# Quantitative social group dynamics on a large scale

Gergely Palla[†‡], Albert-László Barabási[†], and Tamás Vicsek[†‡]

[†]Statistical and Biological Physics Research Group of HAS, Pázmány P. stny. 1A, H-1117 Budapest, Hungary,

[‡] Dept. of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary.

[†]Dept. of Physics, University of Notre Dame, IN 46566, USA.

**The rich set of interactions between individuals in the society [1, 2, 3, 4, 5, 6] results in complex community structure, capturing highly connected circles of friends, families, or professional cliques in a social network [3, 7, 8, 9, 10]. Although most empirical studies have focused on snapshots of these communities, thanks to frequent changes in the activity and communication patterns of individuals, the associated social and communication network is subject to constant evolution [6, 11, 12, 13, 14, 15, 16]. Our knowledge of the mechanisms governing the underlying community dynamics is limited, but is essential for a deeper understanding of the development and self-optimisation of the society as a whole [17, 18, 19, 20, 21, 22]. We have developed a new algorithm based on a clique percolation technique [23, 24], that allows, for the first time, to investigate in detail the time dependence of overlapping communities on a large scale and as such, to uncover basic relationships of the statistical features of community evolution. Our focus is on two networks of major interest, capturing the collaboration between scientists and the calls between mobile phone users, observing that their communities are subject to a number of elementary evolutionary steps ranging from community formation to breakup and merging, representing new dimensions in their quantitative interpretation. We find that large groups persist longer if they are capable of dynamically altering their membership, suggesting that an ability to change the composition results in better adaptability and a longer lifetime for social groups. Remarkably, the behaviour of small groups displays the opposite tendency, the condition for stability being that their composition remains unchanged. We also show that the knowledge of the time commitment of the members to a given community can be used for predicting the community's lifetime. These findings offer a new view on the fundamental differences between the dynamics of small groups and large institutions.**

The data sets we consider contain the monthly roster of articles in the Los Alamos cond-mat archive spanning 142 months, with over 30000 authors [25], and the complete record of phone-calls between the customers of a mobile phone company spanning 52 weeks (accumulated over two week long periods), and containing the communication patterns of over 4 million users. Both type of collaboration events (a new article or a phone-call) document the presence of social interaction between the involved individuals (nodes), and can be represented as (time-dependent) links. The extraction of the changing link weights from the primary data is described in the Supplementary Material. In Fig.1a-b we show the local structure at a given time step in the two networks in the vicinity of a randomly chosen individual (marked by a red frame). The communities (social groups represented by more densely interconnected parts within a network of social links) are colour coded, so that black nodes/edges do not belong to any community, and those that simultaneously belong to two or more communities are shown in red. The two networks have rather different local structure: due to its bipartite nature, the collaboration network is quite dense

and the overlap between communities is very significant, whereas in the phone-call network the communities are less interconnected and are often separated by one or more inter-community nodes/edges. Indeed, while the phone record captures the communication between two people, the publication record assigns to all individuals that contribute to a paper a fully connected clique. As a result, the phone data is dominated by single links, while the co-authorship data has many dense, highly connected neighbourhoods. Furthermore, the links in the phone network correspond to instant communication events, capturing a relationship as is happens. In contrast, the co-authorship data records the results of a long term collaboration process. These fundamental differences suggest that any potential common features of the community evolution in the two networks potentially, represent generic characteristics of community formation, rather than being rooted in the details of the network representation or data collection process.

The communities at each time step were extracted with the Clique Percolation Method (CPM) [23, 24], defining a community as a union of all *$k$-cliques* (complete subgraphs of size $k$) that can be reached from each other through a series of adjacent $k$-cliques (where adjacency means sharing $k-1$ nodes) [24, 26, 27]. When applied to weighted networks, the CPM has two parameters: the $k$-clique size $k$, (in Fig.1a-b we show the communities for $k=4$), and the weight threshold $w^*$ (links weaker than $w^*$ are ignored). The criterion for selecting these parameters is discussed in the Supplementary Material. The key feature of the communities obtained by the CPM are that (i) their members can be reached through well connected subsets of nodes, and (ii) the communities may overlap (share nodes with each other). This latter property is essential, since most networks are characterised by overlapping and nested communities [5, 23].

