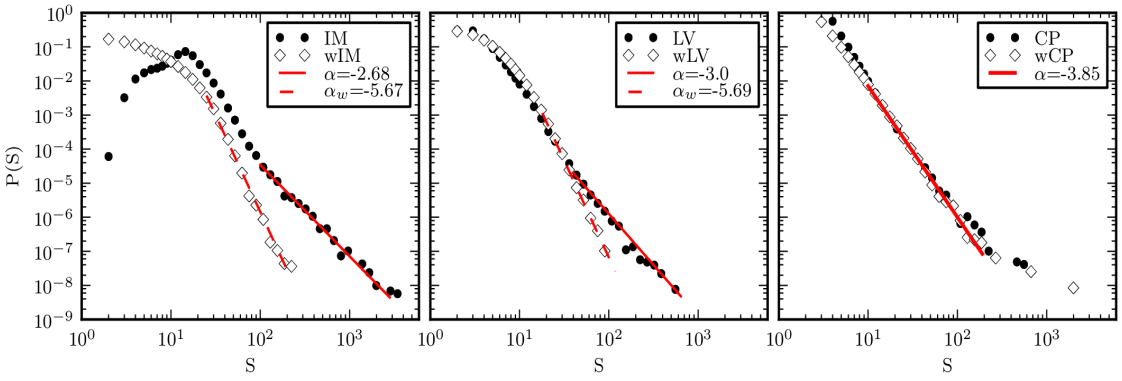


Figure 8: Community size distributions for IM, LV and CP and their weighted versions. The parameter alpha denotes the exponent when the tails are fitted a power-law distribution.

where there was no significant difference). The distributions are all monotonically decreasing, except a single case (IM).

Graph density is normally defined as the proportion of edges out of all possible edges. However, since communities are necessarily connected it is more illustrative to study density relative to the sparsest possible community, a tree. So, density of community *c* is defined as $D_c = L_c/(S-1)$, where S is the size of *c*, and L_c is the number of its edges.
Figure 9 shows the distributions and average values of D_c as function of community size. As expected, CP yields dense communities. On the other hand, IM and LV produces several small treelike communities. The plots for weighted communities in Fig. 9 suggest that weights make the communities more similar across methods. Both wIM and wLV communities are more treelike.
Treelike communities do not fit well either with the idea of social groups, or that of communities in general being dense groups of nodes. The abundance of treelike parts may just be a sampling artifact, as our network does not cover the whole population. One could argue that in treelike regions the network is so sparse that there isn't enough information about community structure. This makes CP's requirement---that nodes must participate in at least one clique to be assigned a community---appear meaningful. On the other hand, CP may yield communities where *cliques* are arranged as chains or starlike patterns, which again does not coincide well with the idea of social groups. Fig. 9 indicates that in CP and wCP there are indeed some communities with densities close to the lower bound. Whatever the interpretation, the detected tree-like structures do provide information about the mesoscopic structure of the network.

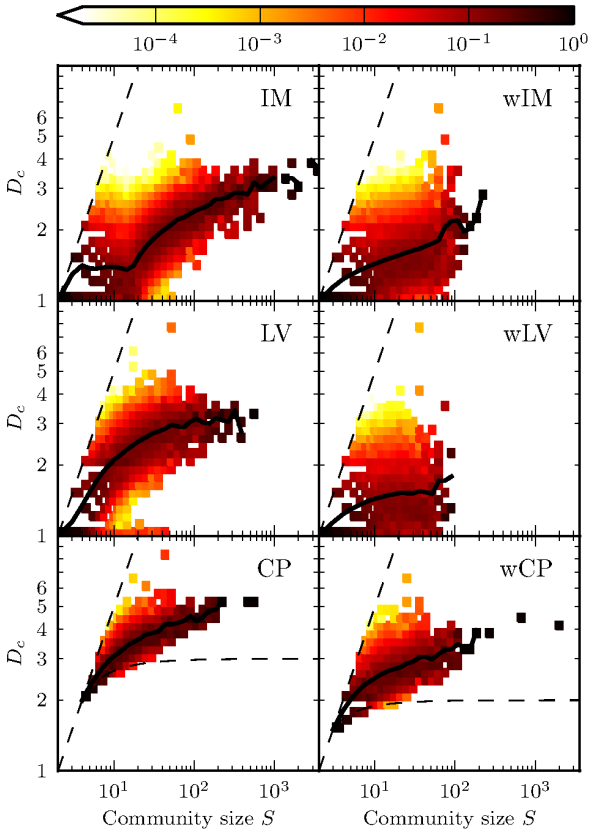


Figure 9. The distributions of relative density for communities from each method. The solid line denotes the average value. The dashed straight line corresponds to the maximal values. For CP, the smallest possible density differs from 1, and is indicated by the curved dashed line.

An important characteristic of communities is the ratio of outside and inside edges. Good communities should have many edges inside and only a few outside. Figure 10 shows the distribution of this ratio for the found communities of all methods. Especially IM can find good communities in this sense. The values for small communities are particularly low, confirming the earlier observation that small IM communities are on the "edges" of the network. Interestingly, including edge weights increases the average ratio.

From earlier studies of mobile phone call networks we know that there is a correlation between edge weight and neighbourhood overlap [15]. As nodes inside communities have overlapping neighbourhoods, we expect the links between communities to be on average weaker than those within communities. Table 1 shows that with all methods this is indeed the case.

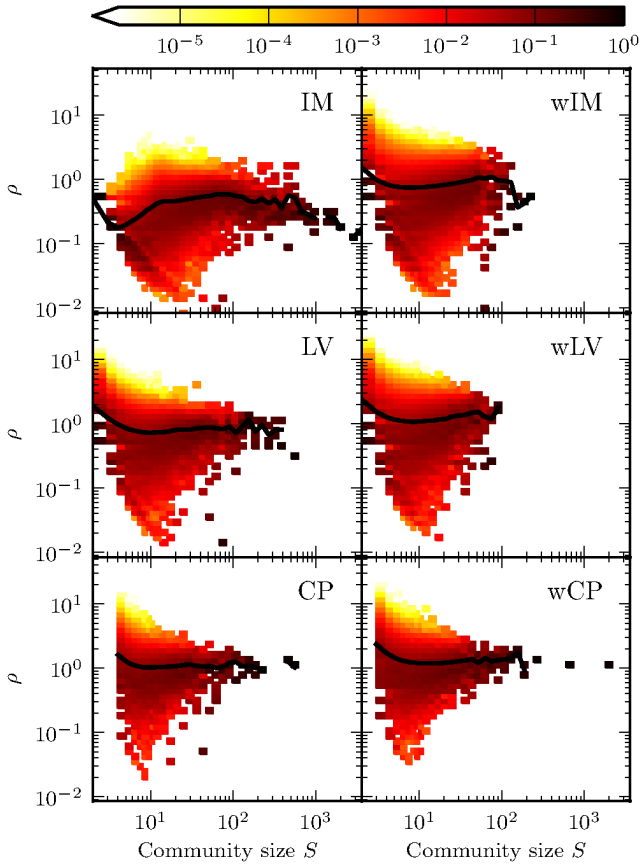


Figure 10. The distributions of the ratio of outside and inside edges as function of community size for each method. Black lines denote the averages.

	$\langle w_{in} \rangle / \langle w \rangle$	$\langle w_{out} \rangle / \langle w \rangle$
IM	1.14	0.69
LV	1.20	0.78
CP	1.20	0.57
wIM	1.65	0.18
wLV	1.92	0.25
wCP	2.57	0.43

Table 1: Edge weights inside and between communities. <w> denotes the average edge weight in the whole network, $\langle w_{in} \rangle$ *the average weight for edges inside communities and* $\langle w_{out} \rangle$ *between communities.*

We have found that all three methods have found the same basic set of communities, but some methods agglomerated them into larger clusters, while others keep them small. By properly defining a measure, tiling imperfection, it is possible to check that assumption. Tiling imperfection measures how well can the resulting communities of one method be tiled by communities of another one. The results show that the tiling imperfection is surprisingly low when tiling IM communities by both LV and CP ones, especially for

small communities. This suggests that the different methods tend to find clusters, which are in a subset-superset relation.
In large sparse networks partitioning methods inevitably identify some questionable regions as communities. The trees, star-like formations and stars detected by IM and LV do, however, bear mesoscopic structural meaning: they too are building blocks of the network. The same topological structure may be considered a community for one purpose but not for some other—a star is hardly a social community but may reasonably be considered as one in for example biochemical networks.

Our conclusion is twofold: First we emphasize the necessity of the use of complementary community detection methods and a comparison of the identified structural features. Second, we draw the attention to take into consideration the existence of different types of mesoscopic structures, as opposed to fixating on a predefined idea of dense communities.

2.3. Using explosive percolation in the analysis of real-world networks

This section is based on R. K. Pan, M. Kivelä, J. Saramäki, K. Kaski, and J. Kertész, *Using explosive percolation in analysis of real-world networks*, Phys. Rev. E 83, 046112 (2011).

Ordinary bond percolation can be considered as a process on an initially empty graph, where edges are selected one by one randomly and set in into the graph leading at some point to a percolation transition, the occurrence of a giant component. In the recently discovered "explosive percolation" [9] m edges are selected from which the one is added to the graph, which minimizes the product or sum of the sizes of the two components that would be merged. This model has attracted considerable interest because the nature of the transition changes.

While the activity on this field had been confined to theoretical studies, we applied the ideas to real world networks and discovered that the transition is closely related to the community structure of the graph. We analyzed the mobile phone call network and the ArXiv co-authorship network. Figure 11 shows different properties as a function of the fraction of already inserted links.

The modularity increases until the threshold indicating that right before it the explosive percolation clusters provide a reasonable community structure. In accord with this, the average overlap on the links drops. Interestingly, the weights behave differently on the two different datasets, indicating the differences in the organizing principles of the corresponding social systems. The mobile call network serves as a proxy of the network of interactions in the society thus the Granovetter picture holds here for the communities (strongly wired modules connected by weak links). In the case of the co-authorship network within the communities the principal investigators are weakly connected to the student and postdoc collaborators, while cooperation between different groups runs through links between professors.

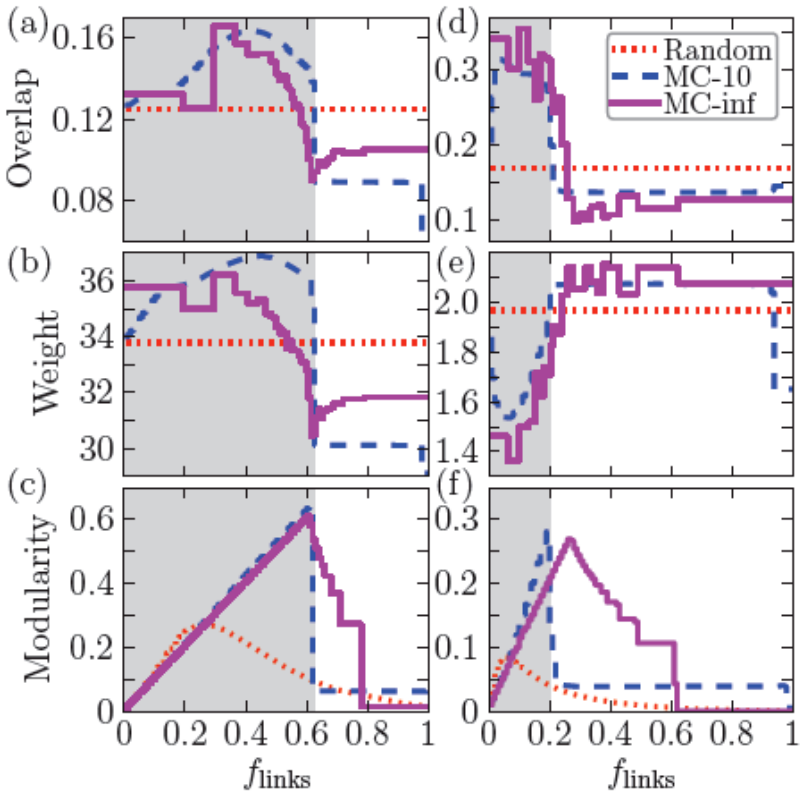


Figure 11. Overlap, weight and modularity as a function of the fraction of links. Top: overlap; middle: average weight; bottom: modularity. Left: mobile call network; right: co-authorship network, MC-10 means m=10, MC-inf means m=infinity.

Figure 12 makes the relation between explosive percolation and community structure clear. In this model networks cliques (obvious communities) are weakly connected to each other.

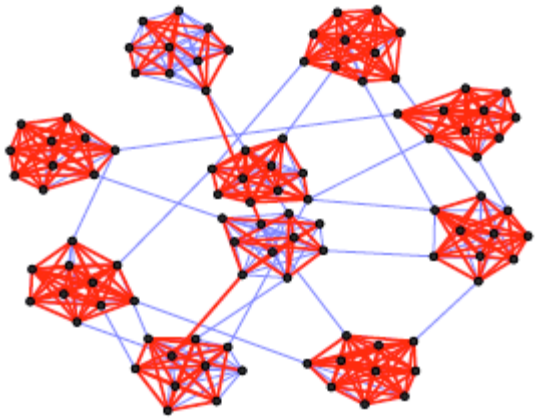


Figure 12. Explosive percolation on a model network (MC-10 rule). Red links are occupied, blue ones unoccupied.

A very important by-product of these studies is that we have shown how finite size scaling on the huge social network of mobile phone calls can be carried out. This is a non-trivial question as we have only one empirical sample, which is extremely inhomogeneous. As we have access to the prescribers' ZIP-codes too, we could study the properties of the network for different sizes of cities. It turned out that this approach leads to a surprisingly good finite size scaling (Fig. 13).

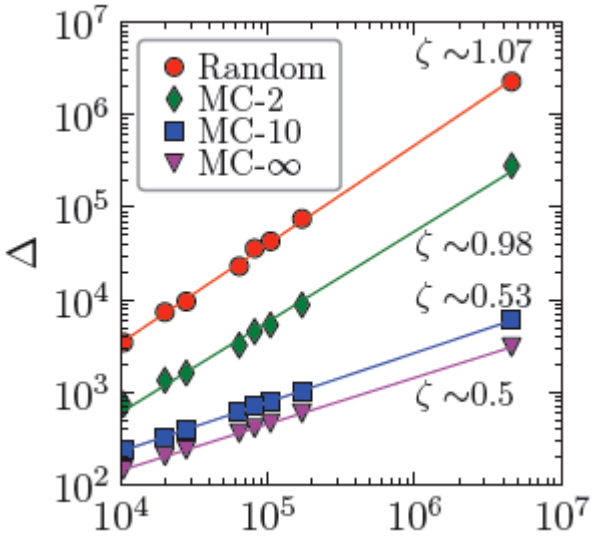


Figure 13. Finite size scaling of the parameter Δ indicating the phase transition. The straightness of the lines in a log-log plot indicates that different size cities can be well used for this purpose.

In conclusion: The study of the explosive percolation model on real world networks has turned out to be very useful. We have uncovered close relationship between the explosive percolation transition and the community structure; we pointed out how the differences in the organizational principle of different social networks are reflected in the properties of the model near to the threshold; we showed that finite scaling can be carried out on mobile call networks by using different size cities as samples.

3. Groups from an egocentric point of view – social signatures and their persistence

This section is based on J. Saramäki, E.A. Leicht, E. Lopez, S. Roberts, F. Reed-Tsochas, R.I.M. Dunbar, *The persistence of social signatures in human communication networks* (manuscript to be submitted, 2011). Work has also been initiated towards a second paper highlighting the differences between communication via text messages and calls.

In order to have a detailed view on the fundamental building blocks of social group structure and the role of mobile communications in maintaining such structure, we have performed a longitudinal study on the data set DS2, where all outgoing phone calls and text messages of 30 students were tracked during their period of transition from high school to university; these record were augmented with self-reported questionnaires on the students' social networks. This work has been conducted as a joint effort between the Aalto and Oxford research groups.

Our main findings can be summarized as follows:

Emotionally close alters are frequently called or texted; thus for large-scale data sets, tie strength is a good proxy of the emotional intensity of a tie and the most important strong ties are well captured by call records. This is in line with the earlier observations of the Social Diary of WP4 (D4.3), where text messages and mobile phone calls, they accounted for 75% of all contacts listed by the participants. Interestingly, when plotting the probability of an alter being called, texted, or called or texted as a function of the emotional closeness to the alter, we see a difference between kin and non-kin alters (Fig 10). For kin alters, text messages seem to play a less important role; this is possibly related to the "generational gap", as text message usage anticorrelates with age. Furthermore, for non-kin alters, it appears that alters with lower emotional intensity are texted more frequently than called; thus the choice of the communication channel is related to the nature of the social tie.

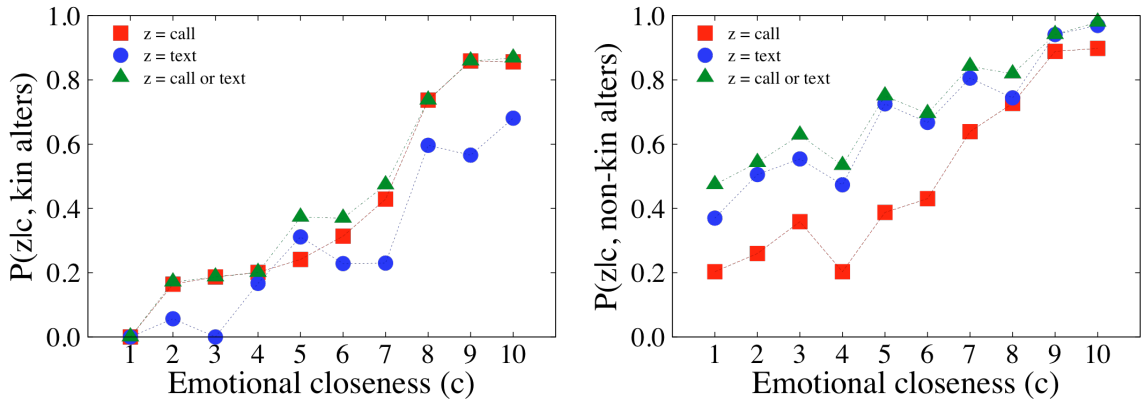


Figure 14: Probability of calling, texting, or calling or texting an alter within the first 6 months of the study period as a function of the self-reported emotional closeness to the alter. Left: kin alters, right: non-kin alters.

When the time allocation patterns of individuals are studied by plotting the fraction of calls/texts going to an alter as a function of the rank of that alter (where the rank is based on the number of calls/texts), we see that the patterns are typically characterized by a broad distribution. This means that a very large number of calls/texts is targeted to a few alters of high emotional closeness, and there is an increasing number of ties associated with fewer calls. However, within this general picture there is still a lot of individual variation; e.g. the fraction of calls/texts going to top alters is far higher than the average for some subjects (see Figure 11).

This observation is broadly in line with the layered network view of the social brain hypothesis, where cognitive constraints are seen to be a limiting factor for the number of high-intensity-high-closeness ties, whereas there are more ties of lower intensity, requiring less maintenance and cognitive efforts.

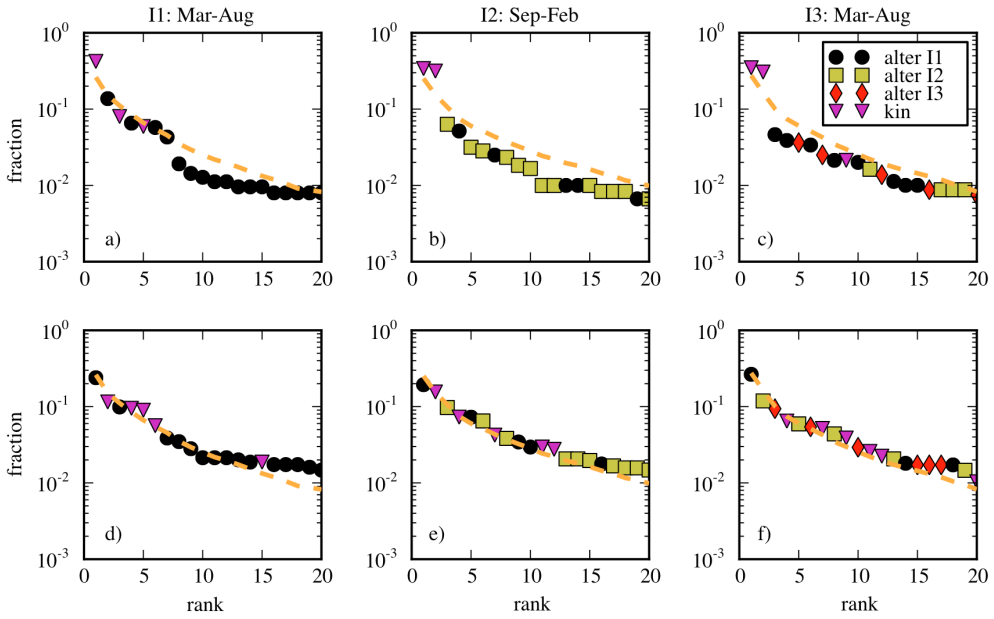


Figure 15. The fraction of calls to an alter as a function of the alter, for two study participants (top and bottom rows) and for three 6-month windows (columns). Symbols denote when alters have first appeared in the networks; from the central and rightmost columns it is evident that the networks undergo major changes. The dashed line indicates average over all participants.

As our data set has been recorded during a period of time where the networks of the participants are undergoing major changes (moving to a different city, first university year, or leaving school and going for a gap year), we detect a large turnover in terms of alters. Despite this large turnover, the time allocation patterns of individuals (measured with calls) are surprisingly persistent during the observation period. That is, if an individual allocates more time to the top-ranking alters than individuals do on average, this time allocation pattern persists even if those alters are replaced by newcomers, i.e. "new best friends". Likewise, if an individual's time allocation pattern is broader and calls are more evenly distributed among alters, this is likely to remain so, even if the composition of the network changes a lot.

For further details, please see the attached draft manuscript, to be submitted for publication in the near future.

References

[1] Kun Zhao, Márton Karsai and Ginestra Bianconi. Entropy of dynamical social networks (submitted).

[2] G. Tibély. Criterions for locally dense subgraphs (*Physica A*, 2011, in press).

[3] G. Tibély, L. Kovanen, M. Karsai, K. Kaski, J. Kertész, J. Saramäki. Communities and beyond: mesoscopic analysis of a large social network with complementary methods. *Phys. Rev. E* **83**, 056125 (2011).

[4] R. K. Pan, M. Kivelä, J. Saramäki, K. Kaski, and J. Kertész. Using explosive percolation in analysis of real-world networks. Phys. Rev. E 83, 046112 (2011).

[5] A.-L. Barabási, The origin of bursts and heavy tails in human dynamics. Nature 435, 207-211 (2005).

[6] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J. F. Pinton, A. Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. PLoS ONE 5:e11596 (2010).

[7] A. Lancichinetti, S. Fortunato, F. Radicchi. Benchmark graphs for testing community detection algorithms. Phys. Rev. E 78, 046110 (2008).

[8] A. Lancichinetti, S. Fortunato. Community detection algorithms: a comparative analysis. Phys. Rev. E 80, 056117 (2009).

[9] D. Achlioptas, R. M. D'Souza, and J. Spencer. Explosive percolation in random networks. Science 323, 1453-1455 (2009).

[10] M. Rosvall, C. Bergstrom. Maps of information flow reveal community structure in complex networks. Proc. Natl. Acad. Sci. USA 105, 1118-1123 (2008).

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre. Fast unfolding of communities in large networks. J. Stat. Mech. P10008 (2008).

[12] G. Palla, I. Derényi, I. Farkas, T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814-818 (2005).

[13] A. Clauset, M. E. J. Newman, C. Moore. Finding community structure in very large networks. Phys. Rev. E 70, 066111 (2004).

[14] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi. Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA 101, 2658-2663 (2004).

[15] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási. Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. USA 104, 7332-7336 (2007).

[16] J. Saramäki, E.A. Leicht, E. Lopez, S. Roberts, F. Reed-Tsochas, R.I.M. Dunbar, The persistence of social signatures in human communication networks (manuscript to be submitted, 2011).

Entropy of dynamical social networks

Kun Zhao [*], **Márton Karsai** [†] **and Ginestra Bianconi** [*]

[*]Physics Department, Northeastern University, Boston MA 02115 USA, and [†]BECS, School of Science, Aalto University, P.O. Box 12200, FI-00076

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Human dynamical social networks encode information and are highly adaptive. To characterize the information encoded in the fast dynamics of social interactions, here we introduce the entropy of dynamical social networks. By analysing a large dataset of phone-call interactions we show evidence that the dynamical social network has an entropy that depends on the time of the day in a typical week-day. Moreover we show evidence for adaptability of human social behavior showing data on duration of phone-call interactions that significantly deviates from the statistics of duration of face-to-face interactions. This adaptability of behavior corresponds to a different information content of the dynamics of social human interactions. We quantify this information by the use of the entropy of dynamical networks on realistic models of social interactions.

entropy | information | social networks | dynamical networks

Networks [1, 2, 3, 4, 5] encode information in the topology of their interactions. This is the main reason why networks are ubiquitous in complexity theory and constitute the underlying structures of social, technological and biological systems. The information encoded in social networks [6, 7] is essential to build strong collaborations [8] that enhance the performance of a society, to build reputation trust and to navigate [9] efficiently the networks. For these reasons social networks are small world [10] with short average distance between the nodes but large clustering coefficient. Therefore to understand how social network evolve, adapt and respond to external stimuli, we need to develop a new information theory of complex social networks.

Recently, attention has been addressed to entropy measures applied to email correspondence [14], static networks [12, 13] and mobility patterns [11]. New network entropy measures quantify the information encoded in heterogenous static networks [12, 13]. Information theory tools set the limit of predictability of human mobility[11]. Still we lack methods to assess the information encoded in the dynamical social interaction networks.

Social networks are characterized by complex organizational structures revealed by network community and degree correlations [15]. These structures are sometimes correlated with annotated features of the nodes or of the links such as age, gender, and other annotated features of the links such as shared interests, family ties or common work locations [16, 17]. In a recent work [18] it has been shown by studying social, technological and biological networks that the network entropy measures can assess how significant are the annotated features for the network structure.

Moreover social networks evolve on many different timescales and relevant information is encoded in their dynamics. In fact social networks are highly adaptive. Indeed social ties can appear or disappear depending on the dynamical process occurring on the networks such as epidemic spreading or opinion dynamics. Several models for adaptive social evolution have been proposed showing phase transitions in different universality classes [19, 20, 21, 22]. Social ties have in addition to that a microscopic structure constituted by fast social interactions of the duration of a phone call or of a face-to-face interaction. Dynamical social networks characterize the social interaction at this fast time scale. For these dynamical networks new network measures are starting to be defined [23] and recent works focus on the implication that the network dynamics has on percolation, epidemic spreading and opinion dynamics [24, 25, 26, 27, 28].

Thanks to the availability of new extensive data on a wide variety of human dynamics [29, 30, 31, 33, 32], human mobility [34, 35, 11] and dynamical social networks [36], it has been recently recognized that many human activities [25] are bursty and not Poissonian. New data on social dynamical networks start to be collected with new technologies such as of Radio frequency Identification Devices [37, 27] and Bluetooth [30]. These technologies are able to record the duration of social interactions and report evidence for a bursty nature of social interaction characterized by a fat tail distribution of the duration of face-to face interactions. This bursty behavior of social networks [38, 39, 37, 27, 40, 41] is coexisting with modulations coming from periodic daily (circadian rhythms) or weakly patterns [42]. The fact that this bursty behavior is observed also in social interaction of simple animals (leeches) [43], in the motion of rodents [44], or in the use of words [45], suggests that the underlying origin of this behavior is dictated by the biological and neurological processes underlying the dynamics of the social interaction. To our opinion this problem remains open: How much can humans intentionally change the statistics of social interactions and the level of information encoded in the dynamics of their social networks, when they are interfacing with a new technology?

In this paper we try to address this question by studying the dynamics of interactions through phone calls and comparing it with face-to-face interactions. We show that the entropy of dynamical networks is able to quantify the information encoded in the dynamics of phone-call interactions during a typical week-day. Moreover we show evidence that human social behavior is highly adaptive and that the duration of face-to-face interaction in a conference follows a different distribution than duration of phone-calls. We therefore have evidence of an intentional capability of humans to change statistically their behavior when interfacing with the technology of mobile phone communication. Finally we develop a model in order to quantify how much the entropy of dynamical networks changes if we allow modifications in the distribution of duration of the interactions.

Results

Entropy of dynamical social networks. In this section we introduce the entropy of dynamical social networks as a measure of information encoded in their dynamics. Since we are

Reserved for Publication Footnotes

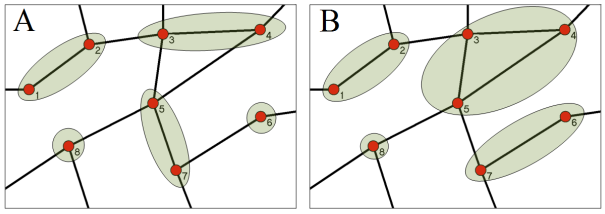


Fig. 1. The dynamical social networks are composed by different dynamically changing groups of interacting agents. In panel (A) we allow only for groups of size one or two as it typically happens in mobile phone communication. In panel (B) we allow for groups of any size as in face-to-face interactions.

interested in the dynamics of contacts we assume to have a quenched social network G of friendships, collaborations or acquaintances formed by N agents and we allow a dynamics of social interactions on this network. If two agents i, j are linked in the network they can meet and interact at each given time giving rise to the dynamical social network under study in this paper. If a set of agents of size N is connected through the social network G the agents $i_1, i_2, \dots i_n$ can interact in a group of size n. Therefore at any given time the static network G will be partitioned in connected components or groups of interacting agents as shown in Fig . In order to indicate that a social interaction is occurring at time t in the group of agents $i_1, i_2, \dots, i_n$ and that these agents are not interacting with other agents, we write $g_{i_1,i_2,\dots,i_n}(t) = 1$ otherwise we put $g_{i_1,i_2,\dots,i_n}(t) = 0$. Therefore each agent is interacting with one group of size $n > 1$ or non interacting (interacting with a group of size $n = 1$). Therefore at any given time

$$\sum_{\mathcal{G}=(i,i_2,\dots,i_n)|i\in\mathcal{G}} g_{i,i_2,\dots,i_n}(t) = 1. \qquad [1]$$

where we indicate with $\mathcal{G}$ an arbitrary connected subgraph of G. The history $\mathcal{S}_t$ of the dynamical social network is given by $\mathcal{S}_t = \{g_{i_1,i_2,\dots,i_n}(t')\,\forall t' < t\}$. If we indicated by $p(g_{i,i_2,\dots,i_n}(t) = 1|\mathcal{S}_t)$ the probability that $g_{i,i_2,\dots,i_n}(t) = 1$ given the story $\mathcal{S}_t$, the likelihood that at time t the dynamical networks has a group configuration $g_{i_1,i_2,\dots,i_n}(t)$ is given by

$$\mathcal{L} = \prod_{\mathcal{G}} p(g_{i,i_2,\dots,i_n}(t) = 1|\mathcal{S}_t)^{g_{i_1,i_2,\dots,i_n}(t)} \qquad [2]$$

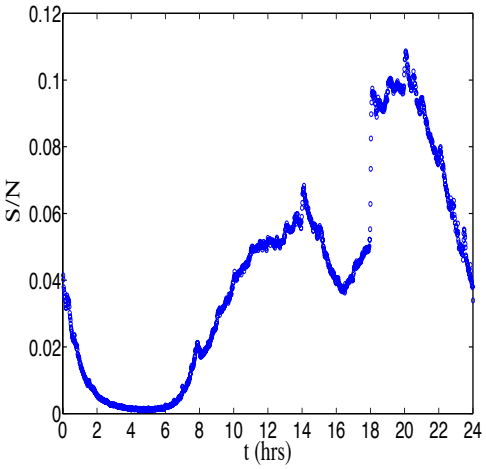


Fig. 2. Mean-field evaluation of the entropy of the dynamical social networks of phone calls communication in a typical week-day. In the nights the social dynamical network is more predictable.

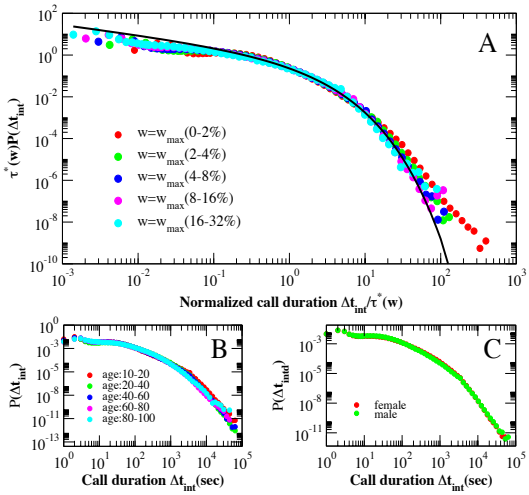


Fig. 3. (A) Probability distribution of duration of phone-calls between two given persons connected by a link of weight w. The data depend on the typical scale $\tau^\star(w)$ of duration of the phone-call. (B) Probability distribution of duration of phone calls for people of different age. (C) Probability distribution of duration of phone-calls for people of different gender. The distributions shown in the panel (B) and (C) do not significantly depend on the attributes of the nodes.

The entropy S characterizes the logarithm of the typical number of different group configurations that can be expected in the dynamical network model at time t and is given by $S = -\langle \log \mathcal{L} \rangle_{|\mathcal{S}_t}$ that we can explicitly express as

$$S = -\sum_{\mathcal{G}} p(g_{i,i_2,\dots,i_n}(t) = 1|\mathcal{S}_t) \log p(g_{i,i_2,\dots,i_n}(t) = 1|\mathcal{S}_t). \qquad [3]$$

According to the information theory results [46], if the entropy is vanishing, i.e. $S = 0$ the network dynamics is regular and perfectly predictable, if the entropy is larger the number of future possible configurations is growing and the system is less predictable. If we model face-to-face interactions we have to allow the possible formation of groups of any size, on the contrary, if we model the mobile phone communication, we need to allow only for pairwise interactions. Therefore, if we define the adjacency matrix of the network G as the matrix a_{ij}, the log likelihood takes the very simple expression given by

$$\mathcal{L} = \prod_i p(g_i(t) = 1|\mathcal{S}_t)^{g_i(t)} \prod_{ij|a_{ij}=1} p(g_{ij}(t) = 1|\mathcal{S}_t)^{g_{ij}(t)} \qquad [4]$$

with

$$g_i(t) + \sum_j a_{ij} g_{ij}(t) = 1, \qquad [5]$$

for every time t. The entropy is then given by

$$\begin{aligned} S \;=\; & -\sum_i p(g_i(t) = 1|\mathcal{S}_t) \log p(g_i(t) = 1|\mathcal{S}_t) \\ & -\sum_{ij} a_{ij} p(g_{ij}(t) = 1|\mathcal{S}_t) \log p(g_{ij}(t) = 1|\mathcal{S}_t). \end{aligned} \qquad [6]$$

Social dynamics and entropy of phone call interactions. We have analyzed the call sequence of subscribers of a major euroepan mobile service provider. In the dataset the users were anonymized and impossible to track. We considered calls between users who at least once called each other during the examined 6 months period in order to examine calls only reflecting trusted social interactions. The resulted event list

consists of 633, 986, 311 calls between 6, 243, 322 users. For the entropy calculation we selected 562, 337 users who executed at least one call per a day during a week period. First of all we have studied how the entropy of this dynamical network is affected by circadian rhythms. We assign to each agent $i = 1, 2$ a number $n_i = 1, 2$ indicating the size of the group where he/she belongs. If an agent i has coordination number $n_i = 1$ he/she is isolated, and if $n_i = 2$ he/she is interacting with a group of $n = 2$ agents. We also assign to each agent i the variable t_i indicating the last time at which the coordination number n_i has changed. If we neglect the feature of the nodes, the most simple transition probabilities that includes for some memory effects present in the data, is given by a probability $p_n = p_n(\tau, t)$ for an agent in state n at time t to change his/her state given that he has been in his/her current state for a duration $\tau = t - t_i$.

We have estimated the probability $p_n(\tau, t)$ in a typical week-day. Using the data on the probabilities $p_n(\tau, t)$ we have calculated the entropy, estimated by a mean-field evaluation (Check Supporting Information for details) of the dynamical network as a function of time in a typical week-day. The entropy of the dynamical social network is reported in Fig. . It significantly changes during the day describing the fact that the predictability of the phone-call networks change as a function of time. In fact, as if the entropy of the dynamical network is smaller and the network is an a more predictable state.

Adaptive dynamics face-to face interactions and phone call durations. In this section we report evidence of adaptive human behavior by showing that the duration of phone calls, a binary social interactions mediated by technology, show different statistical features respect to face-to-face interactions. The distributions of the times describing human activities are typically broad [29, 25, 38, 31, 37, 27], and are closer to power-laws, which lack a characteristic time scale, than to exponentials. In particular in [37] there is reported data on Radio Frequency Identification devices, with temporal resolution of 20s, showing that both distribution duration of face-to-face contacts and inter-contact periods is fat tailed during conference venues.

Here we analysed the above defined mobile-call event sequence performing the measurements on all the users for the entire 6 months time period. The distribution of phone-call durations strongly deviates from a fat-tail distribution. In Fig. we report this distributions and show that these distributions depend on the strength w of the interactions (total duration of contacts in the observed period) but do not depend on the age, gender or type of contract in a significant way. The distribution $P^w(\Delta t_{in})$ of duration of contacts within agents with strenght w is well fitted by a Weibull distribution

$$\tau^*(w)P^w(\Delta t_{in}) = W_\beta\left(x = \frac{\Delta t}{\tau^\star(w)}\right) = \frac{1}{x^\beta}e^{-\frac{1}{1-\beta}x^{1-\beta}}. \quad [7]$$

with $\beta = 0.47...$ The typical times $\tau^*(w)$ used for the data collapse of Figure 3 are listed in Table 1. The origin of this significant change in behavior of humans interactions could be due to the consideration of the cost of the interactions (although we are not in the position to draw these conclusions (See Fig. 11 in which we compare distribution of duration of calls for people with different type of contract) or might depend on the different nature of the communication. The duration of a phone call is quite short and is not affected significantly by the circadian rhythms of the population. On the contrary the duration of no-interaction periods is strongly affected by periodic daily of weekly rhythms. The distribution

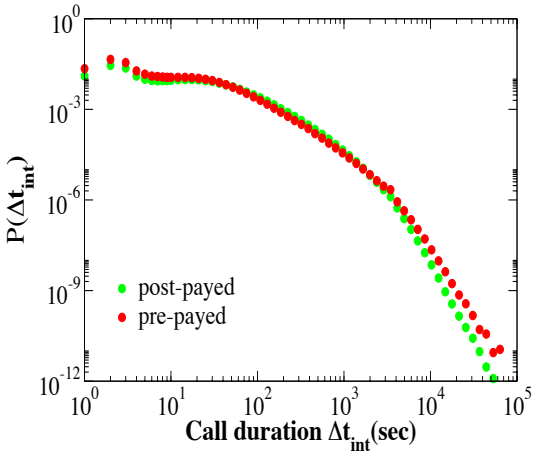


Fig. 4. Probability distribution of duration of phone-calls for people with different types of contract. No significant change is observed that modifies the functional form of the distribution.

of no-interaction periods can be fitted by a double power-law but also a single Weibull distribution can give a first approximation to describe $P(\Delta t_{no})$. In Fig. 5 we report the distribution of duration of no-interaction periods in the day periods between 7AM and 2AM next day. The typical times $\tau^*(k)$ used in Figure 5 are listed in Table 2.

Discussion

The entropy of a realistic model of cell-phone interactions. The data on face-to-face and mobile-phone interactions show that a reinforcement dynamics is taking place during the human social interaction. Disregarding for the moment the effects of circadian rhythms and weakly patterns, a possible explanation of such results is given by mechanisms in which the decisions of the agents to form or leave a group are driven by memory effects dictated by reinforcement dynamics, that can be summarized in the following statements: *i) the longer an agent is interacting in a group the smaller is the probability that he/she will leave the group; ii) the longer an agent is isolated the smaller is the probability that he/she will form a new group.* In particular, such reinforcement principle implies that the probabilities $p_n(\tau, t)$ that an agent with coordination number n changes his/her state depends on the time elapsed since his/her last change of state, i.e., $p_n(\tau, t) = f_n(\tau)$. To ensure the reinforcement dynamics any function $f_n(\tau)$ which is a decreasing function of its argument can be taken. In two recent papers[40, 41] the face-to-face interactions have been realistically modelled with the use of the reinforcement dynamics, by choosing

$$f_n(\tau) = \frac{b_n}{(\tau + 1)}. \quad [8]$$

with good agreement with the data when we took $b_n = b_2$ for $n \geq 2$ and $b_1 > 0$, $b_2 > 0$.

In order to model the phone-call data studied in this paper we can always adopt the reinforcement dynamics but we need to modify the probability $f_n(\tau)$ by a parametrization with an additional parameter $\beta \leq 1$. In order to be specific in our model of mobile-phone communication, we consider a system that consists of N agents. Corresponding to the mechanism of daily cellphone communication, the agents can call each other to form a binary interaction if they are neighbor in the social network. The social network is characterized by a given degree distribution $p(k)$ and a given weight distribution

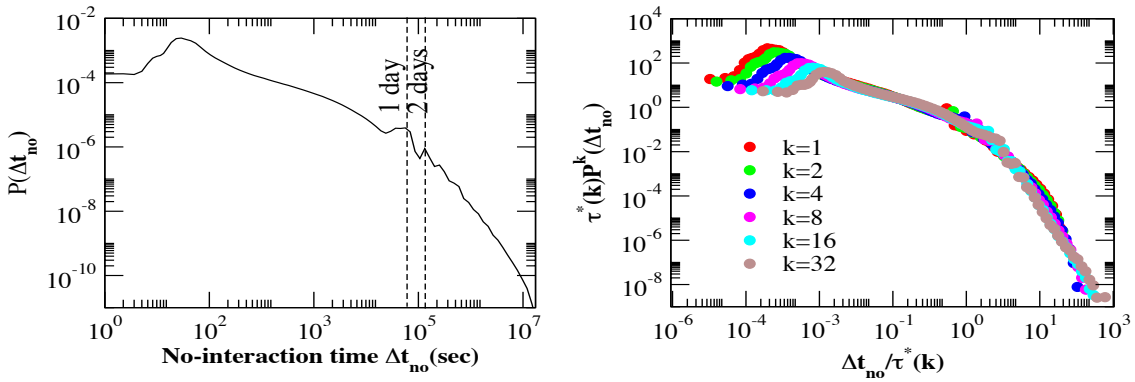


Fig. 5. Distribution of non-interaction times in the phone-call data. The distribution strongly depends on circadian rhythms. The distribution of rescaled time depends strongly on the connectivity of each node. Nodes with higher connectivity k are typically non-interacting for a shorter typical time scale $\tau^\star(k)$.

$p(w)$. Each agent i is characterized by the size $n_i = 1, 2$ of the group he/she belongs to and the last time t_i he/she has changed his/her state. Starting from random initial conditions, at each timestep $\delta t = 1/N$ we take a random agent. If the agent is isolated he/she will change his/her state with probability

$$f_1^\beta(\tau) = \frac{b_1}{(\tau+1)^\beta} \qquad [9]$$

with $\tau = t - t_i$ and $b_1 > 0$. If he/she change his/her state he/she will call one of his/her neighbor in the social network which is still not engaged in a telephone call. A non-interacting neighbor agent will pick up the phone with probability $f_1^\beta(\tau')$ where τ' is the time he/she has not been interacting.

If, on the contrary the agent i is interacting, he/she will change his/her state with probability $f_2^\beta(\tau|w)$ depending on the weight of the link and on the duration of the phone call. We will take in particular

$$f_2^\beta(\tau|w) = \frac{b_2 g(w)}{(\tau+1)^\beta} \qquad [10]$$

where $b_2 > 0$ and $g(w)$ is a decreasing function of the weight w of the link. The distributions $f_1^\beta(\tau)$ and $f_2^\beta(\tau|w)$ are parametrized by the parameter $\beta \leq 1$. As β increases, the distribution of duration of contacts and duration of inter-contact time become broader. These probabilities give rise to either Weibull distribution of duration of interactions (if $\beta < 1$) or power-law distribution of duration of interaction $\beta = 1$. Indeed for $\beta < 1$, the probability $P_2^w(\tau)$ that a conversation between two nodes with link weight w ends after a duration τ is given by the Weibull distribution (See Supporting Information for the details of the derivation)

$$\tau^*(w) P_2^w(\tau) \propto W_\beta((\tau+1)/\tau^\star(w)) \qquad [11]$$

with $\tau^\star(w) = [2b_2 g(w)]^{-1/(1-\beta)}$. This distribution well capture the distribution observed in mobile phone data and reported in Fig. (for a discussion of the validity of the annealed approximation for predictions on a quenched network see the Supporting Information.)

If, instead of having $\beta < 1$ we have $\beta = 1$ the probability distribution for duration of contacts is given by a power-law

$$P_2^w(\tau) \propto (\tau+1)^{-[2b_2 g(w)+1]}. \qquad [12]$$

This distribution is comparable with the distribution observed in face-to-face interaction during conference venues [40, 41]. The adaptability of human behavior, evident when comparing the distribution of duration of phone-calls with the duration of face-to-face interactions, can be understood as a possibility to change the exponent β regulating the duration of social interactions.

Changes in the parameter β correspond to a different entropy of the dynamical social network. Solving analytically this model we are able to evaluate the dynamical entropy as a function of β and b_1. In Fig. 13 we report the entropy S of the dynamical social network a function of β and b_1 in the annealed approximation and the large network limit. In particular we have taken a network of size $N = 2000$ with exponential degree distribution of average degree $\langle k \rangle = 6$, weight distribution $P(w) = Cw^{-2}$ and function $g(w) = 1/w$ and $b_2 = 0.05$. Our aim in Fig. 13 is to show only the effects on the entropy due to the different distributions of duration of contacts and non-interaction periods. Therefore we have normalized the entropy S with the entropy S_R of a null model of social interactions in which the duration of groups are Poisson distributed but the average time of interaction and non interaction time are the same as in the model of cell-phone communication. From Fig. 13 we observe that if we keep b_1 constant, the ratio S/S_R is a decreasing function of the parameter β indicating that the broader are the distribution of probability of duration of contacts the higher is the information encoded in the dynamics of the networks. Therefore the heterogeneity in the

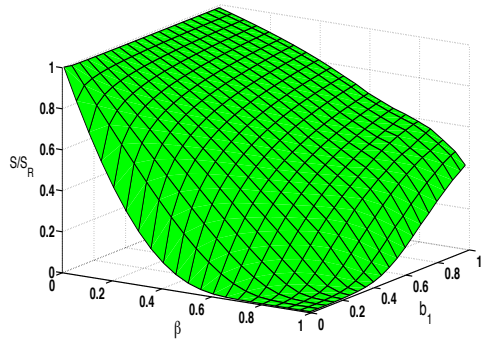


Fig. 6. Entropy S of social dynamical network model of pairwise communication normalized with the entropy S_R of a null model in which the expected average duration of phone-calls is the same but the distribution of duration of phone-calls and non-interaction time are Poisson distributed. The network size is $N = 2000$ the degree distribution of the network is exponential with average $\langle k \rangle = 6$, the weight distribution is $p(w) = Cw^{-2}$ and $g(w)$ is taken to be $g(w) = b_2/w$ with $b_2 = 0.05$. The value of S/S_R is depending on the two parameters β, b_1. For every value of b_1 the normalized entropy is smaller for $\beta \to 1$.

distribution of duration of contacts and no-interaction periods implies higher level of information in the social network. The human adaptive behavior by changing the exponent β in face-to-face interactions and mobile phone communication effectively change the entropy of the dynamical network.

Conclusions

In the last ten years it has been recognized that the vast majority of complex systems can be described by networks of interacting units. Network theory has made tremendous progresses in this period and we have gained important insight into the microscopic properties of complex networks. Key statistical properties have been found to occur universally in the networks, such as the small world properties and broad degree distributions. Moreover the local structure of networks has been characterized by degree correlations, clustering coefficient, loop structure, cliques, motifs and communities. The level of information present in these characteristic of the network can be now studied with the tools of information theory. An additional fundamental aspect of social networks is their dynamics. This dynamics encode for information and can be modulated by adaptive human behavior. In this paper we have introduced the entropy of social dynamical networks and we have evaluated the information present in dynamical data of phone-call communication. By analysing the phone-call interaction networks we have shown that the entropy of the network depends on the circadian rhythms. Moreover we have shown that social networks are extremely adaptive and are modified by the use of technologies. The statistics of duration of phone-call indeed is described by a Weibull distribution that strongly differ from the distribution of face-to-face interactions in a conference. Finally we have evaluated how the information encoded in social dynamical networks change if we allow a parametrization of the duration of contacts mimicking the adaptability of human behavior. Therefore the entropy of social dynamical networks is able to quantify how the social networks dynamically change during the day and how they dynamically adapt to different technologies.

Material and Methods

In order to describe the model of mobile phone communication, we consider a system consisting of N agents representing the mobile phone users. The agents are interacting in a social network G representing social ties such as friendships, collaborations or acquaintances. The network G is weighted with the weights indicating the strength of the social ties between agents. We use $N_1^k(t_0,t)dt_0$ to denote the number of agents with degree k that at time t are not interacting and have not interacted with another agent since time $t' \in (t_0, t_0+1/N)$. Similarly we denote by $N_2^{k,k',w}(t_0,t)dt_0$ the number of connected agents (with degree respectively k and k' and weight of the link w) that at time t are interacting in phone call started at time $t' \in (t_0, t_0+1/N)$. The mean-field equation for this model read,

$$\begin{aligned}\frac{\partial N_1^k(t_0,t)}{\partial t} &= -(1+ck)N_1^k(t_0,t)f_1(t_0,t) + N\pi_{21}^k(t)\delta_{tt_0} \\ \frac{\partial N_2^{k,k',w}(t_0,t)}{\partial t} &= -2N_2^{k,k',w}(t_0,t)f_2(t_0,t|w) + N\pi_{12}^{k,k',w}(t)\end{aligned} \quad [13]$$

where the constant c is given by

$$c = \frac{\sum_{k'}\int_0^t dt_0 N_1^{k'}(t_0,t)f_1(t_0,t)}{\sum_{k'}k'\int_0^t dt_0 N_1^{k'}(t_0,t)f_1(t_0,t)}. \quad [14]$$

In Eqs. (13) the rates $\pi_{pq}(t)$ indicate the average number of agents changing from state $p=1,2$ to state $q=1,2$ at time t. These rates can be also expressed in a self-consistent way and the full system solved for any given choice of $f_1(t_0,t)$ and $f_2(t_0,t|w)$ (See Supporting Information for details).

The definition of the entropy of dynamical social networks of a pairwise communication model, is given by Eq. (6). To evaluate the entropy of dynamical social network explicitly, we have to carry out the summations in Eq. (6). These sums, will in general depend on the particular history of the dynamical social network, but in the framework of the model we study, in the large network limit will be dominated by their average value. In the following therefore we perform these sum in the large network limit. The first summation in Eq. (6) denotes the average loglikelihood of finding at time t a non-interacting agent given a history $\mathcal{S}_t$. We can distinguish between two eventual situations occurring at time t: *(i)* the agent has been non-interacting since a time $t-\tau$, and at time t remains non-interacting; *(ii)* the agent has been interacting with another agent since time $t-\tau$, and at time t the conversation is terminated by one of the two interacting agents.The second term in the right hand side of Eq. (6), denotes the average loglikelihood of finding two agents in a connected pair at time t given a history $\mathcal{S}_t$. There are two possible situations that might occur for two interacting agents at time t: *(iii)* these two agents have been non-interacting, and to time t one of them decides to form a connection with the other one; *(iv)* the two agents have been interacting with each other since a time $t-\tau$, and they remain interacting at time t. Taking into account all these possibilities we have been able to use the transition probability form different state and the number of agents in each state to evaluate the entropy of dynamical networks in the large network limit (For further details on the calculation see the Supporting Information).

ACKNOWLEDGMENTS. We thank A.-L. Barabási for his useful comments and for the mobile call data used in this research. MK acknowledges the financial support from EUs 7th Framework Programs FET-Open to ICTeCollective project no. 238597

1. Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: From biological nets to the Internet and WWW (Oxford Univ Press, Oxford).
2. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:157256.
3. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Phys Rep 424:175308.
4. Caldarelli G (2007) Scale-Free Networks (Oxford Univ Press, Oxford).
5. Barrat A, Barthélemy M, Vespignani A (2008) Dynamical processes on complex networks (Cambridge Univ Press, Cambridge).
6. Granovetter M (1973) The strength in weak ties. Am J Sociol 78:13601380.
7. Wasserman S, Faust K (1994) Social Network Analysis: Methods and applications (Cambridge Univ Press, Cambridge).
8. Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98:404-409.
9. Kleinberg JM (2000) Navigation in a small world. Nature 406:845.
10. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393:440-442.
11. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of Predictability in Human Mobility. Science 327:1018-1021.
12. Bianconi G (2008) The entropy of randomized network ensembles. Europhys Lett 81:28005.
13. Anand K, Bianconi G (2009) Entropy measures for networks: Toward an information theory of complex topologies. Phys Rev E 80:045102.
14. Eckmann JP, Moses E, Sergi D (2004) Entropy of dialogues creates coherent structures in e-mail traffic. Proc Natl Acad Sci USA 101:14333.
15. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Rev Mod Phys 81:591-646.

16. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. Nature 446:664-667.
17. Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466:761-764.
18. Bianconi G, Pin P, Marsili M (2009) Assessing the relevance of node features for network structure. Proc Natl Acad Sci USA 106:11433-11438.
19. Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks. Phys Rev Lett 88:128701.
20. Marsili M, Vega-Redondo F, Slanina F (2004) The rise and fall of a networked society: A formal model. Proc Natl Acad Sci USA 101:1439-1442.
21. Holme P, Newman MEJ (2006) Nonequilibrium phase transition in the coevolution of networks and opinions. Phys Rev E 74:056108. (2006).
22. Vazquez F, Eguíluz VM, San Miguel M (2008) Generic Absorbing Transition in Coevolution Dynamics. Phys Rev Lett 100:108702.
23. Tang J, Scellato S, Musolesi M, Mascolo C, Latora V (2010) Small-world behavior in time-varying graphs. Phys Rev E 81:055101.
24. Holme P (2005) Network reachability of real-world contact sequences. Phys Rev E 71:046119.
25. Vázquez A, Rácz B, Lukacs A, Barabàsi AL (2007) Impact of Non-Poissonian activity patterns on spreading processes. Phys Rev Lett 98:158702.
26. Parshani R, Dickison M, Cohen R, Stanley HE, Havlin S (2010) Dynamic networks and directed percolation. Europhys Lett 90:38004.
27. Isella L, Stehlé J, Barrat A, Cattuto C, Pinton JF, Van den Broeck W (2011) What's in a crowd? Analysis of face-to-face behavioral networks. J Theor Biol 271:166-180.
28. Karsai M, Kivelä M, Pan R K, Kaski K, Kertész J, Barabási A-L, Saramäki J (2011) Small but slow world: How network topology and burstiness slow down spreading Phys Rev E 83:025102.
29. Barabási AL (2005) The origin of bursts and heavy tails in humans dynamics. Nature 435:207-211.
30. Eagle N, Pentland AS (2006) Reality mining: sensing complex social systems. Personal Ubiquitous Comput 10:255-268.
31. Rybski D, Buldyrev SV, Havlin S, Liljeros F, Makse HA (2009) Scaling laws of human interaction activity. Proc Natl Acad Sci USA 106:12640-12645.
32. Malmgren RD, Stouffer DB, Campanharo ASLO, Nunes Amaral LA (2009) On universality in human correspondence activity. Science 325:1696-1700.
33. Malmgren RD, Stouffer DB, Motter AE, Amaral LA (2008) A poissonian explanation for heavy tails in e-mail communication. Proc Natl Acad Sci USA 105:18153-18158.
34. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439:462-465.
35. González MC, Hidalgo AC, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453:779-782.
36. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabàsi AL (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104:7332-7336.
37. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, Vespignani A (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks. PLoS ONE 5:e11596.
38. Hui P, Chaintreau A, Scott J, Gass R, Crowcroft J, Diot C (2005) Pocket switched networks and human mobility in conference environments. Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking (Philadelphia, PA) pp 244-251.
39. Scherrer A, Borgnat P, Fleury E, Guillaume JL, Robardet C (2008) Description and simulation of dynamic mobility networks. Comp Net 52:2842-2858.
40. Stehlé J, Barrat A, Bianconi G (2010) Dynamical and bursty interactions in social networks. Phys Rev E 81:035101.
41. Zhao K, Stehlé J, Bianconi G, Barrat A (2011) Social network dynamics of face-to-face interactions. Phys Rev E 83:056109.
42. Jo HH, Karsai M, Kertész J, Kaski K (2011) Circadian pattern and burstiness in human communication activity. arXiv:1101.0377.
43. Bisson G, Bianconi G, Torre V (in preparation).
44. Anteneodo C, Chialvo DR (2009) Unraveling the fluctuations of animal motor activity. Chaos 19:033123.
45. Altmann EG, Pierrehumbert JB, Motter AE (2009) Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. PLoS ONE 4:e7678.
46. Cover T and Thomas JA (2006) Elements of Information Theory (Wiley-Interscience,Hoboken).

Table 1. Typical times $\tau^\star(w)$ used in the data collapse of Fig. .

Weight of the link	Typical time $\tau^\star(w)$ in seconds (s)
(0-2%) w_{max}	111.6
(2-4%) w_{max}	237.8
(4-8%) w_{max}	334.4
(8-16%) w_{max}	492.0
(16-32%) w_{max}	718.8

Table 2. Typical times $\tau^\star(k)$ used in the data collapse of Fig. 5.

Connectivity	Typical time $\tau^\star(k)$ in seconds (s)
k=1	158,594
k=2	118,047
k=4	69,741
k=8	39,082
k=16	22,824
k=32	13,451

Criterions for locally dense subgraphs

Gergely Tibély

Institute of Physics and HAS-BME Cond. Mat. Group,
Budapest University of Technology and Economics,
Budapest, Budafoki str. 8., H-1111

Abstract

Community detection is one of the most investigated problems in the field of complex networks. Although several methods were proposed, there is still no precise definition of communities. As a step towards a definition, I highlight two necessary properties of communities, separation and internal cohesion, the latter being a new concept. I propose a local method of community detection based on two-dimensional local optimization, which I tested on common benchmarks and on the word association database.

1 Introduction

In the last decade interdisciplinary research on complex networks resulted in spectacular development [1]-[6]. It has become clear that networks constructed from diverse complex systems show remarkably similar features. Several aspects were investigated, like clustering [7], the degree distribution[8], diameter [9], [10], spreading processes [11], diffusion [12], synchronization [13], critical phenomena [14] and game theoretical models on complex networks [15].

One of the most actively researched questions about complex networks is the one of community detection [16]. Community detection aims at finding dense groups in graphs, like circles of friends in social networks, web pages about the same topic, or substances appearing in the same pathway in metabolic reaction networks. Perhaps the strongest motivation behind the research is that dense groups in the topology are expected to correspond to functions performed by the network, such that one can infer from pure topology to function. While the concept of communities seems intuitively plausible, attempts for an algorithmically useful definition have not been successful yet. The global characterization by modularity [17] or by random walks [19, 18], the local "weak" and "strong" definitions [20], the clique percolation approach [21], or the multiresolution methods [22, 23, 24] have all increased out understanding of this complex problem but the proliferation of methods of community detection just indicates the difficulty of this issue [16].

Unfortunately, any precise definition of communities is still lacking, giving rise to innumerable methods using different definitions. Lack of a definition also makes problematic the testing of methods; although there is progress in this issue [25], [26]. Difficulty of the problem is increased by more subtle factors: very

often communities occur on a broad scale, they can be ordered in a hierarchical manner, and they may overlap, which make their identification even harder.

After being the subject of active research for several years, it is getting clear that the following stages appear during community detection:

1 defining the term "community";

2 finding the objects corresponding to the definition;

3 determining the significance of the found communities.

Although from the theoretical perspective stage 1 is clearly a key issue, it is far from being settled. Several different propositions exists, which are evaluated mostly according to their results on a few benchmarks. This is the stage to be improved in the first place in this paper. Stage 2 is a technical issue, often consisting of some combinatorial optimization method. Its choice is usually a result of a trade-off between speed and quality. Stage 3 should give information about how surprising is the existence of a found community in the actual graph, given some characteristics of the graph like edge density or degree distribution. Although this issue also got some attention [27]-[33], it just began to get widespread application [34].

The rest of the paper will focus on the question of definition, so a few remarks about stage 3 are made here. Most community detection methods give no information about the significance of their output, thus forcing the investigator to assume that all results are (equally) significant. This way, the community detection stages 2 and 3 are combined into a single decision whether a particular subgraph is a good enough community or not – effectively pruning the significance test in practice. The other end of the spectrum, represented by [34], builds the definition of communities on statistical significance, which is clearly an improvement. However, it should be noted that the fitness and statistical significance of a subgraph as a community are not synonyms. Statistical significance tells us how surprising a subgraph is, while fitness talks about how close is it to the ideal community. Therefore, the two quantities are complementary and both belong to the description of a community.

2 Local criteria for communities

A fundamental problem of community detection is to define the term "community". There are different approaches to this question. One is the algorithmic approach, giving a computational procedure for finding clusters. This naturally incorporates a mathematically precise definition, although different algorithms usually result in diverse definitions, and there is no theoretical framework currently to help their differentiation. Another possibility is to present a general concept, on which a precise definition can be based. In this paper, the latter approach is taken, although an algorithmic realization is also presented.