As a first step, it is important to check if the uncovered communities correspond to groups of individuals with a shared common activity pattern. For this purpose we compared the average weight of the links inside communities, $w_c$, to the average weight of the inter-community links, $w_{ic}$. For the co-authorship network $w_c/w_{ic}$ is about 2.9, while for the phone-call network the difference is even more significant, since $w_c/w_{ic} \simeq 5.9$, indicating that the intensity of collaboration/communication within a group is significantly higher than with contacts belonging to a different group [28, 29].

While for coauthors the quality of the clustering can be directly tested by studying their publication records in more detail, in the phone-call network personal information is not available. In this case the zip-code and the age of the users provides additional information for checking the homogeneity of the communities. In Fig.1c we show the size of the largest subset of people having the same zip code in the communities, $\langle n_{\rm real}\rangle$, averaged over the time steps, as the function of the community size $s$, divided by $\langle n_{\rm rand}\rangle$, representing the average over random sets of users. The significantly higher number of people with the same zip-code in the CPM communities as compared to random sets indicates that the communities usually correspond to individuals living relatively close to each other. It is of specific interest that $\langle n_{\rm real}\rangle / \langle n_{\rm rand}\rangle$ has a prominent peak at $s \simeq 35$, suggesting that communities of this size are geographically the most homogeneous ones. However, as Fig.1d shows, the situation is more complex: on average, the smaller communities are more homogeneous, but there is still a noticeable peak at $s \simeq 30-35$. In Fig.1c we also show the average size of the largest subset of members with an age falling into a three years wide time window, divided by the same quantity obtained for randomly selected groups of individuals. The fact that the ratio is larger than one indicates that communities have a tendency to contain people from the same generation, and the $\langle n_{\rm rand}\rangle / s$ plot indicates that the homogeneity of small groups is on average larger than that of the big groups. In summary, the phone-call