No definition of communities which is both precise and generally accepted has appeared yet. Currently the description of communities exhausts in the phrase "nodes having more edges among themselves than to the rest of the graph" (or equivalent forms). It can be translated roughly to "statistically significant locally dense subgraphs". Statistical significance is a quite precise expression, the main problem is with the term "locally dense". For an intuitive

picture, it is quite good, but much less than directly transformable to algorithms. Although there is an implicit agreement on that clearly counterintuitive results are not permitted, even a formal list of required properties is missing. However, there are some properties which fit human intuition about locally dense subgraphs[1][2]:

Separation: a good community is well-separated from the rest of the graph;

Cohesion: a good community is homogeneously well connected inside, i.e. it is hard to separate into two communities.[3]

The separation criterion is quite clear, although there is an important remark: separation should be defined locally, involving only the community under investigation and its immediate neighborhood. Global methods, in which distant regions of the graph can modify a community in order to improve a global fitness value, can produce results violating the human perception about clusters. A famous example is the resolution limit of modularity [35, 36].

Although separation is a very intuitive criterion, and famous methods rely on it (see the Appendix), it is not enough in itself. Figs. 1a-1b illustrate that the distribution of links inside the separated region (the "shape" of the subgraph) also matters heavily. Application of current community detection methods to real-world networks confirms that this is a real problem, e.g. tree-like communities can occur, even when the whole network is not tree-like [37], [38].

Both separation and cohesion are required properties of communities. If one neglects cohesion, the result may contain clusters like the one on Fig. 1a. On the other hand, if separation is not taken into account, one may end up chopping a separated subgraph until very cohesive pieces (cliques in the extreme) are obtained, like the triangles on Fig. 1b.

Given the subgraphs on Figs 1a and 1b as proposed communities, most community detection methods' fitness values, to be reviewed in the next section, can not tell the difference between them. This is due to that most methods simply count the internal and/or external edges, which do not tell about the distribution of those edges. The reason why several methods do not fail to assign proper clusters for Fig 1a is that they look for optimal clusters, consequently they compare configurations like Fig 1a in one cluster and in two clusters, and splitting the two cliques into two clusters may improve the partition. But the situation is even worse. In the next Section, we will see that a number of fitness functions are more optimal for a counterintuitive clustering than for the intuitive one (e.g. joining the two cliques on Fig 1a, like modularity for a large enough graph). It should be noted that in such a case, the proper communities might be recovered if the heuristic gets stuck in the proper local optimum, even when that is not the global optimum.

[1]For brevity, the words "community", "group" and "cluster" will be used from this point as synonyms for "locally dense subgraph", omitting the statistical significance from the meaning.

[2]It should be noted that the meaning of the term "community" can depend on the context; consequently a single definition may not be enough. Here the aim is to describe a particularly intuitive one.

[3]The term "cohesion" also appeared in [22], although there it denotes a quantity with an unrelated concept.

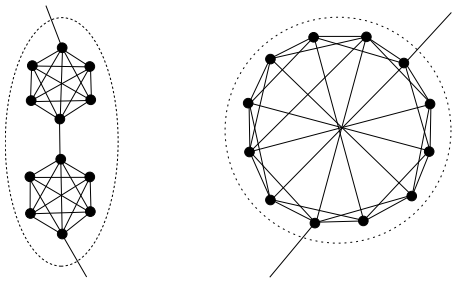


Figure 1: Illustration of the importance of subgraph shape. The two subgraphs have the same number of nodes and the same degrees, i.e. they differ only in the distribution of links. The figure on the left is much less cohesive than the figure on the right, although just a reorganization was applied to the links.

3 Overview of the existing methods

Here, existing community detection methods will be reviewed from the point of view of the previous section, i.e. do they conform the criterions of separation and cohesion. As mini-benchmarks, the examples on Fig. 1 or their simple variations will be used (see the Appendix for details on individual methods). The desired output for 1a is two communities consisting of the two cliques, while 1b should be kept in one piece. In both cases, no nodes from the rest of the graph should be included. For methods optimizing a fitness function, the globally optimal solution will be considered, for other methods, the possible solutions. These solutions will be compared to the desired ones, independently for Fig. 1a and 1b. If a method separates the two cliques of Fig. 1a, then it gets a "+", if it puts all nodes of Fig. 1b into one cluster, then it gets another "+". If there are multiple equally valid solutions (like for label propagation), all solutions are required to conform the preferred result.
For methods optimizing a function, the heuristic realizing the optimization may deviate from the global optimum, presenting worse or even better results (in terms of conformity to separation and cohesion). This will not be investigated, here the focus is on the definition of the communities (following from the choice of the fitness function), not on the practical aspects. Results for methods which can produce a single partition or cover are displayed in Table 1. The large number of published methods makes assembling a complete list nearly impossible. Instead, the emphasis is put on the diversity of the reviewed approaches.

There is a bunch of multiresolution methods, which possess a parameter allowing to tune the cluster sizes from 1 (isolated nodes) to $\mathcal{O}(N)$: the multiresolution modularity of Reichardt and Bornholdt (RB) [22], of Arenas, Fernández and Gómez (AFG) [23], the local fitness method of Lancichinetti, Fortunato and Kertész (LFK) [39], the Potts model of Ronhovde and Nussinov (RN) [43], the Markov autocovariance stability of Delvenne, Yaliraki and Barahona (MAS) [19], the hierarchical likelihood method of Clauset, Moore and Newman (CNM) [56], and the Markov Cluster Algorithm of van Dongen (MCL) [57]. Naturally, these methods are expected to find the proper community assignments both to

method	cohesion test (like Fig. 1a)	separation test (like Fig. 1b)
Lancichinetti et al. [39]	-	+
Labelpropagation [40]	-	-
Infomap [18]	-	+
Clique Percolation [21]	-	-
Estrada & Hatano [41]	-	-
Modularity optimization [17]	-	+
Donetti & Muñoz [42]	-	+
Ronhovde & Nussinov [43]	-	-
Nepusz et al. [44]	+	-
Hofman & Wiggins [45]	-	-
Hastings [46]	-	-
Newman & Leicht [47]	+	-
Wang & Lai [48]	+	-
Bickel & Chen [49]	+	-
Karrer & Newman [50]	+	-
Infomod [51]	-	+
Radicchi et al. [20]	+	-
Chauhan et al. [52]	+	-
Evans & Lambiotte [53]	-	+
Ahn et al. [54]	-	-
ModuLand [55]	-	-

Table 1: Cohesion & separation criterion test results. Tests were done on Fig. 1a and 1b or similar graphs (which are described in the Appendix). + and - are assigned according to whether the fitness function of a method is more optimal for the preferred solution or not. For methods which do not optimize a fitness function, simply the possible solution(s) was (were) analyzed. See the Appendix for details on specific methods.

Fig. 1a and 1b at some parameter values. However, there is no guarantee that these values are also the proper ones for the rest of the graph. Consequently, it is not clear how a resolution parameter should be set: the natural idea is to find the longest interval of the resolution parameter value in which the community structure does not change, but when the optimal parameter value is different for different regions in the graph, the longest stable interval not necessarily reflects the optimal communities.

Furthermore, the fitness values do not help us to tell good clusters from bad ones, like Fig. 1a from Fig. 1b. For most multiresolution methods (RB, AFG, LFK, RN), it is very easy to see that the fitnesses of two clusters are the same given that all nodes has the same in- and outdegrees, independently of the shape of the clusters. Note that it is also true for most single resolution methods. For MAS it is not trivial. Therefore, empirical tests were conducted to check it. According to them, Fig. 1a was found empirically to be at least as good as

1b[4]. Finally, regarding MCL and CNM, they have no fitness function[5], the only accessible quantity about the community structure is the parameter interval in which it is stable.

Finally, there are hierarchical methods, which look for series of smaller and smaller (or larger and larger) clusters hierarchically embedded into the previous ones. Similarly to multiresolution methods, they are expected to contain good clusters in the outputted hierarchy. However, when looking at a graph having a simple one-level community structure, the question how to select the proper levels of the outputted hierarchy arises. The easiest way is to use the lowest level communities. Unfortunately, it is not a reliable procedure, as the lowest-level clusters may be just parts of the communities of the optimal partition or cover (see the Appendix for details). A second idea can be to assign significance scores to the communities on different levels, in the spirit of [29]. Although this approach might reliably qualify the found communities, a new version of statistical significance taking into account the internal cohesion is required. Furthermore, one should be very careful not to impose unnecessary constraints, like prohibiting overlaps, when constructing a hierarchical method.

A further question is whether a method provides information about the shape of the found communities or not. Recent analysis of real-world networks highlights the relevance of this issue [37], [38]. Several methods are based on simply counting the internal and/or external edges, or degrees at most: LFK, Labelpropagation, Infomap, modularity optimization (and equivalents), Hofman & Wiggins, Hastings, Ronhovde & Nussinov, Newman & Leicht, Wang & Lai, Bickel & Chen, Karrer & Newman, Infomod, Ahn et al., OSLOM. Consequently, they do not see any difference in the distribution of the links, e.g. Fig 1a and 1b get the same fitness values. Only Clique Percolation and Radicchi et al.'s[6] method have some very limited requirement about cohesion built in the definition of communities.

The conclusion is that none of the reviewed methods is able to successfully apply both the separation and the cohesion criterions. They susceptible either to glue together well-separated subgraphs or to overpartition a cohesive subgraph. Future network designs should consider cohesion as well as separation.

4 Community detection in a two dimensional parameter space

In this Section, a new method for community detection is introduced. Its main goal is to present a method which takes into account both criterions defined in Sec. 2. First, the LFK method will be reviewed, which will serve as a starting point for the new method. Then, a composite fitness will be constructed which takes into account the separation and cohesion criterions. Finally, a heuristic optimization procedure for the composite fitness will be described, which finds

[4]In this case, only 1 link to the rest of the graph was used. Rest of the graph, represented by a single node having self-loops, was assigned 118 edges inside, resulting in a total of $L = 150$ edges. Stability values were calculated from 0.01 to 100, the step size being 0.01 below 1.0 and 1 above.

[5]CNM does have a fitness function, but it corresponds to a full hierarchical dendrogram, not to any partitions obtained by cutting the dendrogram at some point

[6]If the stopping criterion of their heuristic is considered as part of the definition.

locally dense subgraphs on all scales, and also able to recover hierarchical structures.

4.1 The LFK method

The LFK method [39] optimizes the local fitness function

$$f^C = \frac{K_{in}^C}{(K_{in}^C + K_{out}^C)^\alpha} \tag{1}$$

where C denotes a subgraph, K_{in}^C and K_{out}^C are the total number of inside and outside degrees in C, respectively, and α is a tunable exponent for setting the size scale of the communities to be found. Running the method with large α values result in small clusters, small values in large clusters. The recommended range for α is 0.5-2.

The practical implementation of the optimization works as follows. The communities are found one-by-one, independently of each other. First, a seed node is selected from which the new community will be grown. Then, the node which can best improve the fitness of the cluster is added. This addition is repeated until the fitness reaches a local optimum. After each addition, removal of nodes takes place, if the fitness can be enhanced that way. When the fitness cannot be further increased, the actual subgraph is declared a community. The growth process is repeated for all nodes as seeds, or alternatively, until the found communities cover all nodes in the graph.

Although the resolution parameter α can be tuned continuously, [39] suggested that the relevant community structures should be identified by robustness to changes in α, i.e. which have the longest interval for α values without change. Changes in the community structure were detected by monitoring the mean fitness of the communities, evaluated at a reference value $\alpha = 1$.

4.2 Implementing the criterions

For the separation criterion, the following function will be applied

$$f_S^C = \frac{K_{in}^C}{K_{in}^C + K_{out}^C} \tag{2}$$

where C is a subgraph, K_{in} and K_{out} are the sums of in-community and out-community degrees, respectively. This is the fitness of LFK [39], with the multiresolution parameter being set to one. For detecting hierarchical structures, a different solution will be described. Eq. 2 clearly focuses on the external separation of the clusters, therefore it is suitable as an implementation of the first criterion of the communities.

For the internal cohesion criterion, a possible solution is to consider the second eigenvalue of the Laplacian matrix of the community. The Laplacian of a graph is the matrix $L = A - D$, where A is the adjacency (or weight) matrix, and $D = \mathrm{diag}(k_i)$ is a diagonal matrix containing the degrees (strengths). Its largest eigenvalue is always 0 (corresponding to the trivial eigenvector $(1, 1, \ldots, 1)$). The multiplicity of the largest eigenvalue equals to the number of connected components in the graph. This gives the hint that if two distinct graphs are got connected by a single (weak) link, the Laplacian gets only a slight perturbation

(compared to the case of two connected components), which splits the double degeneracy of the first eigenvalue, such that a new eigenvalue close to zero appears[7]. In fact, it is known that the second eigenvalue of the Laplacian measures "how difficult is to split the graph into two large pieces" [58].

For some important special cases the second eigenvalue can be calculated:

- for full graphs of n nodes (clique), $\lambda_2 = -n$
- for a star-graph, $\lambda_2 = -1$, independently of n
- for a linear chain, $\lambda_2 = -2 + 2\cos(\pi/n) \to 0$ as $n \to \infty$
- for two n-sized cliques attached by a single link (having weight ϵ) (like on Fig. 1a), $\lambda_2 \approx -\frac{2\epsilon}{n+2\epsilon}$, which also goes to 0 as $n \to \infty$.
- for a disconnected graph, $\lambda_2 = 0$. This may seem trivial, but most methods give a finite score for disconnected communities; it is not without precedent that such objects can be produced in reality [38]. Although this problem can be avoided by a properly designed heuristic of a method, disconnected communities should be punished by definition.

Calculation for the two cliques is in the Appendix, other results can be found in [59]. These cases confirm that the second eigenvalue is useful for quantifying the cohesion criterion of the definition of communities. For an illustration, on Fig. 2 a few example graphs with their second Laplacian eigenvalues are shown.

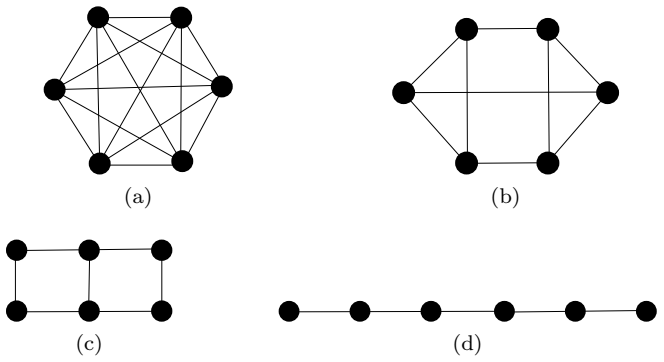


Figure 2: Graphs with different second Laplacian eigenvalues. $\lambda_2^{(a)} = 6$, $\lambda_2^{(b)} = 2$, $\lambda_2^{(c)} = 1$, $\lambda_2^{(d)} = 0.268$. The maximal value of λ_2 is 6 in all cases.

The separation fitness term f_S ranges from zero to one. In order to compose it together with the cohesion fitness, the latter should also be in the interval $[0, 1]$. Therefore, λ_2 needs some transformations before application as fitness. As can be seen from the above examples, for the worst cases $|\lambda_2|$ is of the order of $1/n$, therefore the lowest point of the $|\lambda_2|$-scale will be set to $1/n$. The highest point is trivially given by n. It is reasonable to assume that most subgraphs have $|\lambda_2| = \mathrm{o}(|C|)$. Furthermore, several subgraphs can have worse internal cohesion

[7]The diffusion matrix was also considered, but it prefers star-like graphs too much.

than the star graph, thus having $|\lambda_2| \in [0,1]$. To take into account these effect, $\log|\lambda_2|$ will be more useful than λ_2. So, in order to obtain a quantity between 0 and 1, the minimum will be subtracted and divided by the maximum,

$$\begin{aligned} f_C^C =& \frac{\log|\lambda_2| - \log 1/|C|}{\log|C| - \log 1/|C|} = \frac{1}{2} + \frac{1}{2}\frac{\log|\lambda_2|}{\log|C|}, & \text{if } |C| > 1 \\ =& 0 & \text{if } |C| = 1 \end{aligned} \tag{3}$$

where $|C|$ is the number of nodes in the community. The above measure happens to be 0.5 if λ_2 is 1, e.g. for the star-graph. I wish to emphasize that eq. 3 is only one possible proposition for taking into account the internal cohesion, although a promising one – better measures may exist. The same is true for the choice of f_S.

The cohesion fitness f_C^C opens the way for constructing tests assessing the performance of community detection methods regarding the cohesion of the found communities. One may generate a graph with built-in communities which separation is controlled, like in the LFR benchmark [25], then randomly select pairs of clusters and increase the interconnection between the two members of each pairs to some predefined value, finally calculating f_C^C of the pairs. Running the detection method and measuring the ratio of pairs not split as a function of f_C^C may indicate how strongly focuses the method on cohesion.

The next question is how to combine $f_{separation}^C$ and $f_{cohesion}^C$. Thinking in a two dimensional space of f_S^C and f_C^C, a natural approach is to get as far from the point $(0,0)$ as possible. This implies

$$f^C = \sqrt{(f_S^C)^2 + (f_C^C)^2} \tag{4}$$

so the fitness is the euclidean distance from $(0,0)$. Again, this is just one possibility, better combinations may exist. E.g. the relative weight of f_S and f_C may be adjusted in a more well-grounded way. However, eq. 4 is able to pass the test raised by Fig. 1: for Fig. 1a, $\lambda_2^{\text{2cliques}} = 0.258$, $f_C^{\text{2cliques}} = 0.228$, $f^{\text{2cliques}} = 0.995$ while for a single clique $\lambda_2^{\text{1clique}} = 6$, $f_C^{\text{1clique}} = 1$, $f^{\text{1clique}} = 1.371$. For Fig. 1b, $\lambda_2^{\text{12nodes}} = 3.268$, $f_C^{\text{12nodes}} = 0.738$, $f^{\text{12nodes}} = 1.218$, and for the best subgraph, a triangle, $\lambda_2^{\text{triangle}} = 3$, $f_C^{\text{triangle}} = 1$, $f^{\text{triangle}} = 1.077$.

Beyond enabling one to decide whether a given subgraph is a community or not (by requiring local optimality), the above definition makes it possible to assess how good community it is. This is also possible with another definitions, e.g. by using the modularity function, but here, communities are placed on a 2-dimensional space instead of 1 dimension. This gives rise to an interesting possibility for characterizing the communities, like "very cohesive but densely connected outwards" or "well-separated but poorly interconnected". Considering Fig. 1a, one may think that the latter is not really a community. But for large subgraphs, it may make sense to consider a well-separated subgraph as a community, as common sense says that large communities should be looser than small ones.

4.3 Community detection in reality

In this section, the details of practical implementation of the new method are discussed. Most importantly, in order to actually find the communities, a heuristic

carrying out the optimization of eq. 4 is needed. Furthermore, there is a second problem of detecting communities hierarchically embedded into each other. These two questions will be answered by a common solution.

The heuristic is based on the one of the LFK method [39]. Among its details, the LFK heuristic contains a tunable parameter (denoted as α), which is claimed to be able to recover communities at different hierarchical levels. Lowering this parameter α results in increased community sizes. Hierarchical levels are supposed to be stable against the variation of α, so there should be long intervals for α for which the communities do not change. However, large graphs may lack long stable intervals, as some changes occur around any parameter value (data not shown). Therefore, a new method for investigating hierarchical structure is needed. I dropped the idea of using threshold values of α, corresponding to community structures at different scales, which should be simultaneously valid for all communities, and I will treat each community separately.

Similarly to [39], each community is grown from a seed node. It is important to note that each seed node can result in a series of (successively larger) communities. Growth consists of successively including the neighboring node which increases most the fitness defined by eq. 4. When there is no neighboring node which inclusion can improve the fitness, the stage of node removal begins. Here, the fitness of the cluster is tried to be improved by excluding nodes from it (with the exception of the seed node, which is not permitted to be excluded). It finishes when no further removal can improve the fitness. Then, growth begins again, if possible. The grow-shrink cycle is iterated, as long as the fitness can be improved. When no improvement is possible (there is a local optimum of the fitness), the actual list of nodes is registered as a valid community. After that, the algorithm tries to find a larger community, which contains the current one. This way, hierarchical structures can be revealed. In order to do it, first the growing cluster should escape from the basin of attraction of the current local optimum. Therefore, the cluster is forced to grow, by successively including the neighboring nodes which decrease the fitness the least. After some steps of forced growth, when increasing the fitness becomes again possible, the algorithm turns back to the normal grow-shrink procedure, until a new local optimum is found, signing a new community. The cluster keeps hopping from local optimum to another local optimum until it grows so large that it contains the whole graph. Then a new growth process starts from a new seed node. At the end of its growth process, it includes the whole graph again, unless it encounters a local optimum which has been already found, i.e. the corresponding community has already been registered. In this case, the growth process is stopped. Then, another growth process starts from a not-yet-used seed node. In contrast to [39], all nodes in the graph are used as seed nodes, in order not to miss good communities. When the growth process beginning from the last seed node finishes, the algorithm ends, and the registered communities are written to the output.
There are a few additional tricks. First, if escaping from a local optimum seems to be hard, i.e. after changing from forced growth to the normal grow-shrink stage we still end up in the previous local optimum, the cluster is restored to the state where it had its maximal size (the beginning of one of the removal sessions), then 2 steps of forced growth is applied before the normal grow-shrink cycle begins. A second trick is that when judging the identity of two communities, they are considered identical if at least 80% of the larger community is

a subset of the smaller one[8]. In case of identity, the community which has the higher fitness is kept in the registry.
The algorithm, although based on the one of [39], differs in several points: from one seed, several communities can be reached instead of only the smallest one; node removal occurs when node addition is not possible instead of after each addition (this trick also speeds up the algorithm); seed node is not permitted to be removed; all nodes are used as seeds instead of the not-yet-covered nodes. An algorithm similar in spirit was described in [60]. The results in the next section are obtained using this method, unless stated otherwise explicitly. The software realizing the algorithm is available at
`http://www.phy.bme.hu/~tibelyg/`.

5 Test results

Probably the most frequently used test is Zachary's karate club friendship network[61]. Due to a dispute between two prominent persons (node 1 and 34), the club split into two during sociological observation, and the memberships in the new clubs are known. As the split occurred more or less along a border of two visible communities, new community detection algorithms are usually claimed to pass the test if they reproduce the split. However, the aim is the detection of *topological* modules, not functional ones, so the result of the sociological study is not a strict criterion for judging the output of any community detection method. E.g., node 10 has $1-1$ links to each of the new clubs, so "misplacing" it (compared to the split) may not be considered as a fault. Or node 12, which attaches only to node 1, is hard to be considered as part of a "densely interconnected" cluster.
The algorithm finds 33 groups, containing several non-relevant ones, like pairs of nodes. Therefore, a filtering procedure is required. The statistical significance of the resulting communities [29, 34] is utilized for this purpose. The statistical significance can be sensitive for missing nodes [29], therefore each cluster is allowed to be completed with the neighboring node which optimizes the statistical significance. Then the clusters are ordered according to their statistical significance. The first 3 clusters provide a single-level community structure, corresponding to 3 known communities, with 2 overlapping and 1 homeless nodes (Fig. 3, left panel). Taking a look at the subsequent clusters provides information about the multi-scale structures in the graph. The next few clusters reveal cluster cores and hierarchical decomposition of the network (Fig. 3, right panel). The statistical significance score is quite capable of distinguishing meaningful structures; there is a gap between 0.42 and 0.81, so setting a threshold to 0.5 selects the multi-scale clusters which would be approved by a human investigator. There is only one exception, the almost-full-clique of nodes {1, 2, 3, 4, 8, 14} has significance 0.81, which is probably the consequence of neglecting the internal cohesion by the current form of statistical significance.

The currently most advanced class of benchmarks was introduced by [25]. In these so-called LFR benchmarks, the network size and edge density are freely adjustable, and more importantly, the node degrees and the community sizes are distributed according to power-law distributions, with tunable exponents.

[8]If the criterion were based on some percent of the smaller group, subset-superset pairs would be considered identical.

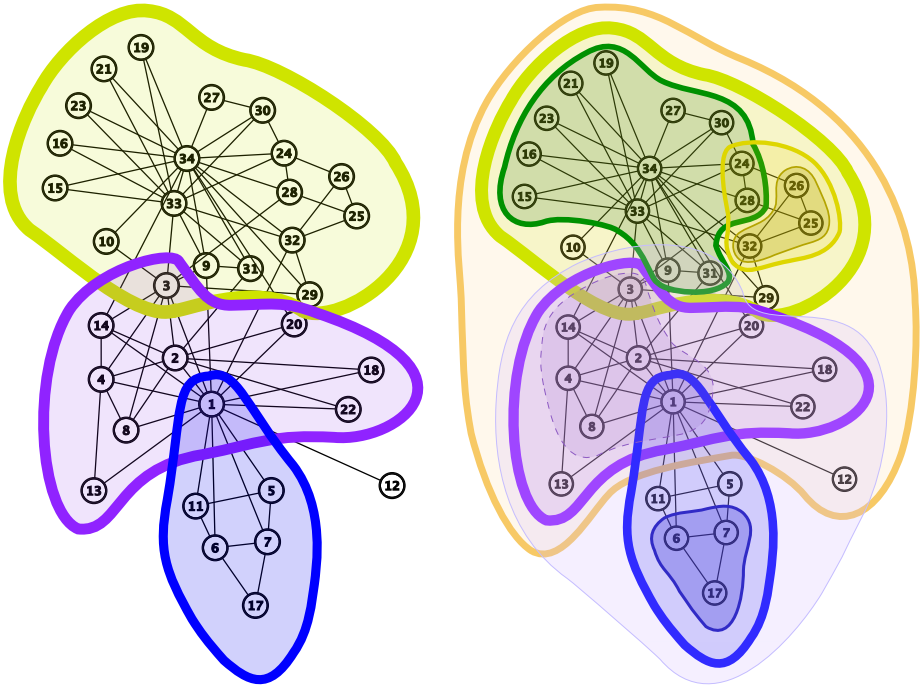


Figure 3: (Color online) The 3 (left) and 10 (right) best found communities of the Zachary karate club. On the right, thicknesses of lines indicate the ordering of the statistical significance values (running from 0.002 to 0.42, plus 0.81 for the dashed line-bordered community). Note that node 12 is contained only by large communities.

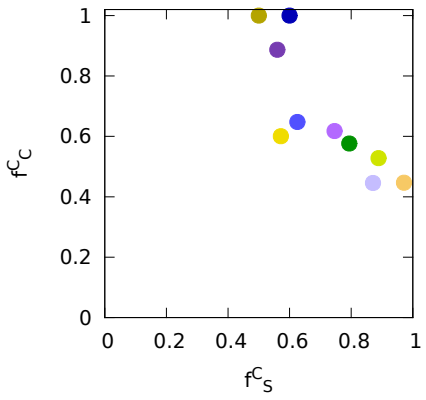


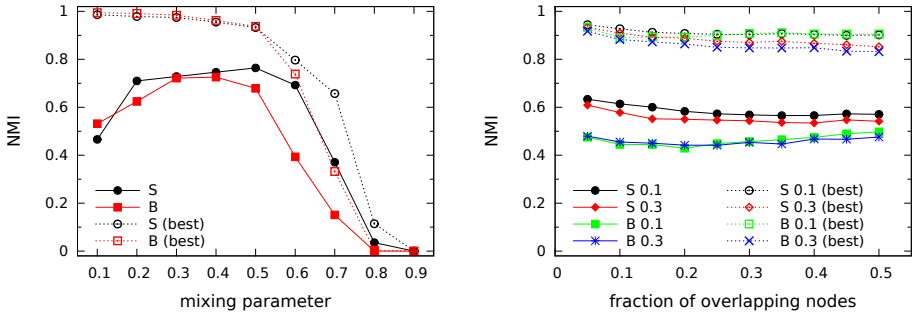
Figure 4: (Color online) Positions of the Zachary communities on the f_S-f_C plane. Small groups tend to cluster at North, and large groups at East.

Communities are defined through a prescribed ratio of inter-community links for each node (mixing ratio, μ), similarly to the preceding GN benchmark class [17]. Generalizations for weighted and directed networks, and for overlapping communities also exist [26].
A wide-scale comparison of different community detection methods using the LFR benchmark was done by [62]. For the ease of comparison, the parameter values of [62] are applied here: the networks consist of 1000 nodes, the average degree is 20, the maximal degree is 50, the exponent of the degree distribution is -2 and the exponent of the community size distribution is -1. There are two types of networks, for the S type the community sizes are between 10 and 50 ("small") and for the B type they are between 20 and 100 ("big"). In [62], networks of 5000 nodes were also investigated. Due to the large computational time, they are omitted here[9]. Also for computational time considerations, the detecting algorithm stopped growing the communities over a predefined size, 120 for the S case and 220 for the B case. All measurement values are obtained from runs on 10 different networks.
Similarity of the built-in and the obtained community structures are quantified by a variant of the normalized mutual information (NMI), which is able to handle overlapping communities [39]. This is the similarity measure applied by [62][10].
Selecting the most relevant communities from the abundant output was done similarly to the previous case. The clusters were completed by 1 neighboring node, if that improved the statistical significance, and sorted with respect to the statistical significance scores. The clusters containing at least 1 uncovered node were accepted one by one until all nodes were covered.
To see the potential of the new method, and check the effect of the output-filtering, the communities corresponding best to the built-in original ones were also selected from the algorithm's output. The results are plotted on Fig. 5 (a). The filtered results are similar to the ones of the lower performing algorithms in [62], while optimal selection provides much better scores, although still not as good as the best methods. The large difference between the optimal and the statistical significance-based results is quite surprising, especially in the light of the fact that statistical significance in itself is able to provide excellent results on the LFR benchmark [34].
The algorithm was also tested on networks with overlapping communities. In this case, clusters having significance score below 0.1 were accepted, similarly to [34]. Fig. 5 (b) shows that the effect of the imperfect output-filtering is again very large, an ideal selection scheme would allow very good results. This is not surprising, as other algorithms based on the one of [39] also give excellent results on overlapping communities [63].

Finally, the new method was applied to a word association graph built from the University of South Florida Free Association Norms [64]. Here, nodes are words and edges show that some people associated the corresponding two words.

[9] it does not mean that a single 5000-sized graph is too large, however, a few hundred of them are

[10] Both the LFR benchmark and the generalized normalized mutual information are freely available from the authors' websites, `http://sites.google.com/site/santofortunato/inthepress2` and `http://sites.google.com/site/andrealancichinetti/software`



(a) (b)

Figure 5: (Color online) Results on the LFR benchmark. Panel (a) corresponds to unweighted, undirected and non-overlapping tests, while panel (b) corresponds to overlapping tests. Overlapping tests were done at two different values of the mixing parameter, at $\mu = 0.1, 0.3$. For both panels: full symbols and lines correspond to the applied filtering, empty symbols with dotted lines correspond to perfect output filtering.

The network has 5018 nodes with mean degree $\langle k \rangle = 22.0$. It is a frequently used example of overlapping community structure [34], [21]. Although edge weights are accessible, the algorithm was applied to the unweighted version of the network. As an illustration, low-level communities around the word *bright* are plotted on Fig. 6. An interesting effect is the appearance of *overlapping edges*, due to the heavy overlap in the network.

In conclusion, although selecting the relevant communities from the output is not an already solved task, the algorithm gives good results on the Zachary karate club, and performs reasonably on the LFR benchmarks. It should be noted however, that due to the internal cohesion criterion, this algorithm's output is not intended to perfectly match benchmarks like GN and LFR, which define communities solely on the basis of external separation. An additional observation is reported here: on GN benchmark graphs[11] with nodes having exactly the prescribed in- and out-degrees, at large mixing ratios communities deviating from the built-in ones but having better-than-designed mixing ratios were found. Note that the new method does not optimize just for external separation, so even better "spontaneous" communities may exist. This phenomenon, although not being a huge surprise, raises the question how to judge precisely a community detection method's output at large mixing ratios, as the known community structure may not be trusted to 100%.

6 Discussion & Conclusions

An important aspect of all community detection methods is the *running time*. In the case of the new method described above, the time requirement is as follows. Starting a new community from each node contributes a factor of N to the CPU

[11] results are omitted, as the presented LFR benchmark is a generalization of the GN.

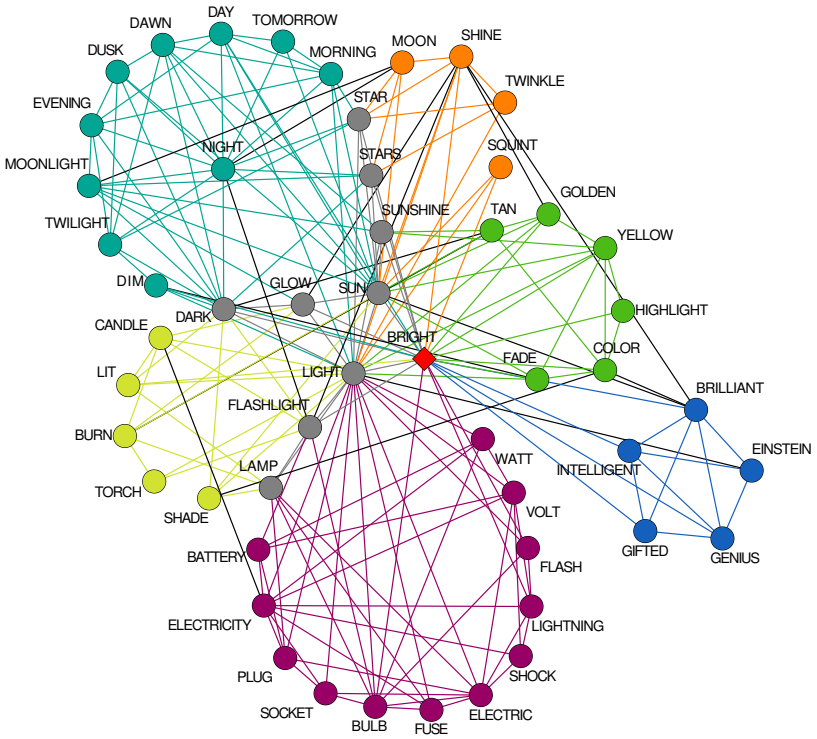


Figure 6: (Color online) Communities around *bright*, on the first hierarchical level. Color denotes communities. Gray shows overlapping nodes and edges. Black edges are between different communities.

time. Evaluating the eigenvalues of a community C plus one extra node takes $2/3\,(|C|+1)^3$. Assuming that C has $\mathrm{const}\cdot\langle k\rangle\cdot|C|$ neighboring nodes (i.e., on average, each node has a constant fraction of its neighbors outside C), running time can be estimated as

$$T \approx N\cdot\sum_{|C|=1}^{|C|_{\max}} \mathrm{const}\cdot\langle k\rangle\cdot|C|\cdot\ (|C|+1)^3 \approx N\cdot\mathrm{const'}\cdot\langle k\rangle\cdot|C|_{\max}^5 \qquad (5)$$

A naive estimate for $|C|_{\max}$ would be N. However, as more and more community growing processes finish, the newly started communities are expected to terminate in a previously discovered community earlier and earlier, on average. Of course, some communities will reach $|C|=N$. Therefore,

$$T\propto N^{5+\delta},\qquad \delta\in[0,1] \qquad (6)$$

which is huge and clearly denies the analysis of even medium-sized graphs ($\mathcal{O}(10^4)$ nodes) without further improvements. Note that graphs of thousands of nodes may be manageable, like the word association graph shown above, which took 56 hours on a single CPU. One possibility is to choose the initial seed more intelligently, starting communities from promising seeds. [63] achieved good results in this aspect. An intelligent seed selection is also important if the number

of communities in a cover is larger than N, or if some communities have only overlapping nodes – in this case, it may happen that all growth processes miss a certain community.
Other important question is the applicability of an advanced eigenvalue solver. Arpack++ [65] and SLEPc [66] were tried. The experience was that – despite their good asymptotic performance in the large matrix limit – for the occurring several small subgraphs the overhead of these complicated machineries was so large that made the final running time much higher than those obtained with the QR-decomposition algorithm.
Doing optimization in a multi-parameter space is a nontrivial task, because different parameters can lie in different ranges. Therefore, an important direction for future research is to investigate the best combination of the parameters in the fitness function, based on the evaluation of empirical data.
Finally, filtering the relevant communities from the found ones is also a challenging task. The natural approach is to apply statistical significance, which should be applied even if filtering was not needed. However, deciding the threshold significance value is not necessarily trivial in all cases. Furthermore, the current form of statistical significance accounts only for the separation of the community, not for its internal cohesion. This manifests itself e.g. in the low score of the almost-full-clique subgraph in the Zachary karate club (the dark purple group on Fig. 3). As the main advantage of the fitness function of eq. 4 is the inclusion of cohesion, it would be important to develop a statistical significance taking it into account.

Conclusions. The community detection problem currently suffers from two fundamental deficiencies. First, there is no definition of community which is precise enough to allow constructing community finding methods. Second, thorough testing a proposed algorithm is problematic, not independently from the previous difficulty. I attempted to improve both issues.

In this paper, I proposed a formal list of required properties for locally dense subgraphs, taking a step towards an applicable definition of the term "community". Two properties, external separation and internal cohesion ("shape") were named. External separation has already been applied by some of the community detection methods, and also by benchmarks. Internal cohesion was not considered explicitly earlier. No current method was found which satisfactorily applies both criterions. I demonstrated on simple examples that both properties are necessary; discarding either of them leads to counterintuitive results. Beyond allowing to construct new methods, these two criterions can also be used as a basis for testing existing ones. They also allow the characterization of a community by two independent quantities, instead of a single scalar.

I proposed a new composite fitness function which takes the two criterions into account. For the quantification of the internal cohesion, the second eigenvalue of the Laplacian matrix is applied, which provides appropriate results on characteristic graphs like cliques or chains. I also proposed a heuristic, by redesigning the LFK heuristic [39], which can find overlapping locally dense subgraphs of all scales, producing much less output than multiresolution methods but with less restrictions than imposed by assuming a hierarchical structure. Runs on the Zachary network and LFR benchmarks showed that the method is able to provide the expected results. Overlapping communities can be detected especially efficiently, similarly to other LFK-based heuristics [63]. However, significant improvements are yet to be implemented; e.g. reducing the running

time, finding a more effective filtering procedure for the output, or fine-tuning the relative weight of the separation and the cohesion terms in the fitness function.

7 Acknowledgments

I wish to thank János Kertész for several useful suggestions. I also thank the Eötvös University for the access to its HPC cluster, and Andrea Lancichinetti, who proposed the Arpack++ package and was extremely helpful about his software. Thanks are due to the authors of the OSLOM [34], LFR benchmark [25], overlapping mutual information [39], Radatools [67] and linegraph-creator [53] softwares for making their code publicly available. I am grateful for the referees for several comments which significantly improved the manuscript. Financial support from EU's 7th Framework Program's FET-Open to ICTeCollective project no. 238597 is acknowledged.

Appendix

A Second eigenvalue of two weakly connected cliques

Assume two cliques of n nodes, edge weights are 1. The two cliques are attached by a single edge having weight ϵ. Then the eigenvalue equations for the Laplacian matrix are

$$\sum_{\substack{j<n \\ j\neq i}} x_j + x_n - (n-1+\lambda)x_i = 0 \qquad \forall i < n \tag{7}$$

$$\sum_{\substack{k>n \\ k\neq i}} x_k + x_{n+1} - (n-1+\lambda)x_i = 0 \qquad \forall i > n+1 \tag{8}$$

$$\sum_{j<n} x_j + \epsilon \cdot x_{n+1} - (n-1+\epsilon+\lambda)x_i = 0 \qquad i = n \tag{9}$$

$$\sum_{k>n+1} x_k + \epsilon \cdot x_n - (n-1+\epsilon+\lambda)x_i = 0 \qquad i = n+1 \tag{10}$$

Adding the last two equations gives

$$\sum_j x_j - x_n - x_{n+1} + \epsilon x_n + \epsilon x_{n+1} - (n-1+\epsilon+\lambda)x_n - (n-1+\epsilon+\lambda)x_{n+1} = 0 \tag{11}$$

The eigenvector corresponding to the first eigenvalue (which is zero) is the constant vector, therefore for all other eigenvectors the sum of components should be zero in order to be orthogonal to the first one. Consequently $\sum_j x_j = 0$. Applying this and a minimal algebra results

$$x_n(\lambda+n) + x_{n+1}(\lambda+n) = 0 \tag{12}$$

$$x_n = -x_{n-1} \qquad \text{if } \lambda \neq -n \tag{13}$$

If $\lambda = -n$ then eqs. 7-8 reduce to $\sum_{j\leq n} x_j = 0$ and $\sum_{k\geq n+1} x_k = 0$. Now consider the eigenspace corresponding to $\lambda = -n$, and look for eigenvectors such that $x_n = x_{n+1} = c$, $\sum_{j<n} x_j = -c$, $\sum_{j>n+1} x_j = -c$. In this eigenspace the number of free parameters are $1 + 2 \cdot (n-2)$, corresponding to c and $x_1 \ldots x_{n-1}, x_{n+2} \ldots x_{2n}$ with two constraints. Altogether the dimension of the eigenspace (the multiplicity of $\lambda = -n$) is $2n-3$. Adding the $\lambda = 0$ case, we are left with at most two unknown eigenvalues.

For $\lambda \neq -n$, we look for the solutions in the form $(a, \ldots, a, b, -b, -a, \mathrm{N}\ldots, -a)^T$. Then the eigenvalue equations are

$$(n-2)a + b - (n-1+\lambda)a = 0 \tag{14}$$

$$(n-1)a - \epsilon \cdot b - (n-1+\epsilon+\lambda)b = 0 \tag{15}$$

After simplifications,

$$-(1+\lambda)a + b = 0 \tag{16}$$

$$(n-1)a - (n-1+2\epsilon+\lambda)b = 0 \tag{17}$$

Expressing λ from these equations reads

$$\lambda = \frac{b}{a} - 1 \tag{18}$$

$$\lambda = -(n-1+2\epsilon) + (n-1)\frac{a}{b} \tag{19}$$

Writing $\lambda = \lambda$ results

$$-n+1-2\epsilon+(n-1)\frac{a}{b} = \frac{b}{a} - 1 \tag{20}$$

$$(-n+1-2\epsilon)\frac{b}{a} + (n-1) = \left(\frac{b}{a}\right)^2 - \frac{b}{a} \tag{21}$$

Introducing $x = b/a$ gives

$$-x^2 + (-n+2-2\epsilon)x + (n-1) = 0 \tag{22}$$

$$x_{1,2} = \frac{n-2+2\epsilon \pm \sqrt{(n-2+2\epsilon)^2 + 4(n-1)}}{-2} \tag{23}$$

The term under the radical symbol can be approximated using the first two terms of the Taylor series $\sqrt{1-x} \approx 1 - x/2$

$$\sqrt{\cdots} = \sqrt{(n+2\epsilon)^2\left(1 - \frac{8\epsilon}{(n+2\epsilon)^2}\right)} \approx \tag{24}$$

$$\approx (n+2\epsilon)\left(1 - \frac{4\epsilon}{(n+2\epsilon)^2}\right) \tag{25}$$

which gives

$$x_1 \approx 1 - \frac{2\epsilon}{n+2\epsilon} \tag{26}$$

$$x_2 \approx 1 - (n+2\epsilon) + \frac{2\epsilon}{n+2\epsilon} \tag{27}$$

which, using eq. 18, leads to

$$\lambda_1 \approx -\frac{2\epsilon}{n+2\epsilon} \tag{28}$$

$$\lambda_2 \approx -(n+2\epsilon) + \frac{2\epsilon}{n+2\epsilon} \tag{29}$$

meaning that the last two eigenvalues of the Laplacian are found.

B Review of current methods

Here, a one-by-one review of methods follows, from the point of view of the separation & cohesion criterions.

Separation-targeted methods

Method of Lancichinetti et al. (LFK) [39] – although being a multiresolution method, it is informative to take a look at it with the resolution parameter (see eq. 1) fixed at $\alpha = 1$. Then the fitness function of a community, which is to be optimized, is simply the sum of in-degrees divided by the sum of degrees of the community members. Thus, this method is a clear implementation of the separation criterion. Consequently, it is not sensitive to the internal distribution of edges (Fig. 1a and 1b get the same fitness). The cohesion criterion is absent, so one clique on Fig. 1a has lower fitness than the union of the two cliques.
Labelpropagation [40] – the communities are defined as sets of nodes such that every node should belong to the community to which the majority of their neighbors do. Labelpropagation does not qualify the communities, just finds partitions obeying the majority rule. Consequently Fig. 1a can be judged as a proper single community, and Fig. 1b can be split by collecting each second node to the same cluster.
Infomap [18] – Infomap aims to minimize the length of the description of a random walk, using clusters. The best description length corresponds to the best trade-off between small cluster sizes (understood in in-degrees) and few links between clusters. It is straightforward to calculate that for the configuration on Fig. 1a, Infomap will properly separate the two cliques unless the number of inter-community links is larger than $6.9 \cdot 10^7$. Although this resolution limit looks practically unimportant, shows that Infomap has some conceptual problems. If 3 edges are placed instead of 1 between the 2 cliques on Fig. 1a, Infomap will merge the two cliques if the number of inter-community edges in the rest of the network is larger than 149, which is more than 5 orders of magnitude smaller than the previous threshold. Two consequences should be drawn: Infomap is quite sensitive to the number of inter-community edges, and, as a consequence, it can produce counterintuitive communities in realistic graphs.
Clique Percolation Method (CPM) [21] – communities are defined as maximal sets of adjacent k-cliques, k being a parameter. Adjacency holds if $k-1$ nodes are shared by two cliques. Although CPM enforces a very strong cohesion locally, it applies only to $\mathcal{O}(1)$-sized subgraphs of communities. Consequently, there are no cohesion requirements on the scale of the whole community. E.g., the cliques of a cluster might form a chain and the method gives no information about the shape of the cluster. Considering Fig. 1a, it is trivial to modify it such that CPM merges the two large cliques into a single cluster, e.g. using 3-cliques. Furthermore, the absence of a single percolating series of neighboring cliques means that a subgraph will not appear as a single community, regardless of its other parameters (see e.g. Fig. 1b applying 4-cliques). Finally, CPM uses the same clique size for the whole network, regardless of local variations in edge density.
Method of Radicchi et al. [20] – there are two possible criterions for communities to choose from: either all community members or only the whole community should have more links inside than outside. Proper communities are found

by iteratively bisecting the network, until no bisection can be carried out without violating the criterion used. So, the effective definition is that a community is a subgraph obeying one of the criterions mentioned above *such that no bisection of it can result proper communities*. Fig. 1b with a minor tweak would be split even using the strong definition, assigning every second node to the same community. The tweak is to place the 2 outside links on the $k_{in} = 6$ nodes.
Method of Estrada and Hatano [41] – as it relies on the eigenvalues and eigenvectors of the whole graph, it is a global method. Therefore whether a set of nodes is judged to be a cluster or not depends also on the rest of the graph. Unfortunately, the behavior of the eigenvalues and eigenvectors of the adjacency matrix of a graph are not well understood. Consequently, empirical tests were conducted. If the method is run on only the 12 nodes of Fig. 1, configuration a) is cut into the two proper sets, but configuration b) is cut into several small (overlapping) clusters, such that all triangles form one. When the 12 nodes are attached to a 100-node ring, in which first and second neighbors on both sides of a node are attached to the node (degrees are 4), then for configuration a) the two clusters expand to the first neighboring nodes in the ring, and configuration b) has the same clusters as in the fully separated case. So, if the rest of the graph is not denser than the set of nodes under investigation, it seems that internal cohesion does matter, however external separation not. If the 100-node ring is two degrees denser (first 3 neighbors are attached, degrees are 6), the 12 nodes coalesce into 1 cluster both for configurations a) and b), incorporating a few nearby nodes from the large ring (8 for a) and 6 for b)). For even denser 100-node rings, the 12 nodes become part of a large cluster containing many nodes from the large ring. So, in conclusion, the global character of the method makes it indefinite concerning its behavior to the configurations on Fig. 1a and 1b.

Stochastic blockmodels and spin-based methods

Modularity optimization [17] – for each community, modularity counts the inside links and their expected values, based on the degrees of the nodes. Due to the well-known resolution limit problem [35, 36], the optimal modularity merge the two cliques on Fig. 1a for sufficiently large graphs.
Laplacian spectral algorithm by Donetti and Muñoz [42] – although the method produces candidate partitions using the spectrum of the Laplacian matrix, the partitions are evaluated using modularity. Consequently, it is equivalent to modularity optimization using a special heuristic, implying all the drawbacks of modularity.
Link partitioning method of Evans and Lambiotte [53] – partitioning is done on the so-called line graph, which nodes correspond to the edges of the original graph, and links are drawn between edges sharing a node in the original graph. Variants of the modularity function are proposed as goal function for the partition. Different variants use different weighting schemes of the edges including the addition of self-loops. As these goal functions are still based on counting intra-community edges and subtracting some expected value, the resolution limit problem should appear for large enough graphs.
Method of Ronhovde and Nussinov (RN) [43] – it proposes a Hamiltonian $\mathcal{H}(\{\sigma\}) = -1/2 \sum_{i \neq j}(a_{ij}A_{ij} - \gamma b_{ij}(1 - A_{ij}))\delta(\sigma_i, \sigma_j)$, $\mathbf{A}$ being the adjacency matrix, a_{ij} and b_{ij} being edge weights. The configuration corresponding to the

minimal Hamiltonian is used as the solution.
The Hamiltonian optimizes simply for the edge densities inside clusters (distorted by the γ resolution parameter), which tends to be the largest for cliques. Consequently Fig. 1b worth to be split into 4 if $\gamma > 19/35$. Similarly, for $\gamma < 1/35$, the two cliques of 1a are merged. The fact that the proper value of γ may vary from cluster to cluster can render the global optimization process locally unsuccessful.
Method of Nepusz et al. [44] – the main goal of the work is to provide community detection framework using fuzzy (soft) memberships, in order to handle overlaps. The proposed realization of the framework, when restricted to conventional hard memberships (and unweighted networks), is equivalent to the previous method with $\gamma = 1$, and with a different heuristic.
Stochastic blockmodel of Hofman and Wiggins [45] – based on the assumption that the community structure can be fitted by a blockmodel in which intra- and inter-cluster nodes are connected with probabilities ϑ_c and ϑ_d respectively, [45] aims to minimize the Hamiltonian

$$H = -\sum_{i<j}(J_L A_{ij} - J_G)\delta_{\sigma_i,\sigma_j} - \sum_{\mu} n_\mu \ln \frac{n_\mu}{n} \tag{30}$$

where σ_i is the cluster of node i, n_μ is the size of cluster μ, $J_G = \ln(1-\vartheta_d)/(1-\vartheta_c)$, $J_L = \ln \vartheta_c/\vartheta_d + J_G$. The number and sizes of clusters, the cluster members, and the probabilities ϑ_c and ϑ_d are determined by minimization. In other words, a community structure should be found which maximizes the edge densities inside the communities (the $J_L A_{ij} - J_G$ term), with the restrictions that 1) each node belongs to exactly one community; 2) the expected intra- and inter-cluster edge densities are both constants. The method of Hastings [46] is a special case of this method, needing J_G and J_L as input, and discarding the last term in equation 30. The method of Ronhovde and Nussinov [43] is also a special case with the same restrictions, i.e. being equivalent to the Hastings method. One can see immediately that equation 30 defines a global method, which is realized by the global J_G and J_L coupling constants. Since the discovery of the resolution limit of modularity it is known that globality leads to counterintuitive local trade-offs. The situation is not different here, a simple calculation for Fig. 1a shows that the two cliques will be merged if $(1-\vartheta_c)/(1-\vartheta_d) > 2^{-12/35}(\vartheta_d/\vartheta_c)^{1/35}$, which can be approximated by $1-\vartheta_c > 0.79(1-\vartheta_d)$, assuming that $(\vartheta_d/\vartheta_c)^{(1/35)} \approx 1$. As ϑ_c corresponds to the intra-cluster edge probability, it is a reasonable criterion. Furthermore, it is similarly simple to show that for Fig. 1b, splitting into four is profitable if $(1-\vartheta_d)/(1-\vartheta_c) > 4^{12/35}(\vartheta_c/\vartheta_d)^{19/35}$.
Mixture model of Newman and Leicht [47] – based on some probabilistic modeling, [47] proposed the following log-likelihood to be maximized:

$$\mathcal{L} = \sum_{i,r} q_{ir} \left(\ln \pi_r + \sum_j A_{ij} \ln \Theta_{rj} \right) \tag{31}$$

where q_{ir} is the probability that node i belongs to cluster r, π_r is the fraction of nodes in cluster r, and Θ_{rj} is the probability that a randomly chosen link originating in cluster r points to node j. Equation 31 is reminiscent of the Hamiltonian of Hofman and Wiggins, although there are important differences. Nodes can have memberships in many clusters simultaneously (with the

constraint that the sum of memberships is 1 for any node). Inter-cluster edges are counted for, while missing edges are never. The coupling strength between neighboring nodes is fine-tuned for each node-cluster pair. Considering Fig. 1b and assuming hard node memberships (i.e. q_{ir} is 0 or 1 for all nodes), it is easy to show that splitting into 4 is favored over putting all nodes into one cluster.
Mixture model of Wang and Lai [48] – Wang and Lai improved the mixture model of Newman and Leicht, arriving to the log-probability

$$\mathcal{L} = \sum_{i,r} q_{i,r} \left(\ln \pi_r + \sum_j A_{ij} \ln \rho_{rj} + \sum_j (1 - A_{ij}) \ln(1 - \rho_{rj}) \right) \tag{32}$$

where ρ_{rj} is the probability that a node in cluster r has a link to node j. Now $\mathcal{L}$ counts also the missing edges. For a hard clustering ($q_{ir} = 0$ or 1) it is easy to calculate that Fig. 1b is preferred in 4 pieces over 1.
Likelihood modularity of Bickel and Chen [49] – the proposition is to maximize

$$Q_{LM} = \frac{1}{2} \sum_{c,d} n_{cd} \left(\frac{O_{cd}}{n_{cd}} \log \frac{O_{cd}}{n_{cd}} + \left(1 - \frac{O_{cd}}{n_{cd}} \right) \log \left(1 - \frac{O_{cd}}{n_{cd}} \right) \right) \tag{33}$$

where $n_{cd} = n_c n_d$ if $c \neq d$, $n_{cc} = n_c(n_c - 1)$, n_c is the size of cluster c, and $O_{cd} = \sum_{i \in c, j \in d} A_{ij}$. The expression is maximal if the clusters are cliques ($O_{cc}/n_{cc} = 1$) which are totally separated ($O_{cd}/n_{cd} = 0$). As Q_{LM} is symmetric with respect to O_{cd}/n_{cd} and $1 - O_{cd}/n_{cd}$, bipartite structures can also get high scores, but here the analysis is restricted to the cluster-based optimum. First, it should be noted that Q_{LM} penalizes clusters in which the edge density deviates significantly from its maximal value. Then, it is easy to calculate that it worth to cut Fig. 1b into 4 clusters.
Stochastic blockmodel of Karrer and Newman [50] – it is similar to the previous case. The main difference in the function to be maximized (compared to equation 33) is the absence of the second logarithmic term representing the missing links, and the application of sums of degrees instead of cluster sizes. Similarly, a simple calculation shows that Fig. 1b gets higher score when split into four.

Other single-scale methods

Infomod [51] – the aim is to compress the description of the graph, while retaining as much information as possible. The description length is given by

$$L = n \log_2 m + \frac{m(m+1)}{2} \log_2 l + \log_2 \prod_{i=1}^{m} \binom{n_i(n_i - 1)/2}{l_{ii}} \prod_{i<j} \binom{n_i n_j}{l_{ij}} \tag{34}$$

where n is the number of nodes, l is the number of edges, m is the number of clusters. As can be seen, it is global method, where trade-offs for a global improvement may spoil local structures. And indeed, a straightforward calculation shows that for all but very small graphs Fig. 1a is preferred as a single community (e.g. $l \geq 128$ and $m \geq 7$).
Method of Chauhan et al. [52] – the idea is interesting, i. e. to maximize the sum of logarithms of the largest eigenvalues of the adjacency matrices of the

individual communities. However, the behavior of the largest eigenvalue of the adjacency matrix is poorly understood. As a counterexample, given a clique of size n, its largest eigenvalue is $n-1$, while when it is cut into two, the product of the first eigenvalues of the two $n/2-1$-sized cliques is $(n/2-1)^2$, which is larger than n if $n \geq 8$ – so it worth to cut a clique into pieces. This is the consequence of using a concave function (log) in the summation, so it can be easily fixed. However, for Fig. 1b, the largest eigenvalue is 5.2, while the largest eigenvalue of a 3-clique is 2. Summing the largest eigenvalues for the two cases (instead of summing their logarithms) results in $5.2 < 8$, so it worth to split Fig. 1b into 4.
Link partitioning method of Ahn et al. [54] – although the described method applies a hierarchical clustering, using an objective function (edge density of clusters) results in a single set of communities. The objective function averages the densities of all clusters, consequently it is a global quantity; its maximum does not guarantee that each cluster is optimal, just the average – nothing prevents the over- or underpartitioning of individual clusters at the global optimum.
Community landscape method of Kovács et al. (ModuLand) [55] – see also at the hierarchical methods. The interesting idea is to give a scalar value to the edges indicating how strongly an edge belongs to communities, then identifying the local maxima & their surroundings ("hills") as the communities. The scalar value for the edges is obtained as the number of appearances of the edges in some auxiliary clusters. From each node (or edge), an auxiliary cluster is grown until its fitness value cannot be increased. Fitness is chosen as simply the average in-degree of the nodes in the growing cluster. After all auxiliary communities are determined this way, each edge is assigned a value equaling the number of times it occurred in the found communities. Edges with the locally highest score are defined as community cores. Membership values are assigned to remaining edges, based on how strongly are they related to the nearby cores. The method, actually a framework for several possible methods, depends heavily on the applied fitness function of the auxiliary clusters. Here the NodeLand auxiliary clustering will be investigated. It is quite easy to engineer graphs in the spirit of Fig. 1 which are misclustered. E.g instead of Fig. 1a, took two 7-cliques, delete 1 link from each, and connect one node with 3 nodes from the other clique, as on Fig 7a. The fitness of one almost-clique is 40/7, is just below the contribution of the node in the other clique (6/1), so the 3 links between the cliques will be included in the community. To be precise, starting a cluster from each node, one almost-clique + the connector node will appear as a cluster $7 + 1/3$ times, and the other almost-clique $6 + 2/3$ times. Fractions correspond to different possibilities when starting from the connector node. In practice, this means that with probability 2/3, all links will have uniform scalar values (perfectly flat landscape, i.e. a single hilltop), and with probability 1/3, a step-like landscape (still identified as a single cluster by the method). Symmetrization to $6 + 1/3 + 2/3$ and $6 + 2/3 + 1/3$ is straightforward, by creating a second bridge node also with 3 links. Similarly, Fig. 1b can be substituted by Fig. 7b. It consists of two 5-cliques with connections such that each node has 2 links to the other clique. ModuLand-NodeLand will tend to separate the two cliques, although their union is much more well-separated from the rest of the graph.

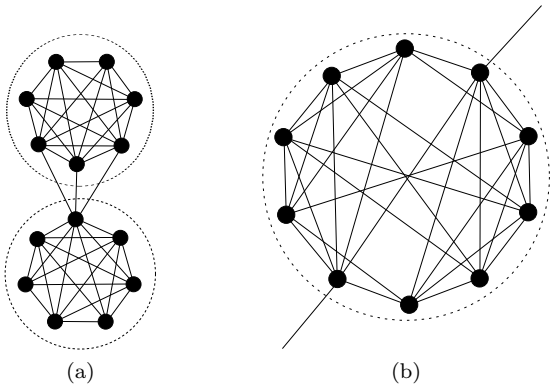


Figure 7: Subgraphs on which ModuLand-NodeLand gives counterintuitive results. Dashed lines show the desired communities.

Hierarchical methods

Here, some hierarchical methods will be investigated. The question is whether the lowest level can be reliably used as an optimal partition (or cover). As a benchmark graph, Fig. 8 will be utilized. The desired output is a single community of 12 nodes, due to their extreme separation from the rest of the graph.

Method of Ruan and Zhang [68] – the proposition is to iteratively run

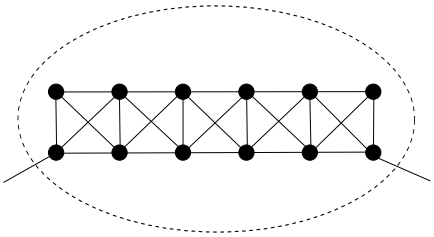


Figure 8: Test case for the lowest level of the hierarchical methods.

modularity optimization in the found clusters, until the best modularity inside a cluster is not larger significantly than those of a corresponding random graph. Numerical calculations show that at the lowest level, Fig. 8 is divided into parts[12].