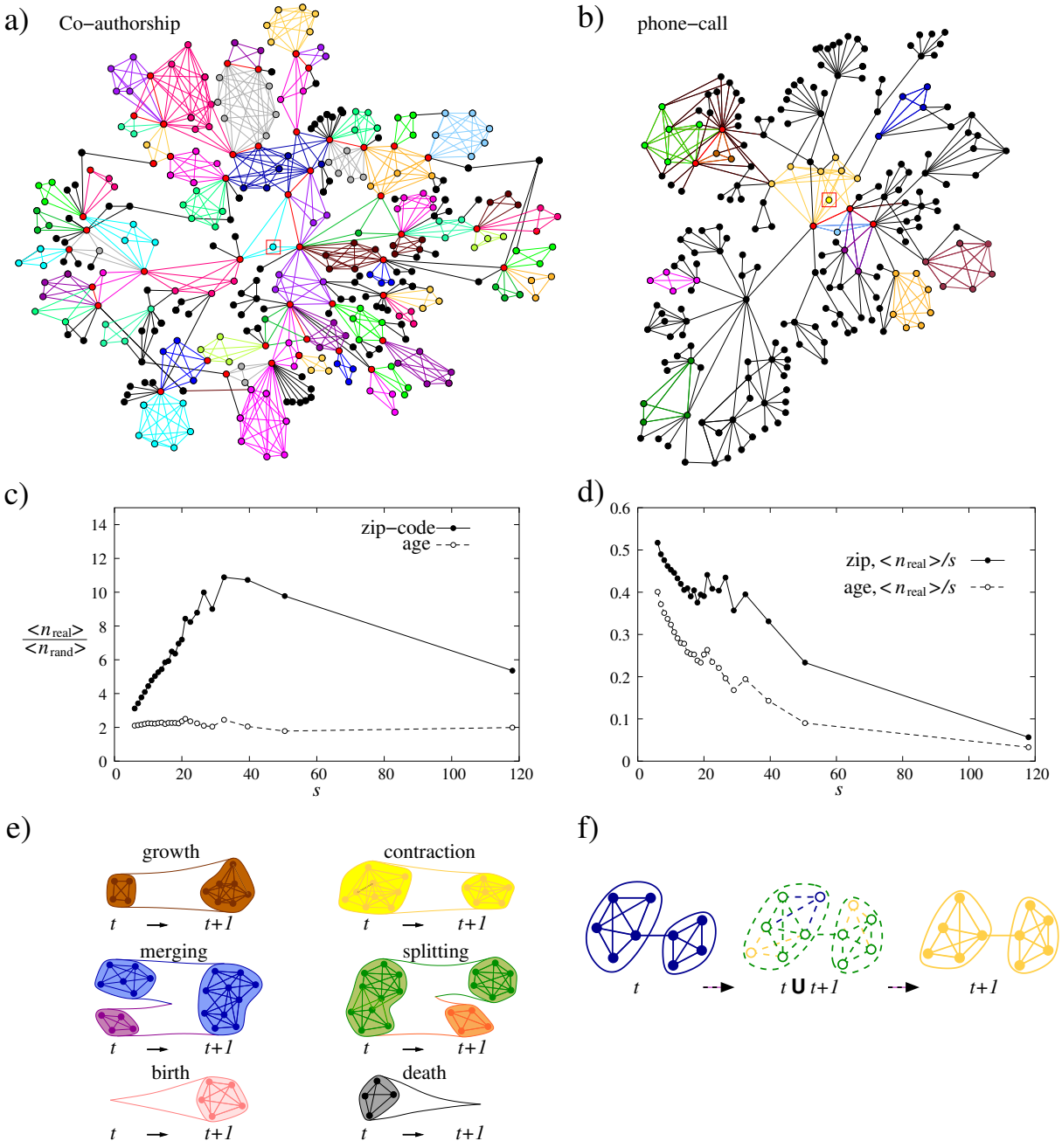


Figure 1: a) The local community structure at a given time step in the vicinity of a randomly selected node in case of the co-authorship network. b) The same picture in the phone-call network. c) The black symbols correspond to the average size of the largest subset of members with the same zip-code, $\langle n_{\rm real} \rangle$, in the phone-call communities divided by the same quantity found in random sets, $\langle n_{\rm rand} \rangle$, as the function of the community size $s$. Similarly, the white symbols show the average size of the largest subset of community members with an age falling in a three year time window, divided by the same quantity in random sets. d) The $\langle n_{\rm real} \rangle / s$ as a function of $s$, for both the zip-code (black symbols) and the age (white symbols). e) Possible events in the community evolution. f) The identification of evolving communities. The links at $t$ (blue) and the links at $t+1$ (yellow) are merged into a joint graph (green). Any CPM community at $t$ or $t+1$ is part of a CPM community in the joined graph, therefore, these can be used to match the two sets of communities.

communities uncovered by the CPM tend to contain individuals living in the same neighbourhood, and with comparable age, a homogeneity that supports the validity of the uncovered community structure. Further support is given in the Supplementary Material.

The basic events that may occur in the life of a community are shown in Fig.1e: a community can grow by recruiting new members, or contract by loosing members; two (or more) groups may merge into a single community, while a large enough social group can split into several smaller ones; new communities are born and old ones may disappear. Given that community finding algorithms extract only static "snap-shots" of the community structure, and that a huge number of groups are present at each time step, it is a significant algorithmic and computational challenge to match communities uncovered at different time steps. The basic idea of the algorithm developed by us to identify community evolution is shown in Fig.1f. For each consecutive time steps $t$ and $t+1$ we construct a joint graph consisting of the union of links from the corresponding two networks, and extract the CPM community structure of this joint network (we thank I. Derényi for pointing out this possibility). Any community from either the $t$ or the $t+1$ snap-shot is contained in exactly one community in the joint graph, since by adding links to a network, the CPM communities can only grow, merge or remain unchanged. Thus, the communities in the joint graph provide a natural connection between the communities at $t$ and at $t+1$. If a community in the joint graph contains a single community from $t$ and a single community from $t+1$, then they are matched. If the joint group contains more than one community from either time steps, the communities are matched in descending order of their relative node overlap (see the Supplementary Material).