Method of Sales-Pardo et al. [69] – it uses the co-occurrence of nodes in different local optima of modularity to construct a new similarity matrix, which is fitted by a block diagonal form. Communities are defined by the blocks. The method is iteratively re-applied to each community until structure deviating from a corresponding random graph is found. Again, running the method on Fig. 8 results in overpartitioning (z-score of the split Fig. 8 is 3.9, the threshold

[12] z-score is 5.3, $Q_{\max} = 0.36$. Z-score is defined as the difference of the modularity of the actual graph and the modularity of a 0-model graph, divided by the variance of the modularity of the 0-model graph, z-score $= (Q - Q_{\text{0-model}})/\sigma_{\text{0-model}}$. Criterion of [68] is z-score ≥ 2, $Q_{\max} \geq 0.3$. Modularities were optimized using the Radatools software [67].

used by [69] is 2.3).
Hierarchical Infomap [70] – this is an extension of the Infomap method [18]. It is easy to calculate that, similarly to the previous cases, splitting Fig. 8 on a lower hierarchical level improves the partition.
ModuLand [55] – ModuLand can also produce hierarchical structures, by iteratively re-running the clustering procedure on the network of clusters (links between clusters are defined by node overlaps). Accordingly, the lowest level clusters are the ones obtained by a simple ModuLand run, which is susceptible to mispartitioning, as described some paragraphs above.
OSLOM [34] – the method applies statistical significance as fitness. Although its output depends to a certain degree on the whole graph, running it on Fig. 8 (as the whole graph) results in a bisection. As the method tries to find the so-called minimal significant clusters, by trying to split already found significant subgraphs while the rest of the graph is neglected, it will divide Fig. 8 independently of the rest of the graph.

References

[1] R. Albert, A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47-97 (2002).

[2] M. E. J. Newman, *SIAM Review* **45**, 167–256 (2003).

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. Hwang, *Physics Reports* **424**, 175-308 (2006).

[4] M. Newman, A.-L. Barabási, D. Watts: The Structure and Dynamics of Networks (Princeton University Press, 2006).

[5] G. Caldarelli, A. Vespignani (eds.): Large Scale Structure and Dynamics of Complex Networks, (World Scientific, 2007).

[6] S. N. Dorogovtsev, J. F. F. Mendes: Evolution of Networks: From Biological Nets to the Internet and WWW, (New York: Oxford University Press, 2003).

[7] D. J. Watts, S. H. Strogatz, *Nature* **393**, 440-442 (1998).

[8] A.-L. Barabási, R. Albert, *Science* **286**, 509–512 (1999).

[9] R. Albert, H. Jeong, A.-L. Barabási, *Nature* **401**, 130-131 (1999).

[10] R. Cohen, S. Havlin, *Phys. Rev. Lett* **90**, 058701 (2003).

[11] R. Pastor-Satorras, A. Vespignani, *Phys. Rev. E* **63**, 066117 (2001).

[12] J. D. Noh, H. Rieger, *Phys. Rev. Lett.* **92**, 118701 (2004).

[13] M. Barahona, L.M. Pecora, *Phys. Rev. Lett.* **89**, 054101 (2002).

[14] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, *Rev. Mod. Phys.* **80**, 1275-1335 (2008).

[15] C Hauert, G Szabó, *Am. J. Phys.* **73**, 405-414 (2005).

[16] S. Fortunato, *Phys. Rep.* **486**, 75-174 (2010).

[17] M. E. J. Newman, M. Girvan, *Phys. Rev. E* **69** 026113 (2004).

[18] M. Rosvall, C. T.Bergstrom, *PNAS* **105** 1118-1123 (2008).

[19] J.-C. Delvenne, S. N. Yaliraki, M. Barahona, *PNAS* **107**, 12755-12760 (2010).

[20] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, *PNAS* **101** 2658-2663 (2004).

[21] G. Palla, I. Derényi, I. Farkas, T. Vicsek, *Nature* **435** 814-818 (2005).

[22] J. Reichardt, S. Bornholdt, *Phys. Rev. E* **74**, 016110 (2006).

[23] A. Arenas, A. Fernández, S. Gómez, *New J. Phys.* **10**, 053039 (2008).

[24] J. Kumpula, J. Saramäki, K. Kaski, J. Kertész, *Fluct. Noise Lett.* **7**, L209 (2007).

[25] A. Lancichinetti, S. Fortunato, F. Radicchi, *Phys. Rev. E* **78**, 046110 (2008).

[26] A. Lancichinetti, S. Fortunato, *Phys. Rev. E* **80**, 016118 (2009).

[27] R. Guimerà, M. Sales-Pardo, L. A. N. Amaral, *Phys. Rev. E* **70**, 025101 (2004).

[28] C. P. Massen, J. P. K. Doye, arXiv:cond-mat/0610077v1 (2006).

[29] A. Lancichinetti, F. Radicchi, J.J. Ramasco, *Phys. Rev. E* **81**, 046110 (2010).

[30] B. Karrer, E. Levina, M. E. J. Newman, *Phys. Rev. E* **77**, 046119 (2008).

[31] Y. Hu, Y. Ding, Y. Fan, Z. Di, arXiv:1002:2007v1 (2010).

[32] M. Rosvall, C. T. Bergstrom, *PLoS ONE* **5**, e8694 (2010).

[33] D. Gfeller, J.-C. Chappelier, P. De Los Rios, *Phys. Rev. E* **72**, 056135 (2005).

[34] A. Lancichinetti, F. Radicchi, J. J. Ramasco, S. Fortunato, *PLoS ONE* **6**, e18961 (2011).

[35] S. Fortunato, M. Barthélemy, *PNAS* **104** 36-41 (2007).

[36] J.M. Kumpula, J. Saramäki, K. Kaski, J. Kertész, *Eur. Phys. J. B* **56**, 41-45 (2007).

[37] A. Lancichinetti, M. Kivela, J. Saramäki, S. Fortunato, *PLoS ONE* **5**, e11976 (2010).

[38] G. Tibély, M. Karsai, L. Kovanen, K. Kaski, J. Kertész, J. Saramäki, *Phys. Rev. E* **83**, 056125 (2011).

[39] A. Lancichinetti, S. Fortunato, J. Kertész, *New J. Phys.* **11** 033015 (2009).

[40] U. N. Raghavan, R. Albert, S. Kumara, *Phys. Rev. E* **76** 036106 (2007).

[41] E. Estrada, N. Hatano, *Phys. Rev. E* **77**, 036111 (2008).

[42] L. Donetti, M. A. Muñoz, *J. Stat. Mech.* P10012 (2004).

[43] P. Ronhovde and Z. Nussinov, *Phys. Rev. E* **80**, 016109 (2009).

[44] T. Nepusz, A. Petróczi, L. Négyessy, F. Bazsó, *Phys. Rev. E* **77** 016107 (2008).

[45] J. M Hofman, C. H. Wiggins, *Phys. Rev. Lett.* **100**, 258701 (2008).

[46] M. B. Hastings, *Phys. Rev. E* **74** 035102 (2006).

[47] M. E. J. Newman, E. A. Leicht, *PNAS* **104** 9564-9569 (2007).

[48] J. Wang, C-H Lai, *New J. Phys.* **10**, 123023 (2008).

[49] P. J. Bickel, A. Chen, *PNAS* **106**, 21068-21073 (2009).

[50] B. Karrer, M. E. J. Newman, *Phys. Rev. E* **83**, 016107 (2011).

[51] M. Rosvall, C. T.Bergstrom, *PNAS* **104** 7327-7331 (2007).

[52] S. Chauhan, M. Girvan, E. Ott, arXiv:0911.2735v1 (2009).

[53] T. Evans, R. Lambiotte, *Phys. Rev. E* **80**, 016105 (2009).

[54] Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, *Nature* **466**, 761-764 (2010).

[55] I. A. Kovács, R. Palotai, M. S. Szalay, P. Csermely, *PLoS ONE* **5**, 12528 (2010).

[56] A. Clauset, C. Moore, M. E. J. Newman, *Nature* **453**, 98-101 (2008).

[57] S. van Dongen, Ph.D. thesis, Dutch National Research Institute for Mathematics and Computer Science, University of Utrecht, Netherlands (2000).

[58] B. Mohar, S. Poljak, *Combinatorial and Graph-Theoretical Problems in Linear Algebra* **50**, 107-151 (1993).

[59] M. Fiedler, *Czechoslovak Math. J.* **23**, 298-305 (1973).

[60] A. Clauset, *Phys. Rev. E* **72**, 026132 (2005).

[61] W. W. Zachary, *J. Anthropol. Res.* **33**, 452-473 (1977).

[62] A. Lancichinetti, S. Fortunato, *Phys. Rev. E* **80**, 056117 (2009).

[63] C. Lee, F. Reid, A. McDaid, N. Hurley, conference paper, Workshop - ACM KDD-SNA (preprint: arXiv:1002.1827v1) (2010).

[64] D. L. Nelson, C. L. McEvoy, T. A. Schreiber, "The university of south florida word association, rhyme, and word fragment norms" (1998) `http://www.usf.edu/FreeAssociation/`.

[65] `http://www.ime.unicamp.br/~chico/arpack++/`

[66] V. Hernandez, J. E. Roman, V. Vidal, *ACM Transactions on Mathematical Software* **31**, 351-362 (2005).

[67] `http://deim.urv.cat/~sgomez/radatools.php`

[68] J. Ruan, W. Zhang, *Phys. Rev. E* **77**, 016104 (2008).

[69] M. Sales-Pardo, R. Guimerà, A. A. Moreira, L. A. N. Amaral, *PNAS* **104**, 15224-15229 (2007).

[70] M. Rosvall, C. T. Bergstrom, *PLoS ONE* **6**, e18209 (2011).

Communities and beyond: Mesoscopic analysis of a large social network with complementary methods

Gergely Tibély,[1] Lauri Kovanen,[2] Márton Karsai,[2] Kimmo Kaski,[2] János Kertész,[1,2] and Jari Saramäki[2,*]
[1]*Institute of Physics and HAS-BME Condensed Matter Group, BME, Budapest, Budafoki út 8., H-1111, Hungary*
[2]*BECS, Aalto University, P.O. Box 12200, FI-00076 Aalto, Finland*
(Received 4 June 2010; revised manuscript received 15 March 2011; published 31 May 2011)

Community detection methods have so far been tested mostly on small empirical networks and on synthetic benchmarks. Much less is known about their performance on large real-world networks, which nonetheless are a significant target for application. We analyze the performance of three state-of-the-art community detection methods by using them to identify communities in a large social network constructed from mobile phone call records. We find that all methods detect communities that are meaningful in some respects but fall short in others, and that there often is a hierarchical relationship between communities detected by different methods. Our results suggest that community detection methods could be useful in studying the general mesoscale structure of networks, as opposed to only trying to identify dense structures.

DOI: 10.1103/PhysRevE.83.056125 PACS number(s): 89.75.Fb, 89.75.Hc, 89.65.−s

I. INTRODUCTION

Large complex networks have different levels of organization. On the microscopic level networks are composed of pairwise interactions, but it is the macroscopic level that has received most attention in recent years. We now know that diverse networks exhibit similarities, for example, in degree distribution, average path length, and clustering coefficient. While the structure is interesting in its own, it also has a significant influence on the dynamic processes taking place on the network, such as spreading, diffusion, and synchronization [1–3].

The intermediate mesoscopic scale has turned out be more elusive to describe. It is this scale where we can identify, for example, motifs [4,5] and dense clusters of nodes commonly known as *communities*. Although communities are relevant for understanding the structure of and the dynamics on networks, even their exact definition is still a controversial issue. Thus it comes as no surprise that the art of *community detection* has grown into a swarming field of diverse methods [6]. Many features of real-world networks add to the complexity of the task. Real networks are often hierarchical, and hence small communities may reside inside larger ones, communities may overlap if nodes participate in several communities, and even more complications arise if we take into account link weights that represent interaction intensity.

Until recently the performance of community detection methods has mainly been tested on small empirical networks with typically no more than 100 nodes, which allows the evaluation of quality by visual inspection. However, several networks of considerable interest are much larger, often with 10^6 nodes or more: data on the World Wide Web, mobile phone call records, electronic footprints of instant messaging users, and networks of social webs such as Facebook, etc. Only a few methods are efficient enough to handle such networks [7–10]—to be successful, a community detection method must be computationally efficient in addition to being accurate.

More systematic comparisons have been recently carried out using synthetic benchmark networks with built-in community structure [11,12]. Although benchmarks are useful in evaluating performance, even their authors acknowledge that they represent only the first step. No benchmark fully incorporates the spectrum of properties commonly observed in real-world networks. Some recent benchmarks do allow heterogeneous distributions for degrees and community sizes, but many other properties are still missing, such as high clustering, existence of cliques [13], overlapping communities [14], assortativity [15], and the prevalence of motifs [16]. This distorts the evaluation of algorithms that depend on (or benefit from) the existence of these features. For example, clique percolation has been successfully used on real-world networks [13,17–19] but does not perform well on synthetic benchmarks—mainly due to its strict requirement for communities to consist of adjacent cliques [12].

In this paper we take three widely applied methods, each based on a different underlying philosophy, and compare their performance on a large real-world social network constructed from mobile phone call records. Unlike with benchmark networks, we do not know the "correct" community structure of the network. Therefore, we introduce new measures that allow us to investigate the differences and similarities of the detected community structures.

The paper is organized as follows. Section II describes the choice of community detection methods, and Sec. III introduces the data set. Section IV presents the results of our analysis where we first analyze the properties and statistics of individual community structures and then turn to a pairwise comparison to quantify the differences between communities. Finally in Sec. V we present conclusions.

II. CHOICE OF METHODS

Because we intend to study a large network, the first selection criterion is only practical: Methods with running time $O(N^2)$ or slower cannot be included. We use three methods that not only fill this requirement but in addition have performed

*jari.saramaki@tkk.fi

well in previous comparisons or in practice: the *Louvain method* [9], the *Infomap* [20], and the *clique percolation* [13].

We consider an undirected network $G = (V, E)$, where V is the set of N nodes and E the set of L edges. The *degree* k_i is the number of neighbors node i has, $k_i = |\{j|(i,j) \in E\}|$. For mathematical purposes a *community* c is simply a set of nodes, $c \subseteq V$, and we denote community size by $S = |c|$. The communities detected by one method constitute a *community structure* $C = \{c_1, \ldots, c_{n_c}\}$. A *partition* P is a special community structure where each node belongs to exactly one community, i.e., $c_i \cap c_j = \emptyset$ if $i \neq j$ and $\bigcup_{i=1}^{n_c} c_i = V$.

All three methods can be extended to handle weighted networks where each edge has a numerical weight w_{ij}. In this paper we consider only positive weights; $w_{ij} = 0$ is equivalent to $(i,j) \notin E$. The weighted counterpart of degree is *node strength*: $s_i = \sum_{(i,j)\in E} w_{ij}$.

The *Louvain* method (LV) [9] was the best of the modularity optimization methods tested in Ref. [8]. Modularity is the expected value of the difference of the number of edges inside communities in the actual network and in a random network with the same degree sequence [21]:

$$Q = \sum_{c \in P} \left[\frac{L_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right], \tag{1}$$

where L_c is the number of edges inside community c and $d_c = \sum_{i \in c} k_i$ its total degree. In the weighted version all quantities are replaced by their weighted counterparts: L_c by the total sum of weights inside a community, d_c by the sum of node strengths, and L by the sum of edge weights in the whole network.

Because modularity optimization is an NP-complete problem [22], LV uses a greedy heuristic to find a local optimum. Each node is initially a separate community, i.e., $c_i = \{i\}$. Neighboring communities[1] are merged in random order so that modularity increases maximally at each step until a local maximum is reached. Resulting communities are then shrunk into "supernodes," and the optimization is repeated on the new "renormalized" network. The two steps—optimization and renormalization—are repeated recursively until no further improvement of modularity is possible.

The local heuristic of LV seems to avoid some of the resolution issues of modularity. In addition, the renormalized networks can be understood as different levels of a hierarchical community structure.

The *Infomap method* (IM) [20] came out on top in a recent state-of-the-art benchmark comparison [8]. The idea is to describe a random walker with a two-level coding scheme: The higher level has a single codebook for communities, and on the lower level each community has its own codebook with a special exit code for moving out of the current community. The optimal partition corresponds to the codebook with the minimum description length: Too small communities increase the description length due to higher frequency of community crossings, while communities containing too many nodes require longer description. In weighted networks the random walks are biased toward edges with higher weight. Since an exhaustive search for the optimal partition is not feasible, IM employs a heuristic similar to the one used in LV.

Clique percolation (CP) [13] has been successfully applied to large empirical graphs, e.g., to study the dynamics of social groups [17]. A k-clique is a fully connected subgraph of k nodes, and two k-cliques are considered adjacent if they share $k-1$ nodes. As the name suggests, clique percolation defines communities as connected k-clique components: A CP community is a maximal set of k-cliques such that there is a path of adjacent k-cliques between them. Different values of k yield different community structures, and communities obtained with a larger value of k reside inside those obtained with a smaller value. To select the best value of k we follow Ref. [13] and use the smallest value for which there is no giant percolating community.

There are significant differences between CP and the other two methods. Both LV and IM use a stochastic optimization scheme while CP is entirely deterministic. In addition, LV and IM yield a partition, but CP does not. With CP the nodes that do not belong to any k-clique are left outside communities, and if a node belongs to several k-cliques it may belong to more than one *overlapping* community. The fact that CP does not provide a partition is not necessarily a bad thing: Sparse regions of the network do not appear as communities, and, e.g., in social networks individuals often do belong to multiple groups, such as family, friends, and colleagues.

To define the weighted clique percolation (wCP) [23] we need the concept of clique *intensity*, defined as the geometric mean of edge weights. In wCP we use a value of k that would give a giant community in the unweighted case but include only those k-cliques that have intensity larger than some predefined threshold $I_>$. Analogously to the unweighted case, $I_>$ is set to the largest value for which there is no giant community.

Notes on applying the methods are given in Appendix A.

III. THE DATA

Our empirical test network is a mobile phone call network constructed from billing records of seven million customers of a single mobile phone operator whose customer base covers about 20% of the population in its country. The records cover a period of 126 days. To ensure anonymity of customers, phone numbers have been replaced by surrogate keys. Data from the same operator have been previously studied in Refs. [17,24,25].

For this study we use only voice calls, and only those that take place between customers of the operator in question. In addition we exclude edges where only one person has made calls to the other during the whole period. We study only the largest connected component, which has $N = 4.9 \times 10^6$ nodes and $L = 10.9 \times 10^6$ edges (mean degree $\langle k \rangle \approx 4.44$).[2]

[1] Two communities are neighbors if there is at least one link between them.

[2] The largest connected component contains 92% of nodes and 98% of edges; the second-largest component has only 47 nodes.

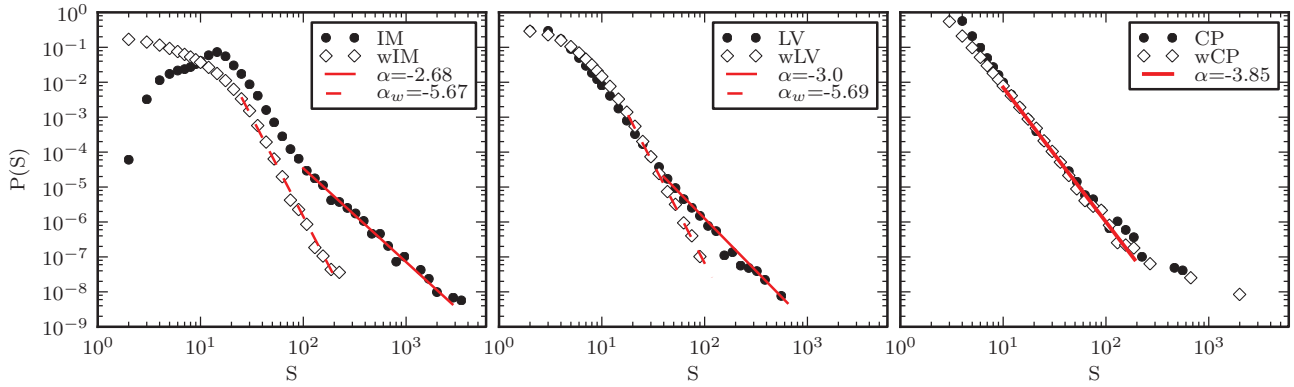


FIG. 1. (Color online) Community size distributions for IM, LV, and CP and their weighted versions. The parameter α denotes the exponent when the tails are fitted a power-law distribution $P(S) \propto S^{\alpha}$; solid lines correspond to the unweighted α and dashed lines to the weighted α_w.

The edge weights in the weighted network are defined as sums of call durations (in seconds) between the two customers. The average weight is $\langle w \rangle \approx 4634$ s.

Using a large social network enables us to relate the findings to known characteristics of such networks [24]. It is known that the overlap of local neighborhoods of adjacent nodes increases with edge weight[3] [26], as conjectured in the "weak ties" hypothesis of Granovetter [27]. This feature should be reflected in correlations between edge weights and communities. We can also study structural features of communities and evaluate whether they represent meaningful social communities.

IV. RESULTS

We analyze *single* community structures detected by each method. Both LV and IM are stochastic methods and therefore give a slightly different partition on every run; however, as shown in Appendix B, the qualitative properties of the communities are stable enough to justify the comparison.

Appendix A contains detailed notes about the application of the three methods. In brief, we use parameters $k = 3$ for CP and $k = 4$ with $I_{>} = 3093$ for wCP—these are the only two methods with explicit parameters—and with LV we study only the first level of the hierarchical community structure since other levels yield communities that are implausibly large in the social context.

A. Community size distributions

Figure 1 shows the community size distributions for all methods. All distributions are broad, as suggested by previous results [10,13,28].

For IM, the tail of the size distribution appears power-law-like. Very small communities are rare. The community structure of wIM is notably different. The weighted communities are smaller, and the distribution is now monotonously decreasing.

Even though the largest LV communities are an order of magnitude smaller than in IM, LV still produces larger communities than its weighted variant wLV. Both LV and wLV have monotonous community size distributions, and small communities are more prevalent than in IM. The power-law exponents for the tails are similar when comparing LV to IM and wLV to wIM.

For CP and wCP the size distributions are well approximated by a power law. This is expected, because the communities are detected close to the critical point where a giant community would emerge. The largest deviation from power-law behavior is in the tail. The largest wCP communities are larger than those in CP because three-cliques are used for wCP and four-cliques for CP. Although these communities partially overlap (see Sec. IV E), the three-clique communities extend far beyond the four-clique communities.

B. Visual observation of small communities

The qualitative properties of small communities can be estimated visually, similarly to evaluating performance on small empirical networks. Figure 2 shows archetypal communities with $S = 5$, 10, 20, and 30, and their immediate network surroundings. Communities larger than this tend to be too complex to visualize in two dimensions.

Of all unweighted methods the CP communities are the least surprising: Larger communities naturally appear only in dense parts of the network. Small LV communities consist of interconnected cliques, which coincides well with the general idea of social groups. The smallest IM communities with $S \leqslant 10$, however, are typically treelike and located at the "edge" of the network—these communities are attached to the rest of the network by only few links. LV covers these sparse parts of the network with much smaller communities (see Fig. 9).

When the weights are taken into account, the partition-based methods wIM and wLV tend to produce even more treelike communities that have the appearance of local "backbones" of the network. This is a natural consequence of the way wIM and wLV use edge weights; however, communities like these do not coincide well with the idea of dense social groups.

[3]Except for the very largest edge weights, where the relation is reversed.

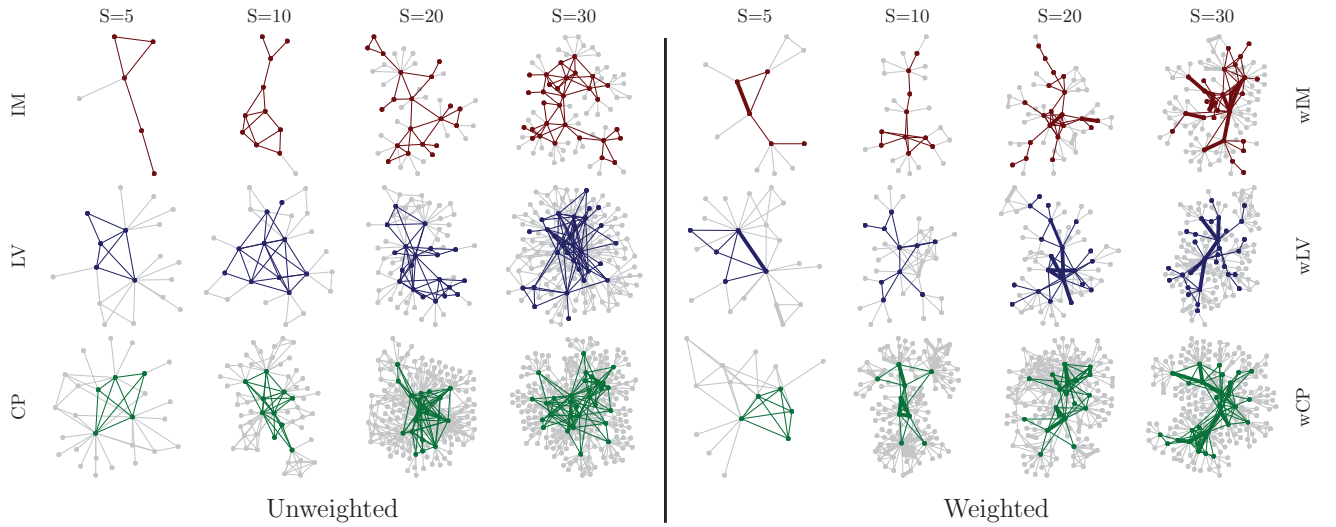


FIG. 2. (Color online) Typical (left) unweighted and (right) weighted communities of different size. These communities have been manually selected from a large random sample of communities with the intention of portraying archetypal examples. Colored (dark gray) nodes and edges denote nodes inside a single community, and the light gray nodes are the first neighbors of the nodes in the community. In weighted communities the edge width is proportional to the logarithm of edge weight, with the restriction that edges with $w_{ij} \leqslant 300$ (5 min) have the minimum width and those with $w_{ij} \geqslant 14\,400$ (4 h) the maximum width.

C. Community density distribution

Since some small communities were already observed to be treelike, we turn to more quantitative characterization of community density. Graph density is normally defined as the proportion of edges out of all possible edges, $L_c/\left[\frac{1}{2}S(S-1)\right]$. However, since communities are necessarily connected, it is more illustrative to study density relative to the sparsest possible community, a tree with $S-1$ edges, as was also done in Ref. [10]: We define density as $D_c = L_c/(S-1)$. In general $1 \leqslant D_c \leqslant S/2$ where the lower bound corresponds to trees and the upper bound to cliques. CP, however, doesn't allow trees; instead, the smallest possible density is reached when each new node adds only $k-1$ edges. In this case $L_c = \binom{k}{2} + (k-1)(S-k)$, which gives $D_c \geqslant (k-1)(S-\frac{k}{2})/(S-1)$. For $S \gg k$ this is approximately $k-1$.

Figure 3 shows the distributions and average values of D_c as function of community size. As expected, CP yields the densest communities. For IM the value of D_c stays close to 1 until ≈ 20, which confirms the observation on the prevalence of small treelike communities. For LV the distribution has a curious bimodal shape in the range $20 < S < 50$: Typical LV communities of this size have D_c from 2 to 4, but there is a small number of LV communities that are trees ($D_c = 1$) but none that are almost trees. A closer inspection (not shown) of these trees reveals that they are stars.

The plots for weighted communities in Fig. 3 suggest that weights make the communities more similar across methods. Both wIM and wLV communities are more treelike, as seen in Sec. IV B.

Treelike communities do not fit well either with the idea of social groups, or that of communities in general being dense groups of nodes. However, if a network contains treelike regions, partition-based methods will correspondingly yield

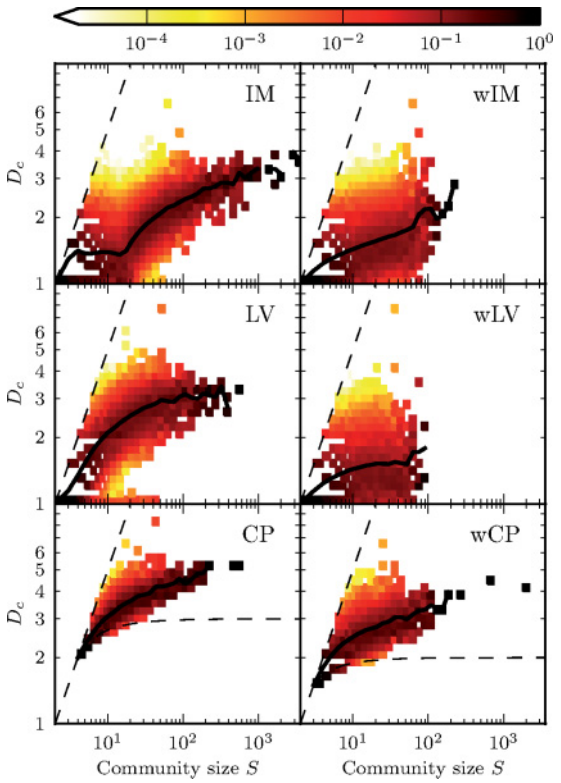


FIG. 3. (Color online) The distribution of relative density $D_c = L_c/(S-1)$ for communities from each method. In all plots each column represents a distribution and is normalized to one, the colors indicating probability density so that the darker the color, the higher the density (see color bar). The thick solid line denotes the average value. The dashed straight line corresponds to cliques, for which $D_c = S/2$. For IM and LV the smallest density is 1, which corresponds to trees. For CP, the smallest possible density is indicated by the curved dashed line (see text).

treelike communities,[4] as also seen in Ref. [10]. The abundance of treelike parts may just be a sampling artifact, as our network does not cover the whole population. Nevertheless, empirical data are rarely perfect, and a good community detection method should deal with this in a sensible way. One could argue that in treelike regions the network is so sparse that there isn't enough information about community structure. This makes CP's requirement—that nodes must participate in at least one clique to be assigned a community—appear meaningful. On the other hand, CP may yield communities where *cliques* are arranged as chains or starlike patterns, which again does not coincide well with the idea of social groups. Figure 3 indicates that in CP and wCP there are indeed some communities with densities close to the lower bound.

Whatever the interpretation, the detected treelike structures do provide information about the mesoscopic structure of the network. In other networks starlike structures can represent meaningful communities: For example, in air transport networks the peripheral airports are connected to local hubs [30].

D. Intra- and intercommunity edges

If the detected partitions are any good, nodes should have more edges to other nodes in the same community than to those in other communities. To measure this we define $\rho(c)$ as the ratio of total out- and in-degree of a community:

$$\rho(c) = \frac{\sum_{i\in c} k_i^{\text{out}}}{\sum_{i\in c} k_i^{\text{in}}} = \frac{1}{2L_c}\sum_{i\in c} k_i^{\text{out}}. \qquad (2)$$

Figure 4 shows the distribution of $\rho(c)$ as function of community size. With respect to this measure IM produces the most clear-cut communities: The majority of IM communities have ρ below one. The values for small communities are especially low, confirming the earlier observation that small IM communities are on the "edges" of the network. LV communities also have $\rho < 1$ on average, except for the smallest communities, but the values are not as low as with IM. Including weights increases the average value of ρ. wLV communities, in fact, have on average more links going outside the community than inside.

Because CP allows nodes to belong to multiple communities, a good community need not have a low value of $\rho(c)$. Also note that with CP a large fraction of edges are attached to noncommunity nodes. For CP (wCP) only 21.4% (18.6%) of edges and 21.8% (25.4%) of nodes are inside communities; 47.6% (43.0%) of edges are between noncommunity nodes.

From earlier studies of mobile phone call networks [24,26] we know that there is a correlation between edge weight and neighborhood overlap, in agreement with the Granovetter hypothesis [27]. Because nodes inside communities have overlapping neigborhoods, we expect the links between communities to be on average weaker than those within communities. Table I shows that with all methods this is indeed the case. With weighted methods this result is, of course, not as surprising since weights were used in identifying the communities.

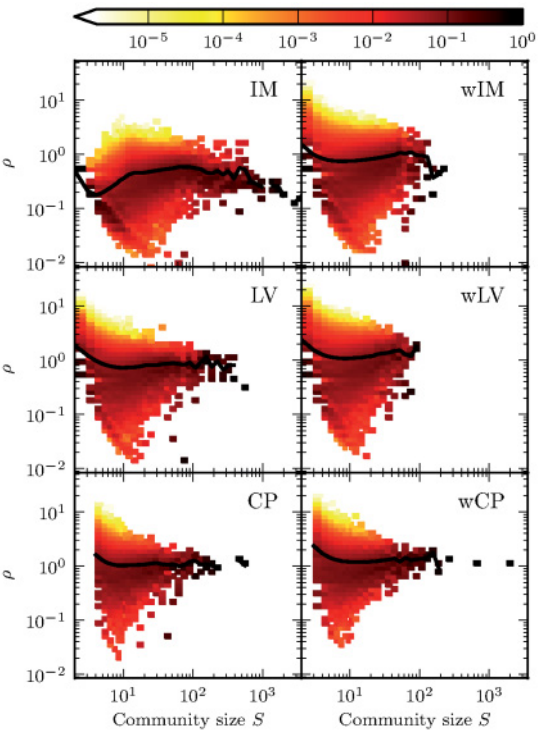


FIG. 4. (Color online) The distribution of $\rho(c)$ [Eq. (2)] as function of community size for each method. The distributions are presented as in Fig. 3, with a similar shading scheme. The black line denotes average value.

To see beyond averages, Fig. 5 displays the normalized average edge weight inside communities as function of community size. Most notably the edge weights in the largest communities are below the network average—even for wIM and wLV.

E. Neighborhood overlap

Neighborhood overlap quantifies the similarity of a node's neighborhood in two community structures. If $\mathcal{N}_i(C_j)$ is the set of those neighbors of node i that belong to its community

TABLE I. Edge weights inside and between communities. $\langle w\rangle$ denotes the average edge weight in the whole network, $\langle w_c\rangle$ the average weight for edges inside communities, and $\langle w_{c-c}\rangle$ between communities. CP also has noncommunity nodes; $\langle w_{c-n}\rangle$ denotes the average weight between community and noncommunity nodes and $\langle w_{n-n}\rangle$ between two noncommunity nodes.

	$\langle w_c\rangle/\langle w\rangle$	$\langle w_{c-c}\rangle/\langle w\rangle$	$\langle w_{c-n}\rangle/\langle w\rangle$	$\langle w_{n-n}\rangle/\langle w\rangle$
IM	1.14	0.69		
LV	1.20	0.78		
CP	1.20	0.57	0.80	1.06
wIM	1.65	0.18		
wLV	1.92	0.25		
wCP	2.57	0.43	0.57	0.73

[4]It has been shown that if there are nodes with a single link, for modularity optimization they should always belong to the community of the node to which they are connected [29]. By construction, this holds for IM as well.

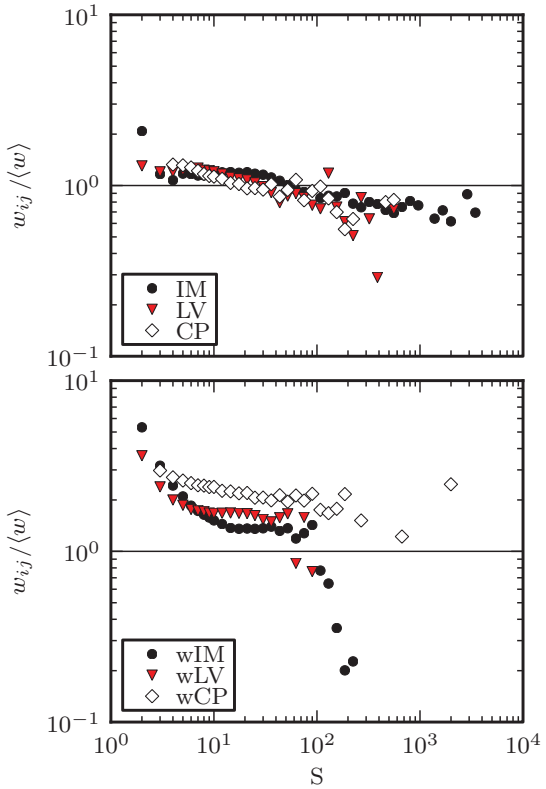


FIG. 5. (Color online) Average edge weights $w_{ij}/\langle w\rangle$ inside communities as a function of community size S, normalized by the network average.

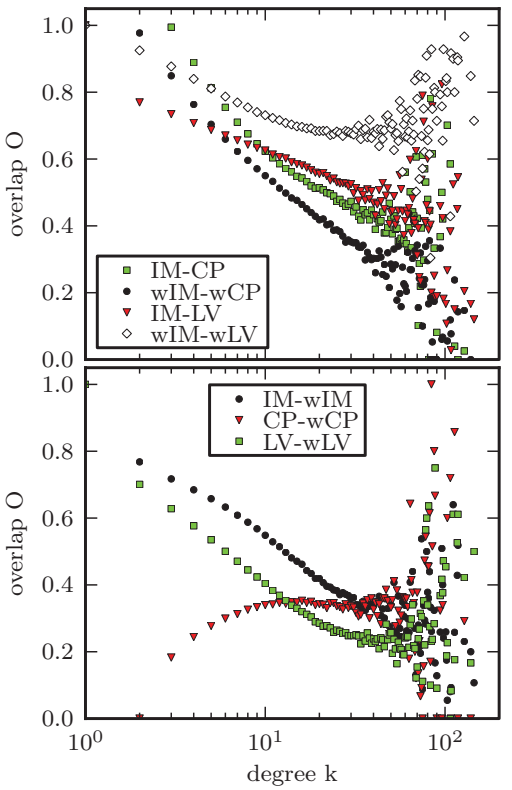


FIG. 6. (Color online) Average neighborhood community overlap O as a function of node degree k, between different methods (top) and unweighted and weighted versions of the same method (bottom).

in C_j, the neighborhood overlap is defined as the Jaccard index of $\mathcal{N}_i(C_1)$ and $\mathcal{N}_i(C_2)$:

$$O_i(C_1,C_2) = \frac{|\mathcal{N}_i(C_1) \cap \mathcal{N}_i(C_2)|}{|\mathcal{N}_i(C_1) \cup \mathcal{N}_i(C_2)|}. \quad (3)$$

Thus $O_i = 1$ if the same neighbors of i belong to its own community in both methods and $O_i = 0$ if the sets do not overlap. In the case of CP we consider only nodes that participate in at least one community; for nodes that participate in several, we assign the node to the community where most of its neighbors reside.

Figure 6 displays the average neighborhood overlap as function of degree for selected method pairs.[5] Nearly all pairs show a decreasing trend and thus in general community neighborhoods of low-degree nodes are more similar. The IM-CP and wIM-wCP overlaps decrease the fastest, because the underlying philosophies are different and the large number of nodes not appearing in any CP community reduces the overlap. wIM and wLV show a better match than their unweighted counterparts, suggesting a similar and fairly strong response to edge weights. On the other hand, overlaps for IM-wIM and LV-wLV become small for large k, which suggests that taking weights into account considerably changes the partitions for these methods. With CP-wCP the opposite behavior occurs because wCP is based on three-cliques and many nodes that are included in a three-clique are not included in any four-clique.

[5]Instead of showing the results for all 15 method pairs we present only the most interesting cases.

F. Nested communities

The above analysis shows that the three methods do not detect the same communities. It is possible, however, that they detect only different levels of a hierarchical community structure. If this is true, then the communities from one method should be the subset of another.

To address this question quantitatively we calculate how accurately a single community $c' \in P_i$ can be *tiled* by the communities of another partition P_j. The best tiling is reached with set $T \subseteq P_j$ that minimizes the sum of *external faults*

$$F_{\text{ext}}(c',T) = \sum_{c_j \in T} |c_j| - |c' \cap c_j|, \quad (4)$$

which equals the number of nodes in T but outside c', and *internal faults*

$$F_{\text{int}}(c',T) = |c'| - \sum_{c_j \in T} |c' \cap c_j|, \quad (5)$$

which equals the number of nodes in c' but outside T. As illustrated in Fig. 7, the minimum of $F_{\text{ext}} + F_{\text{int}}$ is reached when T contains only those communities for which $|c' \cap c_j| > \frac{1}{2}|c_j|$, i.e., those $c_j \in P_j$ that share at least half of their nodes with c'. To allow comparing communities of different size we

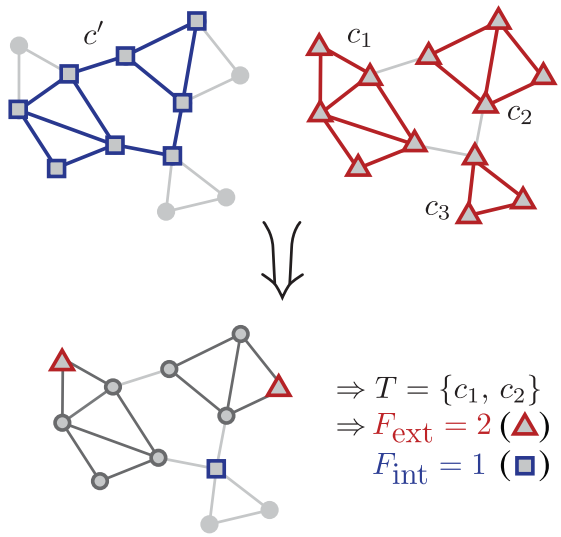


FIG. 7. (Color online) Illustration of tiling imperfection. The eight nodes in c' are spread over three different communities in another partition. Using $T = \{c_1, c_2\}$ gives the best tiling; including c_3 would reduce F_{int} to 0 but increase F_{ext} by 2. The value of tiling imperfection is $\mathcal{I} = 3/8$.

define *tiling imperfection* $\mathcal{I}(c', P_j)$ as the ratio of this minimum total fault and community size:

$$\mathcal{I}(c', P_j) = \frac{\min(F_{\text{ext}} + F_{\text{int}})}{|c'|}. \tag{6}$$

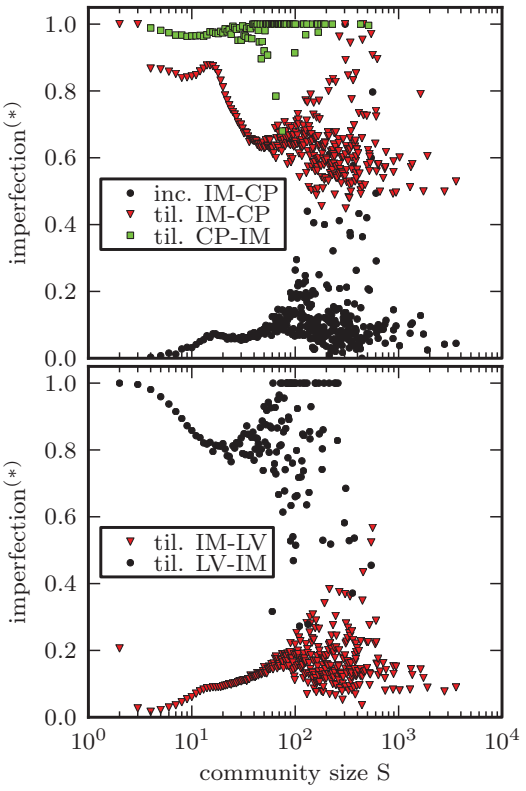


FIG. 8. (Color online) (Top) Tiling imperfection $\mathcal{I}$ and inclusion imperfection $\mathcal{I}^*$ between IM and CP. (Bottom) Tiling imperfection $\mathcal{I}$ between IM and LV.

Note that the aim of this measure is to quantify the subset-superset relationships of communities, which cannot be done with symmetric measures such as mutual information.

It is possible to generalize this measure also for general community structures,[6] such as the one produced by CP, but this is not advisable: If c' would have nodes that are not included in any community of C_j, these nodes would automatically be internal faults, and the tiling imperfection would be misleadingly high. To correct for this we define *inclusion imperfection* $\mathcal{I}^*(c', C_j)$ similar to tiling imperfection, but nodes may be counted as internal faults only if they are covered by both community structures.

Results for tiling measures are shown in Fig. 8. Comparing the tiling and inclusion imperfections for IM-CP, especially for small communities, illustrates the difference of these two measures: *Tiling* imperfection is high since small IM communities are treelike and therefore not included in any CP community; low values of *inclusion* imperfection, however, show that CP communities tend to be subsets of IM communities. High values of CP-IM tiling imperfection shows that the reverse is not true.

The low tiling imperfection for IM-LV and high for LV-IM shows that IM communities tend to be supersets of LV communities. The extreme values for small communities indicate that nearly all small IM communities can be perfectly tiled with LV communities, while small LV communities can

[6] If $T^* = \cup_{j \in T} c_j$, the generalized tiling is defined by $F_{\text{ext}}(c', T) = |T^*| - |T^* \cap c'|$ and $F_{\text{int}}(c', T) = |c'| - |T^* \cap c'|$. The optimal T can now be constructed by first including (as before) the communities that share at least half of their nodes with c', but then adding also those communities that contain more uncovered nodes of c' (i.e., those in $c' \backslash T^*$) than new nodes outside c'. Here, however, we use the same definition of T as for partitions to make the values more comparable.

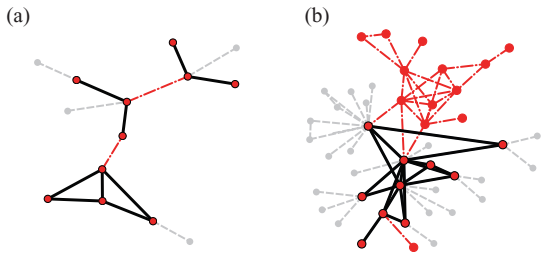


FIG. 9. (Color online) Typical cases of tiling with IM (red or dark gray) and LV (black) communities of size $S = 10$. Light gray nodes are the first neighbors of the community to be tiled. (a) Example of perfect tiling $\mathcal{I} = 0$ when IM community (red nodes) is tiled with LV communities (black edges). A typical IM community with $S = 10$ is located in a treelike region of the network, and LV covers such regions with very small communities. (b) Example of tiling imperfection $\mathcal{I} = 1$ when LV community (black edges) is tiled with IM communities (in red). A typical LV community with $S = 10$ is in a somewhat denser part of the network, where the IM communities are much larger.

almost never be tiled with IM communities.[7] A typical tiling of small IM and LV communities is shown in Fig. 9.

V. CONCLUSIONS AND DISCUSSION

Benchmarks are helpful if the methods are to be tested for sensitivity to particular properties, such as hierarchical structure or broad distribution of community sizes. Real-world networks, however, are incomparably more complicated, often inhomogeneous in many respects and usually contain many different kinds of mesoscopic structures. Good performance on benchmark graphs does not ensure that communities identified in real data are meaningful. Our analysis of the Infomap, Louvain, and clique percolation methods applied to a large social network reveals that although all the three methods do detect reasonable communities in some respects, they still come up short in others.

With all these methods the edge weights were higher inside communities than between them, in accordance with the Granovetter hypothesis [27]; distributions of community sizes were broad, as expected; and tiling imperfection revealed that although IM and LV produce different partitions, they have a hierarchical relation where LV communities tend to be inside IM communities. On the other hand, both IM and LV yield treelike communities that do not coincide well with the notion of a social community, and using edge weights makes the communities even sparser. In contrast, CP clusters are always found in dense regions of the graph and are therefore often meaningful; as a downside CP may end up discarding some important parts of communities.

A natural question is how well our findings can be generalized to other types of networks. Analysis of multiple datasets is beyond the scope of this work, but some speculation can be done. Broad community size distributions have already been observed in a number of studies [7,10,13]. Considering the numerous treelike communities, similar sparse regions occur in other networks as well. For example, the authors of Ref. [10] found that the density of communities can vary widely across different network types; e.g., the Internet has very sparse communities, whereas information networks (such as arXiv citations) have dense ones. The similarity of IM and LV may hold too because both partition the network and their heuristics are similar. The authors of Ref. [10] observed that two very different partitioning methods resulted in similar communities in terms of statistical properties. On the other hand, the difference between CP and the partition-based methods is likely to manifest itself for various networks.

In large sparse networks partitioning methods inevitably identify questionable regions as communities. The trees, starlike formations, and stars detected by IM and LV do, however, bear mesoscopic structural meaning: They too are building blocks of the network. The same topological structure may be considered a community for one purpose but not for some other—a star is hardly a social community but may reasonably be considered as one in, for example, biochemical networks [10].

It would seem that the analysis of large empirical networks would benefit from the use of complementary community detection methods and a comparison of the identified structural features. Instead of just devising ever more efficient community detection methods it might be more beneficial to take into consideration the existence of different types of mesoscopic structures, as opposed to fixating on a predefined idea of dense communities.

ACKNOWLEDGMENTS

The project ICTeCollective acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number 238597. We acknowledge support by the Academy of Finland, the Finnish Center of Excellence program 2006-2011, project no. 129670. J.K. thanks OTKA K60456 and TEKES for partial support. We thank Albert-László Barabási for the data used in this research.

APPENDIX A: NOTES ON APPLYING THE METHODS

The Louvain method. The LV agglomeratively builds larger communities until no improvement in modularity can be achieved. Our data yielded very large communities with sizes up to $S \simeq 5 \times 10^5$ nodes for both LV and wLV, and hence we adopted the view that the different renormalization levels correspond to different levels of hierarchical organization,[8] as suggested in Ref. [9]. To obtain meaningful, smaller social communities and to be able to compare results with other methods, we chose to use the first level, i.e., before the first merger of communities was made. This step revealed another feature of LV: While the modularity value is quite similar regardless of the order in which the nodes are processed, the size of the largest community varies greatly. We use a partition where the size of the largest community is around 10^3 because this makes sense in the social context. Because LV uses a local heuristic and we are dealing with a very large network, it is reasonable to assume that the statistical properties of the partitions are on average similar and do not vary as much as the size of the largest community. For a detailed description of the stability of both LV and IM, see Appendix B. In addition the LV algorithm can in some cases produce *disconnected* communities. Only a few such communities were encountered, and we dealt with this by turning each connected component into a community. Code for the algorithm is available for download [31].

The Infomap method. The implementation code for Infomap is available for download [32]. No changes to the code were required.

Clique percolation. For CP we need to select the value of k such that there is no percolating cluster. For our data, $k = 3$

[7]Note that $\mathcal{I}$ may take only values that are fractions of community size, e.g., with $S = 5$ the smallest nonzero value is 0.2, and to get an *average* value of $O(10^{-2})$ the vast majority of IM communities must have $\mathcal{I} = 0$.

[8]Note, however, that this assumption has not yet been verified, e.g., with benchmarks.

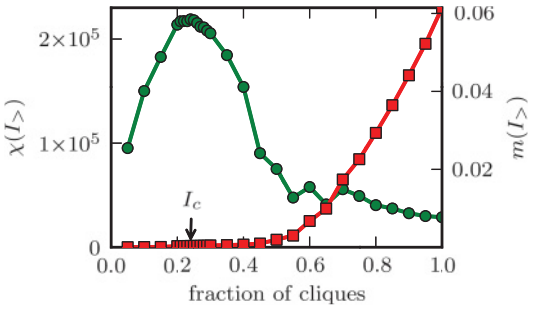


FIG. 10. (Color online) To find the critical threshold I_c for wCP we build up communities by adding cliques in descending order of intensity I, and monitor the largest component size $m(I_>)$ (□) and susceptibility $\chi(I_>)$ (○). The transition occurs when about 24% of cliques have been added ($I_> \approx 3093$).

gives rise to a giant community but $k = 4$ does not, and thus we select $k = 4$.

For the weighted wCP we start with $k = 3$ and find the threshold intensity $I_>$ for which the giant community disappears.[9] Thus we look for the percolation point using clique intensity as the control parameter [23] and set the intensity threshold $I_>$ slightly below the critical point. This point can be identified by the maximum of the susceptibility-like quantity

$$\chi = \sum_{S_\alpha \neq S_{max}} S_\alpha^2 / \left(\sum_\beta S_\beta \right)^2 , \qquad \text{(A1)}$$

where S is community size and α and β index the communities. We varied $I_>$ while monitoring the order parameter $m(I_>)$ and the susceptibility $\chi(I_>)$ (see Fig. 10). When 24% of the cliques have been added in order of descending intensity, a giant cluster emerges, while susceptibility shows a pronounced peak. This point corresponds to the critical intensity $I_c \approx 3093$, which was chosen as our threshold. For CP and wCP, we applied the fast algorithm introduced in Ref. [33]. A sample implementation can be found at Ref. [34].

The running times of all the algorithms used are displayed in Table II. LV and CP are extremely fast, while Infomap takes a few days to complete. All runs were done on a standard desktop machine, utilizing a single processor.

[9]Note that with $k = 4$ the weighted communities would be identical to the unweighted ones, because in the absence of percolation the intensity threshold would be set to 0. Using $k = 2$, on the other hand, would correspond to simply using a weight threshold on single edges.

TABLE II. Running times of the different algorithms on our data set of $N = 4.9 \times 10^6$ nodes and $L = 10.9 \times 10^6$ links.

	Unweighted	Weighted
Louvain	2 min 7 s	1 min 30 s
Infomap	46 h 44 min	3 h 20 min
Clique percolation	2 min 10 s	4 min 52 s

TABLE III. Comparison of the stability of stochastic algorithms. We generated 20 partitions with each method using different random seeds and present the smallest and largest observed values of $|P_i|$ and $S_{\max}$ over all 20 runs and of $f_{\rm pm}^{\rm pair}$ over the 20 ordered pairs (P_i, P_j) with $|i - j| = 1$. The value of $f_{\rm pm}^{\rm all} = |C_{\rm pm}(\{P_i\}_{i=1}^{20})|/|P_j|$ depends on the partition only through $|P_j|$ and is therefore also very stable; we list the value corresponding to the largest $|P_j|$.

	$\lvert P_i\rvert$		$S_{\max}$		$f_{\rm pm}^{\rm pair}$		$f_{\rm pm}^{\rm all}$
IM	280000	280516	2964	3672	42.1%	42.6%	13.2%
LV	1293903	1298256	811	11390	72.1%	72.8%	36.7%
wIM	674587	674727	209	247	97.4%	97.5%	92.5%
wLV	1155557	1155985	73	112	95.8%	95.9%	90.1%

APPENDIX B: STABILITY OF THE STOCHASTIC METHODS

Both IM and LV are stochastic methods, and therefore the partitions produced by different runs will not be identical. To see how stable the algorithms are, we ran each method 20 times with different random seeds to generate partitions $P_i = \{c_{j,i}\}, i = 1, \ldots, 20$, and study the stability of the number of communities found ($|P_i|$), the size of the largest community ($S_{\max} = \max_j\{|c_{j,i}|\}$), and the stability of identified communities across the runs. Let $\mathcal{P} = \{P_1, P_2, \ldots\}$ be a set of partitions and denote by $C_{\rm pm}(\mathcal{P}) = \cap_{P\in\mathcal{P}} P$ the set of communities that appear in all partitions, i.e., the set of perfectly matching communities. For any $P_i \in \mathcal{P}$ the fraction of perfect matches is $f_{\rm pm}(P_i; \mathcal{P}) = |C_{\rm pm}|/|P_i|$. We denote by $f_{\rm pm}^{\rm pair}$ the fraction of perfect matches when $\mathcal{P}$ consists of two partitions, and by $f_{\rm pm}^{\rm all}$ the fraction of perfect matches when $\mathcal{P}$ consists of all 20 partitions generated by a single method.

The results are summarized in Table III. It turns out that both weighted methods are very stable not only with respect to $|P_i|$ and $S_{\max}$, but also with respect to the identity of communities: With both wIM and wLV we get $f_{\rm pm}^{\rm all} > 0.9$, which means that over 90% of communities are identical in all 20 runs. The variation comes mostly from large communities.

In the unweighted case both IM and LV are stable with respect to $|P_i|$, and IM also with respect to $S_{\max}$. The identity of communities found, however, exhibits more variation: e.g., only 13% of communities found by a single run of IM appear in all 20 runs. Furthermore, looking at the unmatched communities for any pair [i.e., those in $P_i \backslash C_{\rm pm}(\{P_i, P_j\})$], in IM about 32% have tiling imperfection $\mathcal{I} < 0.2$, and the average tiling imperfection is 0.46; in LV only 17% of such communities have $\mathcal{I} < 0.2$, with average tiling imperfection of 0.57. Thus the remaining communities are in general not even close matches. As with weighted methods, small communities are more likely to match perfectly than larger ones.

Instability of a method is of course problematic for anyone wanting to identify the "true" communities of a given network. It is, however, premature to judge IM and LV because of this: The network topology is inherently noisy and does not necessarily contain enough information to uniquely identify the communities. Including weights made both methods much more stable, which suggests that the link weights contain information beyond the network topology. Note that there is information even in the instability: Any two IM partitions share

42% of their communities, but if these shared communities were chosen uniformly at random, only $0.42^{20} \approx 10^{-6}\%$ of the communities would appear in all 20 partitions—much less that the actual value of 13.2%.

The high stability of wIM and wLV may be partly explained by the fat-tailed distribution of call lengths in a mobile call network [26]. Since both methods are based on using probabilities proportional to the edge weights, an edge with a weight several orders of magnitude larger than the average will be placed inside a community almost independently of the network topology. On the other hand, in wCP the definition of intensity as the geometric average takes well into account the fat-tailed degree destribution and is equivalent to using weights $w^*_{ij} = \log w_{ij}$, the arithmetic mean for intensity and the intensity threshold $I^*_> = \log I_>$. Although one could use logarithmic weights also with wIM and wLV, this is problematic because the ratio of log weights is not scale invariant, and therefore the result would depend on the unit used to measure call length.

Finally, as suggested by the stability of $|P_i|$ and $S_{\max}$, the qualitative properties of the communities are very stable even though the exact identity of communities are not. For example, IM repeatedly produces treelike communities even if the communities are not made up of the same nodes. Because of this statistical stability no error is made by comparing the methods by using only single realizations from each method.

[1] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).

[2] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).

[3] M. Barahona and L. M. Pecora, Phys. Rev. Lett. **89**, 054101 (2002).

[4] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).

[5] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, Phys. Rev. E **71**, 065103 (2005).

[6] S. Fortunato, Phys. Rep. **486**, 75 (2010).

[7] A. Lancichinetti, S. Fortunato, and J. Kertész, New J. Phys. **11**, 033015 (2009).

[8] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. (2008) P10008.

[10] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato, PLoS ONE **5**, e11976 (2010).

[11] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).

[12] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 016118 (2009).

[13] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).

[14] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature (London) **466**, 761 (2010).

[15] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).

[16] R. Milo, S. Itzkovitz, and N. Kashtan *et al*. Science **303**, 1538 (2004).

[17] G. Palla, A.-L. Barabási, and T. Vicsek, Nature (London) **446**, 664 (2007).

[18] P. F. Jonsson, T. Cavanna, D. Zicha, and P. A. Bates, BMC Bioinformatics **7**, 2 (2006).

[19] M. Herrera, D. C. Roberts, and N. Gulbache, PLoS ONE **5**, e10355 (2010).

[20] M. Rosvall and C. T. Bergstrom, PNAS **105**, 1118 (2008).

[21] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).

[22] U. Brandes, D. Delling, and M. Gaertler *et al*. IEEE Trans. Knowl. Data Eng. **20**, 172 (2008).

[23] I. Farkas, D. Ábel, G. Palla, and T. Vicsek, New J. Phys. **9**, 180 (2007).

[24] J.-P. Onnela, J. Saramäki, and J. Hyvönen *et al*. Proc. Natl. Acad. Sci. U.S.A. **104**, 7332 (2007).

[25] M. C. González, C. A. Hidalgo, and A.-L. Barabási, Nature (London) **453**, 779 (2008).

[26] J.-P. Onnela, J. Saramäki, and J. Hyvönen *et al*. New J. Phys. **9**, 179 (2007).

[27] M. Granovetter, Am. J. Sociol. **78**, 1360 (1973).

[28] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).

[29] A. Arenas, J. Duch, A. Fernandéz, and S. Gómez, New J. Phys. **9**, 176 (2007).

[30] R. Guimerá, S. Mossa, A. Turtschi, and L. A. N. Amaral, Proc. Natl. Acad. Sci. U.S.A. **102**, 7794 (2005).

[31] [http://sites.google.com/site/findcommunities/].

[32] [http://www.tp.umu.se/~rosvall/code.html].

[33] J. M. Kumpula, M. Kivelä, K. Kaski, and J. Saramäki, Phys. Rev. E **78**, 026109 (2008).

[34] [http://www.lce.hut.fi/~mtkivela/kclique.html].

Using explosive percolation in analysis of real-world networks

Raj Kumar Pan,[1] Mikko Kivelä,[1] Jari Saramäki,[1] Kimmo Kaski,[1] and János Kertész[2,1]
[1]*BECS, Aalto University School of Science, P.O. Box 12200, FI-00076 Aalto, Finland*
[2]*Institute of Physics and HAS-BME Condensed Matter Research Group, BME, Budafoki út 8, H-1111 Budapest, Hungary*
(Received 15 October 2010; revised manuscript received 4 March 2011; published 15 April 2011)

We apply a variant of the explosive percolation procedure to large real-world networks and show with finite-size scaling that the university class, ordinary or explosive, of the resulting percolation transition depends on the structural properties of the network, as well as the number of unoccupied links considered for comparison in our procedure. We observe that in our social networks, the percolation clusters close to the critical point are related to the community structure. This relationship is further highlighted by applying the procedure to model networks with predefined communities.

DOI: 10.1103/PhysRevE.83.046112 PACS number(s): 89.75.Fb, 64.60.ah, 89.75.Hc, 89.75.Da

I. INTRODUCTION

The percolation process realized by the *Achlioptas procedure* [1] is different from classical percolation. This "explosive percolation" begins with a graph of isolated nodes and at each step, two potential edges are chosen at random. Then, the edge that minimizes the product or sum of the sizes of the two components that would be merged is added to the graph. This procedure eventually leads to an explosive percolation transition that appears discontinuous (first order). However, it has recently been argued that in reality the transition is continuous and belongs to a new universality class with a very small exponent of the order parameter [2]. The above or similar procedures have been applied to various model networks ranging from regular lattices [3] to scale-free networks [4]. Several papers have painted an intuitive picture of the mechanisms behind this behavior such as local cluster aggregation [5], formation of many large components before percolation transition [6], or inhibition of growth of the largest cluster [7]. Other criteria for the growth process have also been suggested, such as choosing edges proportionally to a weight determined by their cluster sizes [8].