We first consider two basic quantities characterising a community: its size $s$ and its age $\tau$, representing the time passed since its birth. $s$ and $\tau$ are positively correlated: larger communities are on average older (Fig.2a), which is quite natural, as communities are usually born small, and it takes time to recruit new members to reach a large size. Next we used the auto-correlation function, $C(t)$, to quantify the relative overlap between two states of the same community $A(t)$ at $t$ time steps apart:

$$C_A(t) \equiv \frac{|A(t_0) \cap A(t_0+t)|}{|A(t_0) \cup A(t_0+t)|}, \tag{1}$$

where $|A(t_0) \cap A(t_0+t)|$ is the number of common nodes (members) in $A(t_0)$ and $A(t_0+t)$, and $|A(t_0) \cup A(t_0+t)|$ is the number of nodes in the union of $A(t_0)$ and $A(t_0+t)$. Fig.2b shows the average time dependent auto-correlation function for communities born with different sizes. We find that in both networks, the auto-correlation function decays faster for the larger communities, indicating that the membership of the larger communities is changing at a higher rate. On the contrary, small communities change at a smaller rate, their composition being more or less static. To quantify this aspect of community evolution, we define the *stationarity* $\zeta$ of a community as the average correlation between subsequent states:

$$\zeta \equiv \frac{\sum_{t=t_0}^{t_{\max}-1} C(t,t+1)}{t_{\max}-t_0-1}, \tag{2}$$

where $t_0$ denotes the birth of the community, and $t_{\max}$ is the last step before the extinction of the community. In other words, $1-\zeta$ represents the average ratio of members changed in one step; larger $\zeta$ corresponds to smaller change (more stationary membership).

We observe a very interesting effect when we investigate the relationship between the lifetime $\tau^*$ (the number of steps between the birth and disintegration of a community), the stationarity and the community size. The lifetime can be viewed as a simple measure of "fitness": communities having higher

fitness have an extended life, while the ones with small fitness quickly disintegrate, or are swallowed by another community. In Fig.2c-d we show the average life-span $\langle \tau^* \rangle$ (colour coded) as a function of the stationarity $\zeta$ and the community size $s$ (both $s$ and $\zeta$ were binned). In both networks, for small community sizes the highest average life-span is at a stationarity value very close to one, indicating that for small communities it is optimal to have static, time independent membership. On the other hand, the peak in $\langle \tau^* \rangle$ is shifted towards low $\zeta$ values for large communities, suggesting that for these the optimal regime is to be dynamic, i.e., a continually changing membership. In fact, large communities with a $\zeta$ value equal to the optimal $\zeta$ for small communities have a very short life, and similarly, small communities with a low $\zeta$ (being optimal at large sizes) are disappearing quickly as well.

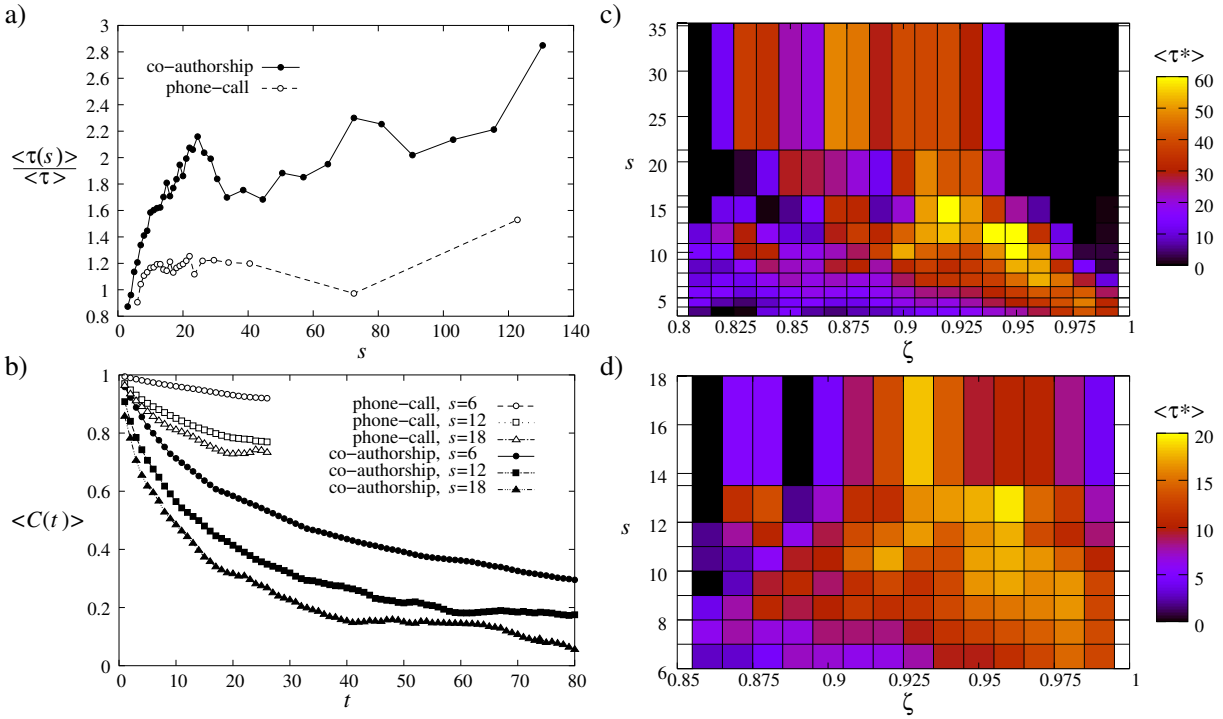


Figure 2: a) The average age $\tau$ of communities with a given size (number of people) $s$, divided by the average age of all communities $\langle \tau \rangle$, as the function of $s$, indicating that larger communities are on average older. b) The average auto-correlation function $C(t)$ of communities with different sizes (the unit of time, $t$, is one month). The $C(t)$ of larger communities decays faster. c) The average life-span $\langle \tau^* \rangle$ of the communities as the function of the stationarity $\zeta$ and the community size $s$ for the co-authorship network. The peak in $\langle \tau^* \rangle$ is close to $\zeta = 1$ for small sizes, whereas it is shifted towards lower $\zeta$ values for large sizes. d)Similar results found in the phone-call network.

To illustrate the difference in the optimal behaviour (a pattern of membership dynamics leading to extended lifetime) of small and large communities, in Fig.3. we show the time evolution of four communities from the co-authorship network. As Fig.3. indicates, a typical small and stationary community undergoes minor changes, but lives for a long time. This is well illustrated by the snapshots of the community structure, showing that the community's stability is conferred by a core of three individuals representing a collaborative group spanning over 52 months. While new co-authors are added occasionally to the group, they come and go. In contrast, a small community with high turnover of its members, (several members abandon the community at the second time step, followed by three new members joining in at time step three) has a lifetime of nine time steps only (Fig.3b). The opposite is