While explosive percolation has triggered a considerable amount of theoretical and simulation work, its application to real-world networks or processes has been limited [9]. The topological characteristics of real-world networks, such as high clustering, degree correlations, community structure, and weight-topology correlations, are far from those of regular or random model graphs [10]. Such features play a role in the characteristics of classical percolation that has earlier been successfully applied to investigate real-world network structure. Here we ask if they also play a crucial role in explosive percolation, and if monitoring the percolation process itself yields important information about the network structure. As a prerequisite, we establish that proper link addition rules yield explosive percolation transitions when applied to real-world networks. However, this depends on both the network structure and the details of the evolution rules.

II. DATA AND METHODS

For our empirical networks, we have chosen a mobile phone call (MPC) network [11] and a large arXiv coauthorship (CA) network [12]. Both networks are social, so that nodes represent people and ties their interactions, and are large enough for percolation studies. They also share features common to social networks, such as community structure and assortativity [10]. For the MPC, it has been shown that tie strengths relate to network topology: Strong ties are associated with dense network neighborhoods (communities) [13]. Such weight-topology correlations are reflected in classical percolation behavior. For the CA, to the best of our knowledge, weight-topology correlations have not been studied in detail before.

The MPC data consist of 325×10^6 voice calls over a period of 120 days. We construct an aggregated undirected weighted network of edges with bidirectional calls between users, weights representing the total number of calls. The largest connected component (LCC) is then extracted, with 4.6×10^6 nodes and 9.1×10^6 edges. The collaboration data is from the arXiv [14] and contains all e-prints in "physics" until March 2010. There are 4.8×10^5 article headers, from which we extract the authors. In the CA network two authors are connected if they have coauthored articles, whose number determines the link weight. We then extract the LCC, with 1.8×10^5 nodes and 9.1×10^6 edges. In addition, we construct a filtered version of the CA, where articles with more than 10 authors ($\sim$2% of articles) are ignored. This is to remove the very large cliques from papers with $\sim 10^3$ authors in fields such as hep-ex or astro-ph, where the principles behind collaboration network formation appear different. The LCC of the resulting small collaboration coauthorship (SCA) network has 1.5×10^5 nodes and 9.1×10^5 edges. Note that, although the number of nodes is not much smaller than for the CA, the number of edges is an order of magnitude less.

For the percolation process, we use the min-cluster (MC-m) sum rule with different values of m, defined as follows. Initially, all the edges of the empirical network are considered unoccupied. Then, at each time step, m unoccupied edges are drawn at random. Of these, the edge that would minimize the size of the component formed if the edge were occupied is chosen. Intracomponent edges are always favored against intercomponent edges as they do not increase the size of any cluster. When comparing two intercomponent edges, we select the one for which the sum of cluster sizes that it connects is minimized. Ties are resolved randomly. We also study the limiting case ($m = \infty$), where all unoccupied edges are considered at each step. This leads to a semideterministic process where all intracluster links get occupied before the

cluster grows in size. The only source of randomness is the existence of clusters of same size during the process [15].

III. RESULTS

A. Percolation analysis

Let us first monitor the behavior of the order parameter; that is, the relative size of the largest cluster, $s_{\max}/N$, as the fraction of occupied edges f_{links} is increased. As intracluster edges do not affect cluster growth, we consider the number of intercluster edges τ instead of f_{links} [16]. We apply three variants of the MC rule, MC-2, MC-10, and MC-∞, as well as random link percolation for comparison. Figures 1(a), 1(b) and 1(c) show the variation of the fraction $s_{\max}(\tau)/N$ against the scaled number of intercomponent edges, τ/N. For all three networks, the transition of the order parameter is smooth for the random case, while for the extreme case, MC-∞, the transition appears abrupt. However, for MC-2 and MC-10, the situation is more complicated, and we study them in detail.

To determine the nature of the transition, Achlioptas *et al.* [1] studied the dependence of the width of the transition window on system size. This width can be quantified as $\Delta \equiv \tau(N/2) - \tau(\sqrt{N})$, where $\tau(N/2)$ and $\tau(\sqrt{N})$ are the lowest values of τ for which $s_{\max} > N/2$ and $s_{\max} > \sqrt{N}$, respectively. In general the width scales as a power law with the system size, $\Delta \propto N^{\zeta}$. For classical percolation, $\zeta = 1$. It was argued that for explosive percolation $\zeta < 1$ and the rescaled width of the transition region, $\Delta/N \propto N^{\zeta-1}$, vanishes in the limit of large N. While recent results [2] argue that the transition region is in reality finite, the very small exponent of the order parameter guarantees that in practice it is vanishingly small even for large systems.

For applying finite-size scaling to empirical networks, samples of different sizes are needed. In general, unbiased sampling of a network is difficult. Here, we take advantage of the known properties of our networks. Call networks are geographically embedded [18], and we extract subnetworks of users in chosen cities, based on postal codes of their subscriptions. For the CA networks, we extract subnetworks of authors with articles in the same subject class. We see that for all networks $\Delta\tau \propto N^{\zeta}$, with $\zeta \sim 1$ for random and $\zeta \sim 0.5$ for the MC-∞ case [Figs. 1(d), 1(e), and 1(f)]. Thus, the exponent ζ clearly differentiates the explosive transition from random-link percolation. Further, for all three networks, $\zeta \sim 1$ for the MC-2, resembling an ordinary percolation transition. However, for MC-10, the scaling exponent behaves differently

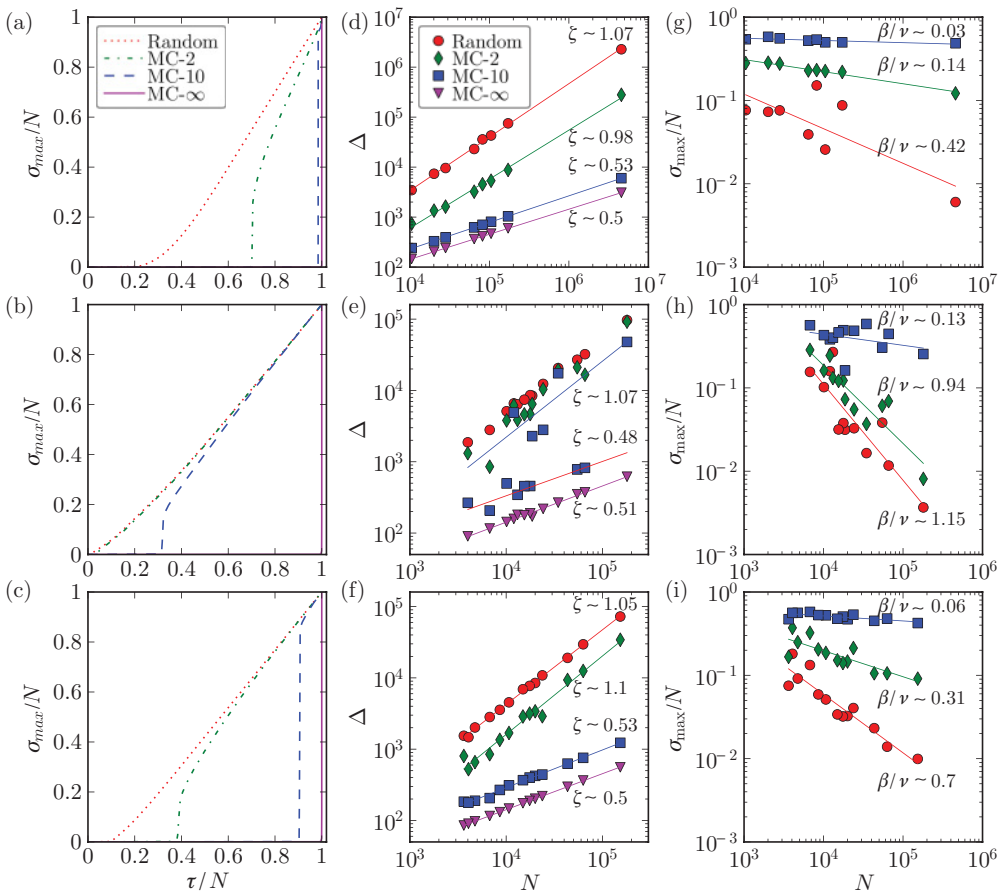


FIG. 1. (Color online) Variation of the relative size of giant component, $s_{\max}/N$, with scaled number of intercluster edges τ/N for the (a) MPC, (b) CA, and (c) SCA network. The corresponding variations in the gap, $\Delta \equiv \tau(N/2) - \tau(\sqrt{N})$, as a function of system sizes are shown in (d), (e), and (f), for the Random, MC-2, MC-10, and MC-∞ rules. Solid lines indicate fitted scaling exponents ζ. The variation of the order parameter, $s_{\max}/N$ as a function of the system size N is shown for (g) MPC, (h) CA, and (i) SCA networks. For each system the order parameter is calculated at the critical point. The solid line indicates the best fit obtained and the exponent β/ν. All curves are averaged over 10^3 runs.

for the three networks. For the MPC and SCA networks, $\zeta \sim 0.5$, indicating explosive percolation. For the CA, at first it appears that the data points do not follow scaling. However, a closer inspection shows that they cluster around two straight lines with $\zeta \sim 1$ and $\zeta \sim 0.5$. Indeed, for subnetworks with large collaborations (e.g., hep-ex, hep-ph) $\zeta \sim 1$, whereas for other subject classes (e.g., cond-mat, math-ph), $\zeta \sim 0.5$.

In addition, we have performed a finite-size scaling analysis of the order parameter $s_{\max}/N$ [19]. The scaling relation for $s_{\max}/N$ is given by

$$\frac{s_{\max}}{N} = N^{-\beta/\nu} F[(\tau - \tau_c)N^{1/\nu}], \tag{1}$$

where F is some universal function, τ is the control parameter, τ_c is the critical point of transition, β is the critical exponent of the order parameter, and ν that of the correlation length. We choose the critical value τ_c of the control parameter as the value of τ where the susceptibility, that is, average cluster size has its maximum. Note that τ_c could also be chosen as the point where the cluster size distribution becomes a power law [2]; however, since our range of network sizes includes fairly small networks, this would be too inaccurate as in some cases there are not enough clusters for determining the shape of the distribution.

For the MPC network [Fig. 1(g)], we find that the scaling at τ_c of the order parameter $s_{\max}/N$ yields a very small exponent $\beta/\nu \sim 0.03$ for the MC-10 case, while for MC-2 and random percolation, the exponents are larger, $\beta/\nu \sim 0.14$ and $\beta/\nu \sim 0.42$, respectively. The exponents for the SCA network behave similarly [Fig. 1(h)], with a low value $\beta/\nu \sim 0.06$ for the MC-10 case and relatively high values $\beta/\nu \sim 0.31$ and $\beta/\nu \sim 0.70$ for MC-2 and random percolation, respectively. In contrast, for the CA network, the exponents have high values for all cases [Fig. 1(i)], $\beta/\nu \sim 0.13$, $\beta/\nu \sim 0.94$, and $\beta/\nu \sim 1.15$, for MC-10, MC-2, and random percolation, respectively.

In order to compare our results to the existing literature, we follow the relations for the critical exponents given in Ref. [2]: $\beta/\nu = \beta/(4\beta + 1)$. The value for the exponent $\beta \sim 0.0555$ given in Ref. [2] yields $\beta/\nu \sim 0.0455$. This value is consistent with our observation that the transition for MC-10 is explosive in the MPC and SCA networks, while it is ordinary in the CA network. Note that such small but finite values of the exponent are consistent with a second-order transition; however, because we are dealing with single, finite-size networks, we cannot make definite conclusions. Further, in all the three systems MC-2 behaves similar to the ordinary random percolation.

Thus, our percolation analysis on CA and SCA networks reveals a difference between collaboration structures in different fields. One possible explanation is the broad degree distribution for the CA network, whose tail can be approximated with a power law with exponent ~1.7 in contrast to SCA, which decays as ~4.3. Hence, in this respect, the SCA network structure resembles the social network of the MPC. Further, it is clear that the nature of the transition depends both on the number of edges m considered in the percolation process and structural features of the network.

For the rest of this paper we focus only on the complete MPC and SCA networks and first study their cluster size distributions around the critical point, τ_c. For the following, we have chosen τ_c as the point at which $P(s)$ is a power law for the full region of s [2]. The complete networks are large enough to choose τ_c this way, giving us in this case a more precise value than the susceptibility peaks. Then we sweep the value of τ around this point and monitor the distribution of cluster sizes. Figures 2(a) and 2(b) show the cluster size distributions $P(s)$ around τ_c for the MCP, for MC-2, MC-10, and MC-∞. For MC-2, $P(s)$ behaves as usual for ordinary percolation, becoming a power law at τ_c and then turning exponential. For MC-10, the situation is different: For $\tau < \tau_c$, there is a bump in the tail of the distribution, in line with theoretical predictions for explosive percolation [2]. Immediately above t_c, the smallest remaining clusters get depleted from the distribution as they are the first to join the giant cluster. For the semideterministic MC-∞ (inset), the cluster size distribution resembles exponential for $\tau < \tau_c$. The cluster size distributions for SCA are qualitatively similar.

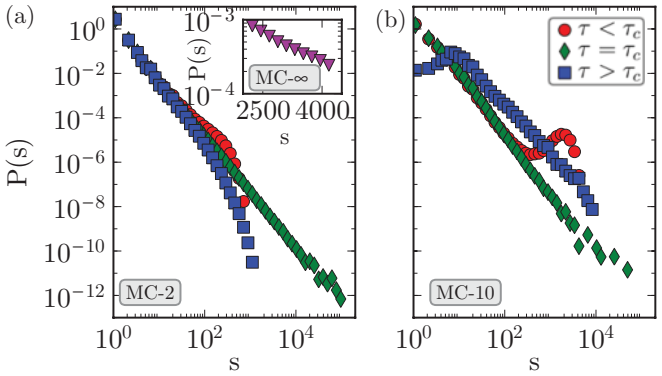


FIG. 2. (Color online) Cluster size distributions around the critical τ_c for the MPC for MC-2 (a), MC-10 (b), and MC-∞ (inset).

B. Percolation clusters, weight-topology correlations, and communities

Next, we investigate the evolution of the percolation clusters and their relationship to communities and the weight-topology correlations. We study the overlap of the neighborhoods of end-point nodes i and j of a link, defined as

$$O_{ij} = n_{ij}/(k_i - 1 + k_j - 1 - n_{ij}), \tag{2}$$

where n_{ij} is the number of neighbors common to both nodes, and k_i and k_j are their degrees [11]. This measure quantifies the extent by which two connected nodes share their neighborhoods: If i and j have no common neighbors, then $O_{ij} = 0$, and if i and j share all of their neighbors, $O_{ij} = 1$. Thus, if there are dense communities in the network, links inside the communities have high values of overlap, whereas links acting as "bridges" connecting separate communities have low overlap values.

Figure 3(a) displays the results for the MPC network. As expected, for random link addition, the overlap and the time when edges are added in the percolation process are uncorrelated. For MC-10 and MC-∞, edges with high overlap and weight are added first [Figs. 3(a) and 3(b)]. This indicates that dense regions of the network, that is communities, get percolated first. Both quantities show an abrupt drop at the transition point. This fits well with the Granovetterian

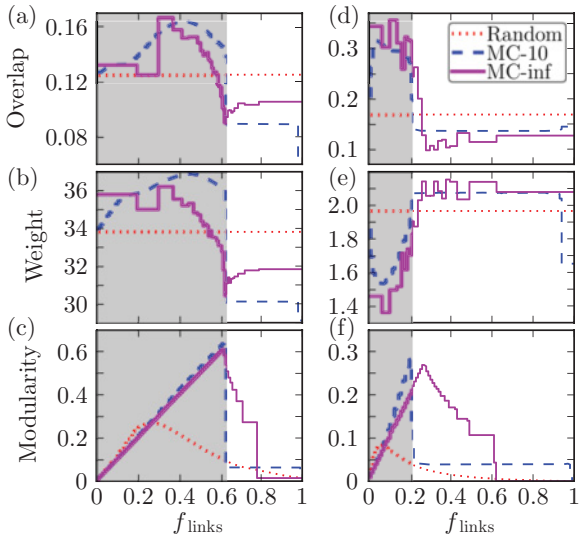


FIG. 3. (Color online) Variation of the overlap, edge weight, and modularity as a function of the fraction of links added, for MPC (a),(b),(c) and SCA (d),(e),(f). The shaded area denotes the nonpercolating regime for MC-10.

weight-topology correlations observed earlier [11]. However, the behavior of the SCA network is different. Although high-overlap edges are added first [Fig. 3(d)], their weights are low [Fig. 3(e)]. This points toward fundamentally different weight-topology correlations, where strong links act as bridges between communities of weaker links. A likely explanation is that communities organize around senior scientists (hubs), with whom junior researchers are linked. The latter has a small number of joint publications with the local hubs, as they are only temporarily connected. The hubs, in turn, are linked via long-lasting collaborations and many coauthored papers.

The relationship to community structure is confirmed with the behavior of the modularity [20] of percolation clusters, defined as

$$\mathcal{M} = \sum_c [(L_c/L) - (d_c/2L)^2], \tag{3}$$

where the sum runs over clusters, L is the number of links in the network, L_c is the number of links within cluster c, and d_c is the sum of the degrees of nodes in c. High values of $\mathcal{M}$ correspond to a good community partition; hence, a high value of modularity calculated for percolation clusters indicates that they match well with communities. As for the other quantities, we calculate $\mathcal{M}$ as a function of the fraction of links added f_{links}. As seen in Figs. 3(e) and 3(f), the peak of $\mathcal{M}$ and the following sharp transition match the transition points well for MC-10. For the semideterministic MC-∞, the peak also matches the percolation point although the transition is less sharp.

C. Analysis of network model with communities

It appears that the explosive percolation process follows community structure when applied to a network where such structure exists. Communities in real-world networks are, however, hard to define unambiguously, and therefore we turn to a simple model with built-in community structure [20,21]. In

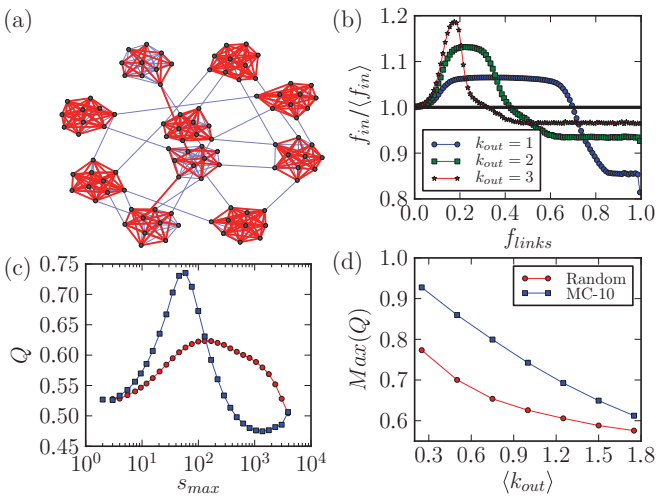


FIG. 4. (Color online) (a) Occupied (red or thick) and unoccupied (blue or thin) edges before the critical point in the model network with the MC-10 rule. Here, $N = 100$, $M = 10$, $k_{\text{in}} = 9.6$, and $k_{\text{out}} = 0.4$. (b) The fraction of intracommunity links f_{in} during the percolation process normalized by the average fraction $\langle f_{\text{in}}\rangle$ for random link addition. (c) Matching quality Q against largest cluster size, s_{max} for the model network. (d) Maximum of the quality, Q_{max}, as a function of k_{out}. All curves are shown for the model with $N = 4096$, $M = 128$, and $k_{\text{out}} + k_{\text{in}} = 16$. For (b) and (c) $k_{\text{out}} = 1$.

this model, N nodes are arranged into M communities of equal size, and edges are placed at random such that on average each node has k_{in} intra-community links and k_{out} intercommunity links. When applying the MC-m sum rule to this network, we find that mostly intracommunity edges are occupied before the transition point [Fig. 4(a)]. We quantify this by measuring the fraction of intracommunity links that have been added during the process, normalized by the respective fraction for random link addition. It is evident from Fig. 4(b) that the MC rules prefer intracommunity links early on in the process, and intercommunity links only get added toward the end.

To quantify the match between percolation clusters and the model communities, we consider the confusion matrix with elements

$$n_{kk'} = |C_k \cap C'_{k'}|, \forall\, k,k', \tag{4}$$

where C_k is the kth cluster and $C'_{k'}$ is the k'th community. Hence, the element $n_{kk'}$ represents the number of nodes in the intersection of cluster C_k and community $C'_{k'}$.

There is a perfect match if clusters are subsets of communities and vice versa, that is, clusters equal communities. The extent to which clusters are subsets of communities can be measured by the projection number of C on C', defined as

$$p_C(C') = \sum_k \max_{k'} n_{kk'}, \tag{5}$$

that is, the sum of the maximum of each row in the confusion matrix. $p_C(C')$ increases with cluster size, reaching its maximum when there is a single cluster that overlaps with all communities. For the reverse case, communities as subsets of clusters, one can define a similar projection number $p_{C'}(C)$, that is, the sum of the maximum of each column in the matrix. This number is maximized when the clusters are

as small as possible, that is, single nodes, and decreases with increasing cluster size [22]. The *quality* of matching can now be quantified with the normalized average of both projection numbers,

$$Q = [p_C(C') + p_{C'}(C)]/2N, \tag{6}$$

reaching its maximum when the match between clusters and communities is optimal.

Figure 4(c) shows the behavior of Q for the model network as a function of the size of the largest observed cluster $s_{\max}$ spanned by the added links. Here we use the largest cluster size $s_{\max}$ instead of f_{links} because this provides us with a more detailed view on what happens around the transition point; the cluster sizes change only a little beyond this region. It is seen that Q initially increases and then decreases as a function of $s_{\max}$, reaching its maximum before the formation of the giant component and merging of clusters. The percolation clusters coincide well with the model communities below and around τ_c compared to random link addition. We next study the behavior of the maximum of quality $Q_{\max}$ as we make the community structure more smeared-out by increasing k_{out} while keeping the average total degree fixed [Fig. 4(d)]. Although $Q_{\max}$ decreases as k_{out} increases for both the MC-10 and random addition, its higher value for the MC-10 process indicates better match with the built-in communities.

We also obtain qualitatively similar results by using normalized the mutual information (NMI) instead of the matching quality Q (not shown). The mutual information [23] can be defined using the confusion matrix as

$$I(C,C') = \sum_{k,k'} \frac{n_{kk'}}{N} \log \frac{n_{kk'}N}{n_k n_{k'}}, \tag{7}$$

where $n_k = \sum_{k'} n_{kk'}$ and $n_{k'} = \sum_k n_{kk'}$ are the size of the kth community and k'th cluster, respectively. The normalized mutual information is then defined as

$$\text{NMI}(C,C') = \frac{2I(C,C')}{H(C) + H(C')}, \tag{8}$$

where $H(C) = -\sum_k n_k/N \log(n_k/N)$ is the entropy of the community C, and $H(C')$ is the entropy of the cluster C'. In our case, where we have a large number of small communities, the NMI does not, however, work as well as the matching quality. This is because the NMI values are high already at the beginning of the percolation process when all the nodes are isolated forming their own clusters. In this case, $\text{NMI}(C,C') = 2(\frac{\log N}{\log M} + 1)^{-1}$, which approaches 1 if the model network size is increased keeping the community sizes, N/M fixed. In contrast, the initial value of the quality is $Q = \frac{1}{2} + \frac{M}{2N}$, which is independent of the number of communities.

IV. SUMMARY AND CONCLUSIONS

To summarize, we have shown that the Achlioptas procedure can give rise to an explosive percolation transition when the rules are applied to empirical real-world social networks. We have used a variant of the minimum cluster (MC) rule, where the number of links compared during the link addition process is a parameter, and shown that both the network structure and the number of links compared have an influence on the universality class (ordinary or explosive) of the observed percolation transition. In order to show this, we have carried out finite-size scaling using subnetworks, chosen on the basis of known external properties of the empirical networks. This is an important but nontrivial task when percolation analysis is applied to empirical networks where only a single "realization" is available. The resulting values for critical exponents are in line with the view that the explosive percolation transition is, in fact, second order; however, one cannot make definite conclusions since we are dealing with singe, finite-size networks.

In addition, we have illustrated a connection between links selected by the MC rule during the percolation process and community structure—at the critical point, the cluster structure arising from the application of the MC rule reflects the community structure of the network. This is confirmed by the analysis of single-link properties (the overlap, link weight), and modularity for the empirical networks, and by detailed studies of the match between clusters and built-in community structure of model networks.

ACKNOWLEDGMENTS

Financial support from EU's 7th Framework Program's FET-Open to ICTeCollective Project No. 238597 and by the Academy of Finland, the Finnish Center of Excellence program 2006–2011, Project No. 129670, as well as by TEKES (FiDiPro) are gratefully acknowledged. We thank A.-L. Barabási for the MPC data used in this research.

[1] D. Achlioptas, R. M. D'Souza, and J. Spencer, Science **323**, 1453 (2009).
[2] R. A. da Costa, S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. Lett. **105**, 255701 (2010).
[3] R. M. Ziff, Phys. Rev. Lett. **103**, 045701 (2009).
[4] F. Radicchi and S. Fortunato, Phys. Rev. Lett. **103**, 168701 (2009); Y. S. Cho, J. S. Kim, J. Park, B. Kahng, and D. Kim, *ibid.* **103**, 135702 (2009).
[5] Y. S. Cho, B. Kahng, and D. Kim, Phys. Rev. E **81**, 030103 (2010); R. M. D'Souza and M. Mitzenmacher, Phys. Rev. Lett. **104**, 195702 (2010).
[6] E. J. Friedman and A. S. Landsberg, Phys. Rev. Lett. **103**, 255701 (2009).
[7] N. A. M. Araújo and H. J. Herrmann, Phys. Rev. Lett. **105**, 035701 (2010).
[8] S. Manna and A. Chatterjee, Physica A **390**, 177 (2011).
[9] H. D. Rozenfeld, L. K. Gallos, and H. A. Makse, Eur. Phys. J. B **75**, 305 (2010).
[10] M. Newman, A. L. Barabasi, and D. J. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, 2006).

[11] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. L. Barabási, Proc. Natl. Acad. Sci. USA **104**, 7332 (2007).
[12] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[13] M. Granovetter, Am. J. Soc. **78**, 1360 (1973); J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész, New J. Phys. **9**, 179 (2007).
[14] [http://arxiv.org/].
[15] A similar rule has been introduced independently in J. S. Andrade *et al.* [e-print arXiv:1010.5097], which appeared after submission of this paper.
[16] In literature, f_{links} has also been used as the control parameter, with slightly different MC rules; for those, intracluster links do play a role [17].
[17] A. A. Moreira, E. A. Oliveira, S. D. S. Reis, H. J. Herrmann, and J. S. Andrade, Phys. Rev. E **81**, 040101 (2010).
[18] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel, J. Stat. Mech. (**2009**) L07003; Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature (London) **466**, 761 (2010).
[19] F. Radicchi and S. Fortunato, Phys. Rev. E **81**, 036110 (2010).
[20] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[21] R. K. Pan and S. Sinha, Europhys. Lett. **85**, 68006 (2009).
[22] S. Van Dongen, Center for Mathematics and Computer Science (CWI), Amsterdam, Technical Report No. INS-R0012, 2000; M. Meila, J. Multivariate Anal. **98**, 873 (2007).
[23] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech. (**2005**) P09008.

The persistence of social signatures in human communication

Jari Saramäki (1), Elizabeth Leicht (2), Eduardo Lopez (2), Sam Roberts (3,4), Felix Reed-Tsochas (2), Robin Dunbar (4)

(1) Department of Biomedical Engineering and Computational Science, School of Science, Aalto University, 00076 Aalto, Espoo, Finland

(2) CABDyN Complexity Centre, Said Business School, University of Oxford, Oxford OX1 1HP, UK

(3) Department of Psychology, University of Chester, CH1 4BJ

(4) Institute of Cognitive and Evolutionary Anthropology, University of Oxford, Oxford OX2 6PN, United Kingdom

Abstract

We have carried out a longitudinal study on the social networks collected over an 18-month period during a major transition in social context when relationships are known to change. Our findings are from an empirical setting that allows us to unify auto-record data, in the form of calls made from mobile phones, with traditional survey data. We look for persistent social signatures in the egos' personal networks at a time of large turnover in the networks. We show that typically, a small number of emotionally close alters in the networks receive a disproportionately large fraction of calls, in line with the layered network view of the social brain hypothesis. Such time allocation patterns display individual variation, and are surprisingly persistent even when the alter composition of the networks undergoes major changes.

1. Introduction

Human beings are above all social animals [1], and having strong and supportive relationships is essential for our health and wellbeing [2]. Whilst there is considerable instability in individual social relationships [3], there is some evidence to suggest a greater level of stability at the level of personal networks – the set of ties an individual (ego) has to their family and friends (alters) [4, 5]. Each ego can therefore be said to have a 'social signature' – the characteristics of their personal network and the communication patterns in that network. However, detailing the nature of this social signature, and particularly how stable this social signature is over time, is complicated by the fact that until recently communication within personal networks has been studied using a survey approach, based on questionnaires or interviews. This has severely limited the level of detail it is possible to collect about personal networks, as well as the reliability of this information and the sample size [12, 13].

Over the last two decades, the increasing use of communication technology has revolutionized the study of social relationships. As each transaction leaves a digital trace, the new field of 'computational social science' [14], is able to analyze communication patterns on a scale, and level of detail, not remotely possible with traditional survey approaches. This has provided important new insights into structure and dynamics of large-scale social networks, involving millions of individuals [8, 15-21]. However, these datasets typically have very limited information on the attributes of, or relationships between, individuals in these networks; consequently it is difficult to differentiate between qualitatively different types of social interactions, purely on the basis of communication patterns. Further, for mobile phone records specifically, the datasets capture only a subset of communication between mobile users on one specific network, and do not capture calls to landlines. Thus, in terms of understanding the persistence of social signatures in personal networks, even computational social science has its limitations.

In this paper, we use a unique 18 month longitudinal dataset, which combines detailed data on communication patterns from mobile phone records with questionnaire data, to explore changes in the personal networks of participants undergoing a major social transition: the move from school to university. We examine whether there is a stable social signature in the egos' personal networks, even when there is a large turnover of individual alters. The nature of this social signature gives novel insights into the underlying structure of personal networks, and the forces shaping that structure.

The social brain hypothesis [1], building on work in primates [6], suggests that there are time and cognitive constraints on the number of relationships an individual can maintain at particular levels of emotional intensity [7, 8]. The operation of these constraints results in a layered structure to personal networks, such that an individual ego can be envisaged as sitting in the centre of a series of concentric circles of acquaintanceship, with the relationships in these layers increasing in number but decreasing in emotional intensity [7, 9]. One way humans sustain social relationships is through regular verbal communication [10]; consequently, unique signatures of the stratified personal network an individual maintains can be observed in communication records.