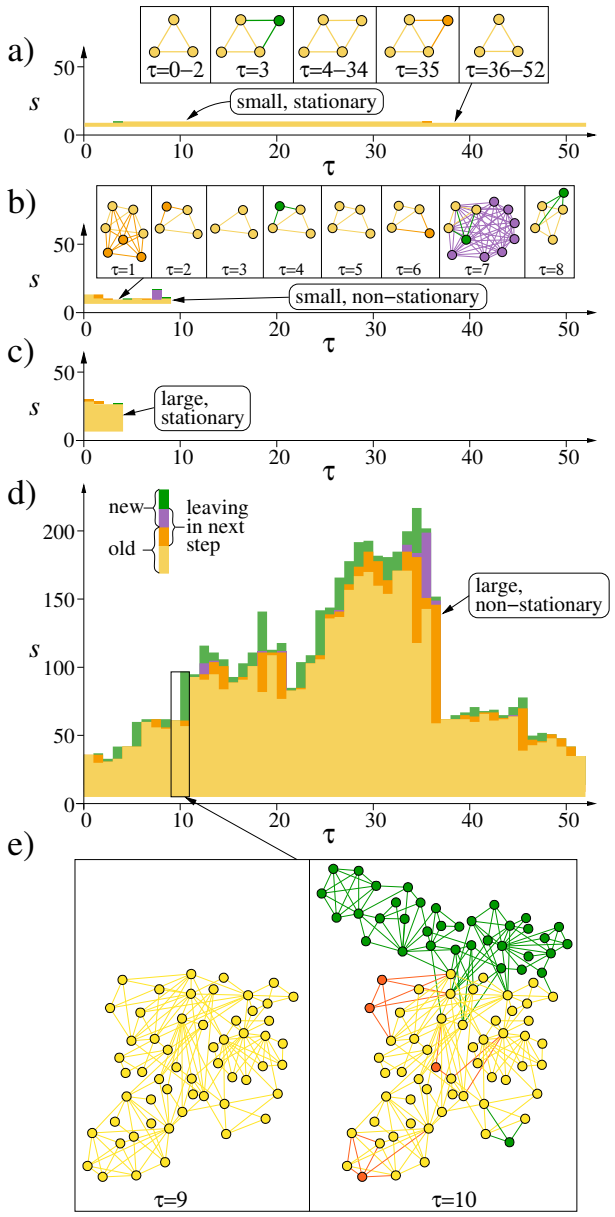


Figure 3: Time evolution of four communities in the co-authorship network. The height of the columns corresponds to the actual community size, and within one column the yellow colour indicates the number of ”old” nodes (that have been present in the community at least in the previous time step as well), while newcomers are shown with green. The members abandoning the community in the next time step are shown with orange or purple colour, depending on whether they are old or new. (This latter type of member joins the community for only one time step). From top to bottom, we show a small and stationary community (a), a small and non-stationary community (b), a large and stationary community (c) and, finally, a large and non-stationary community (d). A mainly growing stage (two time steps) in the evolution of the latter community is detailed in panel e).

seen for large communities: a large stationary community disintegrates after four time steps (Fig.3c). In contrast, a large non-stationary community whose members change dynamically, resulting in significant fluctuations in both size and the composition, has quite extended lifetime (Fig.3d). Indeed, while the community undergoes dramatic changes, gaining (Fig.3e) or loosing a high fraction of its membership,

it can easily withstand these changes.

The quite different stability rules followed by the small and large communities raise an important question: could an inspection of the community itself predict its future? To address this question, for each member in a community we measured the total weight of this member's connections to outside of the community ($w_{\rm out}$) as well as to members belonging to the same community ($w_{\rm in}$). We then calculated the probability that the member will abandon the community as a function of the $w_{\rm out}/(w_{\rm in} + w_{\rm out})$ ratio. As Fig.4a shows, for both networks this probability increases monotonically, suggesting that if the relative commitment of a user is to individuals outside a given community is higher, then it is more likely that he/she will leave the community. In parallel, the average time spent in the community by the nodes, $\langle \tau_{\rm n} \rangle$, is a decreasing function of the above ratio (Fig.4a inset). Individuals that are the most likely to stay are those that commit most of their time to community members, an effect that is particularly prominent for the phone network. As Fig.4a shows, those with the least commitment have a quickly growing likelihood of leaving the community. Taking this idea from individuals to communities, we measured

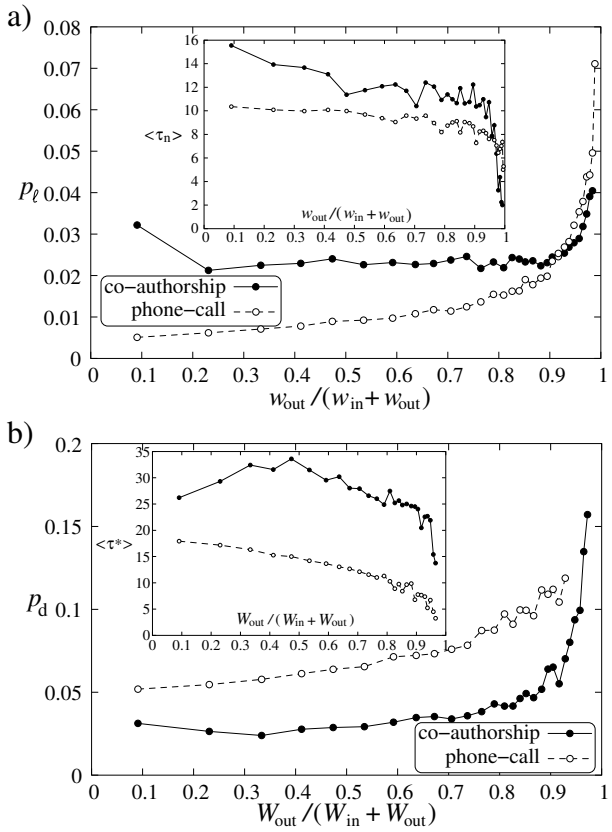


Figure 4: a) The probability $p_\ell$ for a member to abandon its community in the next step as a function of the ratio of its aggregated link weights to other parts of the network ($w_{\rm out}$) and its total aggregated link weight ($w_{\rm in} + w_{\rm out}$). The inset shows the average time spent in the community by the nodes, $\langle \tau_{\rm n} \rangle$, in function of $w_{\rm out}/(w_{\rm in} + w_{\rm out})$. b) The probability $p_{\rm d}$ for a community to disintegrate in the next step in function of the ratio of the aggregated weights of links from the community to other parts of the network ($W_{\rm out}$) and the aggregated weights of all links starting from the community ($W_{\rm in} + W_{\rm out}$). The inset shows the average life time $\langle \tau^* \rangle$ of communities as a function of $W_{\rm out}/(W_{\rm in} + W_{\rm out})$.

for each community the total weight of links (a measure of how much a member is committed) from the members to others, outside of the community ($W_{\rm out}$) , as well as the aggregated link weight inside the

community ($W_{\rm in}$). We find that the lifetime of a community decreases with the $W_{\rm out}/(W_{\rm in}+W_{\rm out})$ ratio (Fig.4b inset), indicating that self-focused communities have a significantly longer lifetime than those that are open to the outside world. Taken together, these results suggest that a tracking of the individual's as well as the community's relative commitment to the other members of the community provides a clue for predicting the community's fate: communities whose members limit most of their "bandwidth" to members of their own community have a higher chance of survival.

In summary, our results indicate the significant difference between smaller collaborative or friendship circles and institutions. At the heart of small cliques are a few strong relationships, and as long as these persist, the community around them is stable. Such social groups can afford to add and loose members, as long as the core is not perturbed. For this reason, they do show some decay in their membership correlation with time, but the correlation function stabilises at the core membership. It appears to be almost impossible to maintain this strategy for large communities, however. Thus we find that the condition for stability for large communities is continuous changes in their membership, allowing for the possibility that after some time practically all members are exchanged. Such loose, rapidly changing communities are reminiscent of institutions, that can continue to exist even after all members have been replaced by new members. For example, in a few years most members of a school or a company could change, yet the school and the company will be detectable as a distinct community at any time step during its existence. Thus our results indicate that the key to stability for small groups is continuity in membership. In contrast, the key to stability for large groups and institutions is an ability to change by constantly accepting new members, and letting old members leave. We expect that our approach, allowing the quantitative analysis of overlapping social group dynamics for very large networks for the first time, will be useful in establishing a qualitative classification of social groups as well. In addition, results like those presented in Fig.2c-d can be used to predict the expected lifetime of a given group which could have great practical implications. In particular, if a large community evolves into a more stationary state (its structure and membership becomes more "rigid") it is likely to become more fragile and eventually disappear.