In studying these communication patterns, by unifying data from mobile phone records with traditional survey data our approach extends previous work in this area in three key ways. First, we have a complete records of all calls an ego made to alters in their personal network over 18 months (including calls to landline numbers), rather than a subset of calls an ego made to alters who happened to be on the same mobile network as them, as has usually been the case in previous work. Having this complete personal network, rather than the partial personal network revealed in typical data based on mobile phone records, is crucial in examining the stability of social signatures over time. Second, by combining information from the phone records and questionnaire data, we are able to uncover the structure of personal networks in more detail, in terms of how the nature of social relationships relates to calling patterns. Finally, we are able to determine the proportion of an ego's personal network captured by the phone records, as well as the characteristics of the alters present in personal networks but *not* present in the phone records. Thus, we are able to establish each ego's social signature at the beginning of the study, and then determine whether this social signature persists during a period of flux for social relationships with many alters both entering and leaving the network [10, 22, 23]. Specifically, we examine whether there is evidence from the communication patterns for the layered structure in personal networks, and whether this layered structure shows persistence over time, despite the turnover of individual alters in the network.

2. Methods

2.1. Personal network survey and call records

We used longitudinal data on the social networks of thirty participants (15 males and 15 females, aged 17 to 19 years old: mean±SD age 18.1±0.48) in their last year of secondary school, collected over an 18-month period during the transition from school to university (for full details, see Roberts & Dunbar [17]). Participants completed a questionnaire on their active personal network at three points in time: at the beginning of the study (t_1), at 9 months (t_2) and at 18 months (t_3). The analysis in this study is based on the 25 participants (12 males, 13 females) who completed all three questionnaires.. To elicit their personal network, participants were asked to list all unrelated individuals "for whom you have contact details and with whom you consider that you have some kind of personal relationship (friend, acquaintance,

someone you might interact with on a regular basis at school, work or university)". The participants were also asked to list all their known relatives. For all individuals listed, participants were asked to provide both landline and mobile phone numbers. In each survey (t_1, t_2 , t_3), for both kin and friends/acquaintances, the participants were asked to indicate the emotional intensity of the relationship by providing an emotional closeness score, measured on a 1-10 scale, where 10 is someone 'with whom you have a deeply personal relationship'. In addition to the social network questionnaire, participants also completed a personality questionnaire at t_1, t_2, and t_3. This measured the 'Big-Five' personality domains – Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Intellect– using the 50-item International Personality Item Pool version of the Revised NEO Personality Inventory [23,24].

At t_1, all participants lived in the same large UK city ('City A'). At month 4 of the study, the participants took their final exams at school ('A-levels') and left the school. Of the 25 participants who completed all three questionnaires, six participants stayed in City A and worked, not going to University. Eight of the participants went to university in City A (which has two large universities) and the remaining 11 participants went to university elsewhere in England.

In compensation for participating in the study, participants were given a mobile phone, with an 18-month contract from a major UK mobile telephone operator. The line rental for the mobile phone was paid for, and included 500 free monthly voice minutes (to landlines or mobiles) and unlimited free text messages. For each participant, we obtained itemized, electronic monthly phone invoices that listed all outgoing calls (recipient phone number, time and duration of calls). The electronic PDF invoices were parsed into machine-readable form. The questionnaire data and this call data form the main basis for our analysis.

2.2. Constructing ego-centric call networks

For each participant in the study (ego), we used the list of kin and friends/acquaintances (alters) generated in response to the three social network questionnaires and combined it with the electronic phone invoices to construct a set of ego-centric call networks. If an alter was listed as having multiple phone numbers, a mobile and a fixed line number, a call by the ego to either number was recorded as a

call between ego and alter. Phone numbers appearing on the invoices but not listed in the questionnaire responses were treated as unique alters; however, service numbers (such as those with 0800- suffixes) were filtered out. The 18-month observation period of electronic phone invoices was divided into three consecutive intervals of 6 months each (I_1: March-August, I_2: September-February, I_3: March-August). For each ego in each of the three intervals, we counted the total number of his/her outgoing calls and the number of calls made to each alter. Comparing the ego-alter relationships, as reported by the egos via emotional closeness scores from the survey data, with the egos' real calling behaviour, we determined the fraction of self-reported ego relationships appearing in the calling records. Using the alter-call-counts per interval, we ranked the egos from most called to least called, calculated a time allocation pattern (Zipf plot) depicting the total fraction of calls to an alter as a function of the alter's rank, and calculated average emotional closeness as a function of alter's rank for all 25 egos.

2.3 Comparison of ego-reported relationships to phone call records

In most previous studies of human communication using auto-recorded data [4-11] an alter appears in the data only if there is communication between the ego and alter. Thus, if communication occurs between ego and alter via a channel not being studied (e.g. landline calls, calls on other mobile networks to the one under investigation) the alter is never known. Here we use the list of alters, kin and friends/acquaintances, from the survey data and the ego-reported emotional closeness score for these alters to understand the characteristics of those alters missing from the data call pattern analysis.

Let us consider the calling behaviour of each ego towards its alters of varying emotional closeness. Let $A(g,c_i,t_i)$ be the set of alters of ego g called in time interval t_i that were categorized during time interval t_i with emotional closeness c_i. Similarly, let $L(g,c_i,t_i)$ be the set of alters of specified with emotional closeness c_i during time interval t_i that were "callable" by ego g. An alter was "callable" during time interval t_i if the alter was first listed in the survey data corresponding to interval t_j or was in the set $A(g,\bullet, t_j)$ where $t_i \geq t_j$. The fraction of alters called by g with emotional closeness c_i in time interval t_i is simply,

$$f(g,c_i,t_i)=\frac{|A(g,c_i,t_i)|}{|L(g,c_i,t_i)|},$$

where numerator and denominator are simply cardinality for each set.

2.4. Analyzing the time allocation patterns

We quantify the variation between the sets of alters an ego calls in two time intervals with the Jaccard coefficient,

$$J\left(A_{I_1},A_{I_2}\right)=\frac{\left|A_{I_1}\cap A_{I_2}\right|}{\left|A_{I_1}\cup A_{I_2}\right|},$$

where A_{I1} and A_{I2} are the full sets of alters who were called by the ego in two time intervals I_1 and I_2, respectively. J=1 if the sets are equal, and 0 if the sets have no common alters. For a pairwise comparison of the time allocation patterns between two different egos or two different time intervals for a single ego we measure the Jensen-Shannon divergence (JSD) [18] defined as

$$JSD(P_1,P_2)=H\left(\tfrac{1}{2}P_1+\tfrac{1}{2}P_2\right)-\tfrac{1}{2}\left[H(P_1)+H(P_2)\right],$$

where P_1 and P_2 are two distributions, $P=\{p(r)\}$ and p(r) is the fraction of calls to the alter of rank r ; additionally, $H(P)$ is the Shannon entropy,

$$H(P)=-\sum_{r=1}^{k}p(r)\log p(r),$$

where p(r) is as above and k is the maximum rank, i.e. the total number of alters called. The Jensen-Shannon divergence is a generalized form of the Kullback-Leibler divergence (KLD) such that $JSD(P_1,P_2)\in[0,\infty)$, and $JSD(P_1,P_2)=0$ if and only if the distributions are identical. We chose JSD over KLD due to its capacity to deal with zero probabilities $p(r)$=0. The maximum number of alters called by an ego in a given time interval, k , varies depending on the ego and the interval; therefore, if $k_2>k_1$ is the

larger number, we assign $p_1(r_1)=0$ for $k_1>r_1 \geq k_2$, i.e. zero-pad the series of fractions of calls such that they are of the same length.

3. Results

3.1 Strength of ego-identified relationships and real calling behaviour

The calls egos place to their alters are related to the strength of the ego-alter relationship, as measured by the ego-reported emotional closeness score. In plotting the averages number of alters an ego will call in a 6-month time interval, t_i, as a function of the emotional closeness score, c_i, main panel Fig. 1, we see a positive relationship between fraction of alters called and the average alter emotional closeness score. The small sample size does result in large values for standard deviation, illustrated by the shaded regions. Furthermore, we see that on average an ego will place at least one call within a given 6-month time interval to four out of five alters that the ego scored with emotional closeness 8 or higher. Thus, the alters rated as most emotionally close to the ego are likely to appear as the most frequent contacts in auto-record phone data. However, it is also clear the phone data do not document every close ego-alter relationship. To fully capture an ego's interaction with all of its alters we would need to collect data on phone calls, emails, Facebook communications, face-to-face interactions, etc., a daunting undertaking. From this analysis, it is clear that while there may be some alters missing from an ego's data, the majority of important relationships are included.

Moving beyond the binary accounting of whether or not an ego calls an alter given a particular emotional closeness and time interval, we calculate the average emotional closeness scores of alters by their ranked call frequency. The inset plot in Fig. 1 shows the average emotional closeness, and standard deviation via error bars, of the top 40 most called alters for all egos over all three time intervals. In the inset plot we see average emotional closeness decrease with increasing alter rank. It is not that alters with low emotional closeness scores are excluded from the top ranks of the most called alters, but on average the most called alters do have higher emotional closeness scores than those alters called less frequently.

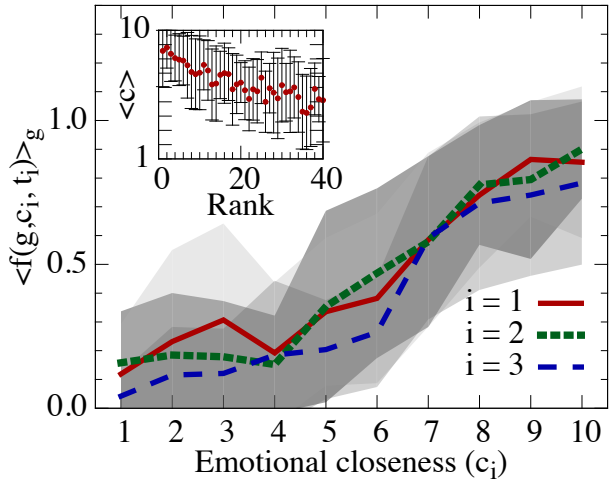


Figure 1 Relationship between call pattern and emotional closeness scores attributed to alters. The main figure illustrates the fraction of alters, averaged over all egos, $\langle f(g,c_i,t_i)\rangle_g$, that are actually called by anego in a 6-month period, t_i, given that the ego scores the alter with emotional closeness c_i. The shaded region indicates the standard deviation. The inset shows the average emotional closeness of alters of varying rank with error bars showing the standard deviation.

3.2. Time allocation patterns and their persistence

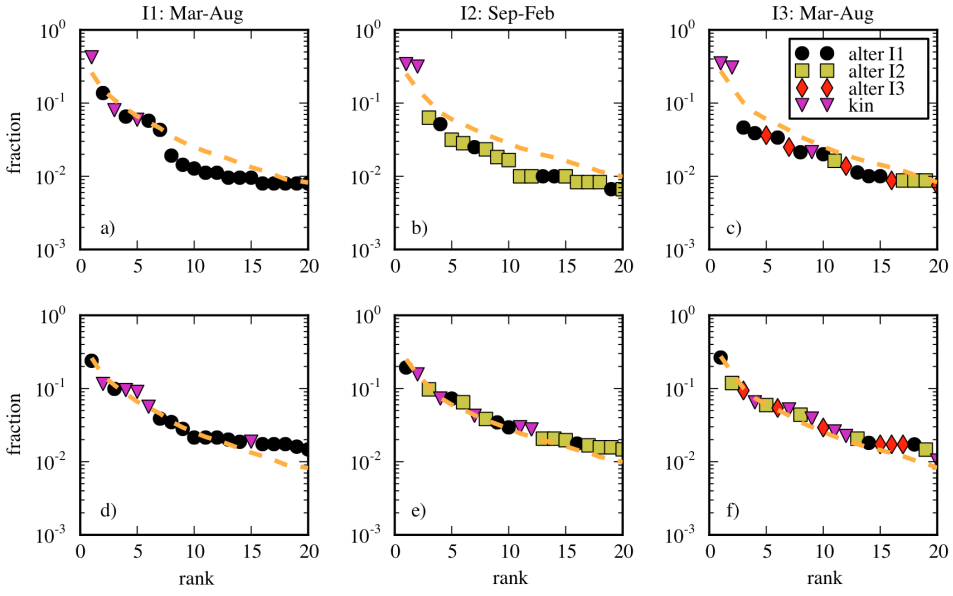


Figure 1 Time allocation patterns for two different egos (survey participants) (top and bottom rows), displaying the fraction of calls to each alter called as a function of alter rank, for the three 6-month time intervals (columns). The symbols correspond to alters observed for the first time in intervals I1 (circles), I2 (squares), and I3 (diamonds), or to kin (triangles) as reported by the egos. The dashed line indicates the time allocation pattern averaged over all 25 egos.

For almost all individuals in the survey, the time allocation patterns are characterized by a heavy tail that decreases slower than exponentially. A large fraction of

communication is typically allocated to a small number of top-ranked alters: for male (female) participants, the fraction of calls to the top alter is on average 0.20 +/- 0.09 (0.26 +/- 0.08), and the fraction of calls to the top three alters is 0.41 +/- 0.12 (0.50 +/- 0.11). A similar tendency to communicate electronically mostly with only a few others has been observed earlier for text messages [10,19] and Facebook [25]. This shape of the time allocation pattern is in line with the layered network view, where the innermost layer contains a small number of alters with close emotional ties that require large maintenance effort.

Figure 2 shows the time allocation patterns for two specific egos for each of the three time intervals, together with a pattern averaged over all 25 egos. The ego whose patterns are depicted in the upper row (panels a to c) is a male who went to university in another city, and the lower row (panels d to f) represents a female student who went to university in City A. For the upper row, the top-ranking alters receive a very large fraction of calls and persistently include two family members (triangles), whereas for the networks in the lower row, the top alters are less dominant, kin are ranked lower, and kin display larger rank fluctuations.

It is also clear on the basis of Figure 2 that the alter composition of the networks undergoes major changes. For both egos shown here, the networks corresponding to the second 6-month interval (I_2) are dominated by newcomers, i.e. alters that were first observed in I_2. This reflects the period of change that the egos are going through: I_2 represents the first six months of the first academic year for those participants who went to university. Overall, as quantified by the Jaccard coefficient, the similarities between the sets of alters in consecutive intervals, averaged over all respondents, are $J(I_1,I_2)$: 0.20 +/- 0.08 and $J(I_2, I_3)$: 0.26 +/- 0.09 for the full set of alters. Thus there is more turnover between intervals I_1 and I_2. However, if we only consider top 20 ranking alters, the similarities are higher: $J(I_1,I_2)$: 0.34 +/- 0.12 and $J(I_2, I_3)$: 0.44 +/- 0.10. Nonetheless, it is clear that the variation is not solely due to high turnover in the lowest ranks.

In order to measure the changes in the time allocation patterns over time, we apply the Jensen-Shannon divergence as a measure of the distance between patterns. In order quantify how similar an individual's patterns for consecutive windows are, we

calculated i) the distances between one ego's pattern for consecutive windows, and ii) the averaged distances between the patterns for the focal ego and all other egos within an interval (see Fig 2 a). We then calculated self and reference distances d_{self} and d_{ref} such that d_{self} was averaged over the two distances between consecutive windows, $d^{i}_{self} = \frac{1}{2}(d^{ii}_{12} + d^{ii}_{23})$, where i indicates the focal ego and sub-indices denote time intervals. Reference distances were averaged for each pair of egos over the three time windows, $d^{ij}_{ref} = \frac{1}{3}(d^{ij}_{11} + d^{ij}_{22} + d^{ij}_{33})$, where j denotes non-focal egos.

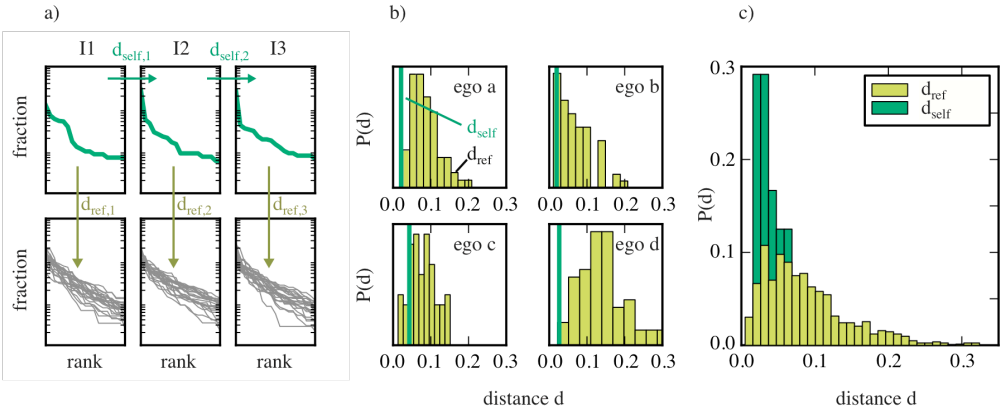


Figure 3 Persistence of time allocation patterns. a) A schematic of how the distances based on Jensen-Shannon divergences are calculated. For the focal participant (top row), self-distances (d_{self}) are calculated for patterns in consecutive intervals and averaged. Reference distances (d_{ref}) are calculated for the focal participant and all other participants within each interval (bottom row) and then averaged. b) Values of the self-distances (d_{self}) and histograms for reference distances (d_{ref}) for four example egos. c) Distributions for self-distances and reference distances for all participants.

The results in Figure 3 (panels b and c) clearly indicate that on average, the shapes of the time allocation patterns of participants (the social signatures) show a tendency of persisting in time, as the distance values d_{self} between one participant's consecutive patterns are on average much lower than the distances d_{ref} to other participants. On average, for each ego, 80%+/-13% of the distances to others were greater than d_{self}. Averaged over all egos, the average self-distance was $<d_{self}>$= 0.037 +/- 0.015 while the average distance to other egos was $<d_{ref}>$ = 0.087 +/- 0.057.

4. Discussion

In this study, we used a unique longitudinal dataset, combining detailed mobile phone call records with three waves of survey data, to examine the personal networks

of participants during a period of natural flux in their social relationships after leaving school. Our key findings can be summarized as follows: first, there is a clear relationship between the emotional intensity of alters and the frequency of calls to them. Second, we have established that the time allocation patterns of call frequency to alters are roughly similar in their signature shape, such that a small number of top-ranked alters receives a disproportionately large fraction of calls. Third, when monitoring the composition of these networks, we find that it undergoes major changes, with many alters entering and leaving the network, and relationships increasing and decreasing in intensity. Importantly, these changes are seen to have surprisingly small effects on the time allocation patterns – thus individuals appear to have a 'social signature' in that they allocate roughly the same amount of time to their alters depending on their rank, independent of who these alters are. Such signature patterns show variation between participants but appear persistent in time for each participant. This constitutes the first direct evidence for the suggestion [Dunbar et al. 1998, Sutcliffe et al. 2011] that social networks are constrained by the time individuals have available for social interaction. Time is an inelastic resource, so if a new alter enters the network, and an ego devotes a lot of time to calling that alter, less time tends to directed towards other alters, and the overall social signature thus remains stable. As such, this confirms one key assumption underpinning the social brain hypothesis thought to be responsible for the layering of social networks [7].

When reflected against the prediction of a layered structure of personal networks from the social brain hypothesis [S1,S7,S9], our observations can broadly speaking be considered in accordance with the main characteristics where the networks comprise a small number of relationships of high emotional intensity, with increasing numbers of relationships of lower emotional intensity. Discrete "layer boundaries" where the communication frequency drops abruptly were observed for some egos (see Fig 2, top row) within some of the time intervals; however, there was a lot of individual variation. This is to be expected: while emotional intensity was seen to correlate with call frequency, calls are only one of the possible communication modalities, and social interactions carried out, e.g., by face-to-face contacts were not included in our analysis.

If the observed overall shape of the time allocation patterns is also assumed to arise from general cognitive constraints, as implied by the social brain hypothesis, how can we interpret the individual variation and the persistence of the ego-specific signature patterns? One possible explanation is that such signatures reflect personality traits. There is a relationship between personality, the size of personal networks and communication patterns within those networks [20-22]. Thus because personality amongst our participants was stable across the study period, this may offer one explanation for the both the heterogeneity of the social signatures we observed, and their stability over time. Indeed, using this dataset, a significant relationship between personality and network size at *t*1 has been demonstrated [Lu et al. (2009].

More broadly, our approach shows the value of combining subjective survey data, (e.g. on the emotional intensity of relationships) with the digital traces of electronically-mediated communication. Both of these sources of data have their limitations, but by combining the two, important insights can be gained about how the objective pattern of communication relates to the nature of our social relationships [e.g. Gilbert & Karahalios, 2009]. Future work could use this combined data to further our understanding of how patterns of communication relate to specific types of social tie. For example, if there clear differences in the patterns of mobile communication between family members, and communication between friends, these can be used to infer social relationships, based soley on communication patterns, from mobile datasets where information on the nature of the social interactions is lacking.

In conclusion, by combining call records and survey data, we demonstrated that distinct patterns of communication - social signatures – show stability over time, despite considerable turnover of individual members of the personal network. This suggests that personal networks in humans, as in primates (Lehmann et al. 2009) are shaped by time and cognitive constraints that limit the number of relationships that can be maintained at each level of emotional intensity, producing a network with a layered structure.

Acknowledgments

The authors wish to thank Renaud Lambiotte for useful discussions. Financial support from EU's 7th Framework Program's FET-Open to ICTeCollective project no. 238597 is acknowledged. JS acknowledges support from the Academy of Finland, the Finnish Center of Excellence program 2006-2011, project no. 129670. SR and RD were supported by the Lucy to Language British Academy Centenary Research Project and the EPSRC/ESRC Developing Theory for Evolving Socio-Cognitive Systems (TESS) project. SR was also supported by an EPSRC Knowledge Secondment Award.

Intro's references:

1. Dunbar R.I.M. 1998 The social brain hypothesis. *Evolutionary Anthropology* **6**(5), 178-190.

2. Holt-Lunstad J., Smith T.B., Layton J.B. 2010 Social relationships and mortality risk: A meta-analytic review. *PloS Medicine* **7**(7). (doi:10.1371/journal.pmed.1000316).

3. Burt R.S. 2000 Decay functions. *Social Networks* **22**(1), 1-28.

4. Lubbers M.J., Molina J.L., Lerner J., Brandes U., Ávila J., McCarty C. 2010 Longitudinal analysis of personal networks. The case of Argentinean migrants in Spain. *Social Networks* **32**(1), 91-104. (doi:10.1016/j.socnet.2009.05.001).

5. Terhell E.L., van Groenou M.I.B., van Tilburg T. 2007 Network contact changes in early and later postseparation years. *Social Networks* **29**(1), 11-24. (doi:10.1016/j.socnet.2005.11.006).

6. Dunbar R.I.M., Shultz S. 2007 Understanding primate brain evolution. *Philosophical Transactions Of The Royal Society B: Biological Sciences* **362**(1480), 649-658. (doi:10.1098/rstb.2006.2001|ISSN 0962-8436).

7. Sutcliffe A., Dunbar R.I.M., Binder J., Arrow H. 2011 Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology*. (doi:10.1111/j.2044-8295.2011.02061.x).

8. Gonçalves B., Perra N., Vespignani A. 2011 Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS One* **6**(8), e22656.

9. Zhou W.X., Sornette D., Hill R.A., Dunbar R.I.M. 2005 Discrete hierarchical organization of social group sizes. *Proceedings Of The Royal Society B-Biological Sciences* **272**(1561), 439-444.

10. Roberts S.G.B., Dunbar R.I.M. 2011 The costs of family and friends: An 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior* **32**(3), 186-197. (doi:10.1016/j.evolhumbehav.2010.08.005).
11. Plickert G., Cote R.R., Wellman B. 2007 It's not who you know, it's how you know them: Who exchanges what with whom? *Social Networks* **29**(3), 405-429. (doi:10.1016/j.socnet.2007.01.007).
12. Marsden P.V. 1990 Network Data And Measurement. *Annual Review Of Sociology* **16**, 435-463.
13. Bernard H.R., Killworth P.D., Sailer L. 1982 Informant accuracy in social network data 5. An experimental attempt to predict actual communication from recall data. *Social Science Research* **11**(1), 30-66.
14. Lazer D., Pentland A., Adamic L., Aral S., Barabasi A.L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., et al. 2009 Computational Social Science. *Science* **323**(5915), 721-723. (doi:10.1126/science.1167742).
15. Onnela J.-P. 2007 Structure and tie strengths in mobile communication networks. *PNAS* **104**(18), 7332-7336.
16. Palla G., Barabasi A.L., Vicsek T. 2007 Quantifying social group evolution. *Nature* **446**(7136), 664-667. (doi:10.1038/nature05670).
17. Lambiotte R., Blondel V.D., de Kerchove C., Huens E., Prieur C., Smoreda Z., Van Dooren P. 2008 Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications* **387**(21), 5317-5325. (doi:10.1016/j.physa.2008.05.014).
18. Liben-Nowell D., Kleinberg J. 2008 Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences* **105**(12), 4633-4638. (doi:10.1073/pnas.0708471105).
19. Kossinets G., Watts D.J. 2006 Empirical analysis of an evolving social network. *Science* **311**(5757), 88-90.
20. Eagle N., Pentland A., Lazer D. 2009 Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America* **106**(36), 15274-15278. (doi:10.1073/pnas.0900282106).
21. Wu Y., Zhou C., Xiao J., Kurths J., Schellnhuber H.J. 2010 Evidence for a bimodal distribution in human communication. *Proceedings of the National Academy of Sciences* **107**(44), 18803-18808. (doi:10.1073/pnas.1013140107).

22. Cummings J.N., Lee J.B., Kraut R. 2006 Communication technology and friendship during the transition from high school to college. In *Computers, phones and the internet: Domesticating information technologies* (eds. Kraut K., Brynin M., Kiesler S.), pp. 265-278. New York, Oxford University Press.

23. Oswald D.L., Clark E.M. 2003 Best friends forever?: High school best friendships and the transition to college. *Personal Relationships* **10**(2), 187-196.

Later References:

1 Lazer, D. et al. 2009 Computational Social Science. *Science* **323**, 721–723. (doi:10.1126/science.1167742)

2 Wasserman, S. & Faust, K. 1994 *Social network analysis: methods and applications*. Cambridge University Press.

3 Marsden, P. V. 1990 Network data and measurement. *Annual Review of Sociology* **16**, 435–463.

4 Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. & Barabási, A.-L. 2007 Structure and tie strengths in mobile communication networks. *PNAS* **104**, 7332–7336. (doi:10.1073/pnas.0610245104)

5 Palla, G., Barabási, A.-L. & Vicsek, T. 2007 Quantifying social group evolution. *Nature* **446**, 664–667. (doi:10.1038/nature05670)

6 Lambiotte, R., Blondel, V. D., Kerchove, C. de, Huens, E., Prieurc, C., Smoredac, Z. & Doorena, P. V. 2008 Geographical dispersal of mobile communication networks. *Physica A* **387**, 5317–5325. (doi:doi:10.1016/j.physa.2008.05.014)

7 Liben-Nowell, D. & Kleinberg, J. 2008 Tracing information flow on a global scale using Internet chain-letter data. *PNAS* **105**, 4633–4638. (doi:10.1073/pnas.0708471105)

8 Kossinets, G. & Watts, D. J. 2009 Origins of Homophily in an Evolving Social Network. *American Journal of Sociology* **115**, 405–450. (doi:10.1086/599247)

9 Eagle, N., Pentland, A. (Sandy) & Lazer, D. 2009 Inferring friendship network structure by using mobile phone data. *PNAS* **106**, 15274–15278. (doi:10.1073/pnas.0900282106)

10 Wu, Y., Zhoud, C., Xiao, J., Kurths, J. & Schellnhuber, H. J. 2010 Evidence for a bimodal distribution in human communication. *PNAS* **107**, 18803–18808. (doi:10.1073/pnas.1013140107)

11 Gonçalves, B., Perra, N. & Vespignani, A. 2011 Validation of Dunbar's number in Twitter conversations. *PLoS ONE* **6**, e22656. (doi:10.1371/journal.pone.0022656)

12 Szella, M. & Thurner, S. 2010 Measuring social dynamics in a massive multiplayer online game. *Social Networks* **32**, 313–329. (doi:10.1016/j.socnet.2010.06.001)

13 Dunbar, R. I. M. 1998 The Social Brain Hypothesis. *Evolutionary Anthropology* **6**, 178–190. (doi:10.1002/(SICI)1520-6505(1998)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8)

14 Dunbar, R. I. M. 1992 Neocortex size as a constraint on group size in primates. *Journal of Human Evolution* **22**, 469–493. (doi:10.1016/0047-2484(92)90081-J)

15 Hill, R. A. & Dunbar, R. I. M. 2003 Social network size in humans. *Human Nature* **14**, 53–72. (doi:10.1007/s12110-003-1016-y)

16 Dunbar, R. I. M. 1993 Coevolution of neocortex size, group size and language in humans. *Behavioral and Brain Sciences* **16**, 681–735. (doi:10.1017/S0140525X00032325)

17 Roberts, S. G. B. & Dunbar, R. I. M. 2011 The costs of family and friends: an 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior* **32**, 186–197. (doi:10.1016/j.evolhumbehav.2010.08.005)

18 Lin, J. 1991 Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **37**, 145–151.

19 Reid, D. J. & Reid, F. J. M. 2006 Textmates and text circles: Insights into the social ecology of SMS text messaging. In *Mobile world: past, present and future* (eds L. Hamill & A. Lasen), New York: Springer.

20 Asendorpf, J.B. and Wilpers, S. 1998 Personality effects on social relationships, Journal of Personality and Social Psychology **74**, 1531-1544

21 Neyer, F.J. and Asendorpf, J.B 2001 Personality-relationship transaction in young adulthood, Journal of Personality and Social Psychology **81**, 1190-1204

22 Pollet, T.V, Roberts, S.G.B, and Dunbar, R.I.M. 2011 Extraverts have larger social layers, Journal of Individual Differences, *in press*

23 Costa, P. T. & McCrae, R. R. 1992 Four Ways Five Factors Are Basic. Personality And Individual Differences **13**, 653-665

24 Goldberg, L. R. et al. The international personality item pool and the future of public-domain personality measures. Journal Of Research In Personality **40**, 84-96 (2006).

25 Marlow, C. (2009) Maintained relationships on Facebook. Retrieved 25. April 2011, from http://overstated.net/2009/03

26 Lu, Y-E., Roberts, S. G. B., Dunbar, R. & Lió, P. & Crowcroft, J. (2009). Size matters: Variation in personal network size, personality and effect on information

transmission. *CSE '09: International Conference on Computational Science and Engineering 2009*, *4,* 188-193
doi: 10.1109/CSE.2009.179
27 Lehmann, J., Korstjens, A. H., & Dunbar, R. I. M. (2007). Group size, grooming and social cohesion in primates. Animal Behaviour, 74, 1617-1629. doi: 10.1016/j.anbehav.2006.10.025|ISSN 0003-3472