## Acknowledgement

We thank I. Derényi for useful suggestions. This work was supported by grants from OTKA Nos: F047203 and T034995.

## References

[1] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

[2] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

[3] Scott, J. *Social Network Analysis: A Handbook*, 2nd ed. (Sage Publications, London, 2000).
*Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

[4] Watts, D. J., Dodds, P. S., & Newman, M. E. J. Identity and search in social networks. *Science* **296**, 1302–1305 (2002).

[5] Faust, K. Using Correspondence Analysis for Joint Displays of Affiliation Networks. *Models and Methods in Social Network Analysis* (Eds Carrington, P., Scott, J., & Wasserman, S.) Ch. 7 (Cambridge University Press, New York, 2005).

[6] Liljeros, F., Edling, Ch. R., Amaral, L. A. N., Stanley H. E., & Aberg, Y. The Web of Human Sexual Contacts. *Nature* **411**, 907–908 (2001).

[7] Shiffrin, R. M. & Börner, K. Mapping knowledge domains. *Proc. Natl. Acad. Sci. USA* **101** 5183–5185 Suppl. 1 (2004).

[8] Newman, M. E. J. Detecting community structure in networks. *Eur. Phys. J. B*, **38**, 321–330 (2004).

[9] Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821-7826 (2002).

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

[11] Barabási, A.-L., Jeong, H. Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. Evolution of the social network of scientific collaborations . *PHYSICA A* **311**, 590–614 (2002).

[12] Holme, P., Edling, Ch. R., & Liljeros, F. Structure and Time-Evolution of an Internet Dating Community *Social Networks* **26**, 155-174 (2004).

[13] Ebel, H., Davidsen, J., & Bornholdt, S. Dynamics of social networks. *Complexity* **8**, 24-27 (2002).

[14] Wagner, C. S., Leydesdorff, L. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy* **34** 1608–1618 (2005).

[15] Yeung Y.-Y. , Liu, T. C.-Y. , Ng P.-H. A social network analysis of research collaboration in physics education. *American Journal of Physics* **73**, 145–150 (2005).

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

[17] Guimerá, R., Danon, L., Diaz-Guilera, A., Giralt, F., & Arenas, A. Self-similar community structure in organisations. *Physical Review E* **68**, 065103 (2003).

[18] Hopcroft, J., Khan, O., Kulis, B., & Selman, B. Tracking evolving communities in large linked networks. *Proc. Natl. Acad. Sci. USA* **101**, 5249–5253 (2004).

[19] Guimerá, R., Uzzi B., Spiro J., & Amaral, L. A. N. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science* **308** 697–702 (2005).

[20] Li, Ch., Maini, Ph. K. An evolving network model with community structure. *Journal of Physics A: Mathematical and General* **38**, 9741-9749 (2005).

[21] Pollner, P., Palla, G., & Vicsek, T. Preferential attachment of communities: The same principle, but a higher level. *Europhys. Lett.* **73**, 478–484 (2006).

[22] Kossinets, G. & Watts, D. J. Empirical analysis of an evolving social network. *Science* **311**, 88–90 (2006).

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

[24] Derényi, I., Palla, G., & Vicsek, T., Clique percolation in random networks. *Phys. Rev. Lett.* **94**, 160202 (2005).

[25] Warner, S. E-prints and the Open Archives Initiative. *Library Hi Tech* **21**, 151–158 (2003).

[26] Everett, M. G. & Borgatti, S. P. Analyzing clique overlap. *Connections* **21**, 49–61 (1998).

[27] Batagelj, V. & Zaversnik, M. Short cycles connectivity. *arXiv* cs.DS/0308011 (2003).

[28] Granovetter, M. S. The strength of weak ties. *American Journal of Sociology* **78**, 1360–1380 (1973).

[29] Csermely, P. *Weak Links*. (Springer Verlag, Heidelberg, Germany, 2006)