Quantifying social group evolution

Supplementary Information

1 Summary of our main results

In Fig.1. we show a schematic illustration of our main results concerning the statistical properties of the community evolution in social networks. The basic events that may occur in the life of a community are shown in Fig.1a: a community can grow by recruiting new members, or contract by losing 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. A question of great interest connected to these phenomena is the connection between the expected life-span of a community and its other statistical properties. 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. This effect is illustrated in Fig.1b-e, where we show the time evolution of four communities from the co-authorship network investigated. As Fig.1b. 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.1c). The opposite is seen for large communities: a large stationary community disintegrates after four time steps (Fig.1d). 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.1e). Indeed, while the community undergoes dramatic changes, gaining or losing a high fraction of its membership, it can easily withstand these changes.

2 Construction of the networks

In our studies the two time dependent networks were constructed from data concerning *collaboration/communication acts* between the people involved. In case of the Los Alamos cond-mat archive [1], the primary data set contained the monthly roster of articles, (altogether 142 months, over 30000 authors), whereas in case of the phone-call network the phone-calls between the customers were aggregated over two week long periods (altogether 26 periods, over 4 million users). In both cases, we assumed that the *social connection* between people had started some time before the collaboration/communication events and lasted for some time after these events as well. (*e.g.,* the submission of an article to the archive is usually preceded by intense collaboration and reconciliation between the authors, which is in most cases prolonged after the submission as well). Collaboration/communication events between the same people can be repeated from time to time again, and higher frequency of collaboration/communication acts usually indicates closer relationship [2]. Furthermore, weights can be assigned to the collaboration and communication events quite naturally: an article with n authors corresponds to a collaboration act of weight $1/(n-1)$ between every pair of its authors, whereas the cost of the phone-calls provide the weight in case of the phone-call network. Based on this, we define the *link weight* between two nodes a and b at time t as

$$w_{a,b}(t) = \sum_i w_i \exp\left(-\lambda \left|t - t_i\right| / w_i\right), \tag{1}$$

where the summation runs over all collaboration events in which a and b are involved (*e.g.,* a phone-call between a and b), and w_i denotes the weight of the event i occurring at t_i. (The constant λ is a decay time characteristic for the particular social system we study). Thus, in this approach the time evolution of the network is manifested in the changing of the link weights. However, if the links weaker than a
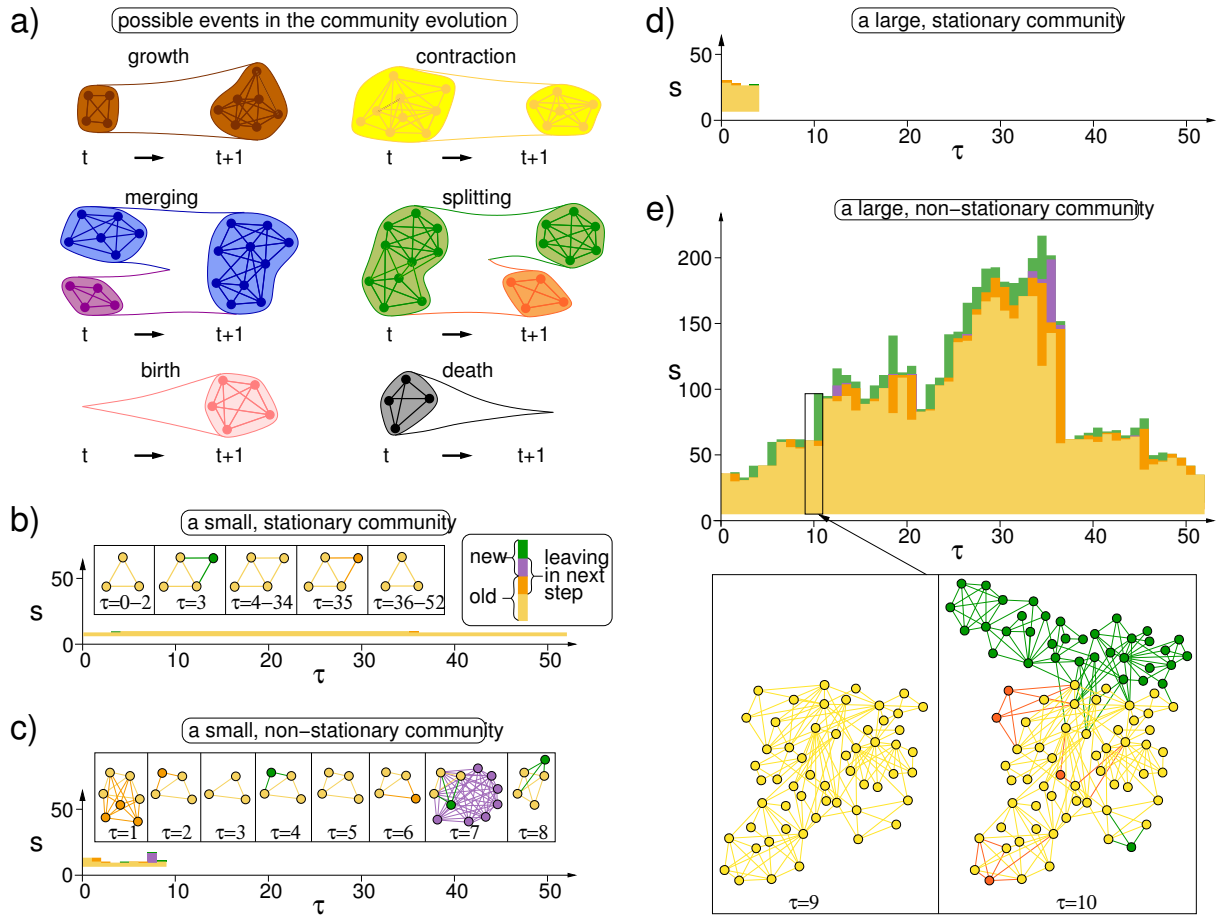
Figure 1: The schematic summary of our main findings related to the evolution communities in social networks. a) The basic events in the community evolution. In panels (b)-(e) we display the time evolution of four communities from the co-authorship network to illustrate the difference in the optimal survival strategy of large- and small communities. 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).

certain threshold w^* are neglected, the network becomes truly restructuring in the sense that links appear only in the vicinity of the events and disappear further away in time, as illustrated in Fig.2. The above method of weighting ties between people is very useful in capturing the continuous time dependence of the strength of connections when the information about them is available only at discrete time steps.

3   The Clique Percolation Method

The communities defined by the Clique Percolation Method (CPM) correspond to k-clique percolation clusters in the network [3, 4]. The k-cliques are complete subgraphs of size k (in which each node is connected to every other nodes). A k-clique percolation cluster is a subgraph containing k-cliques that can all reach each other through chains of k-clique adjacency, where two k-cliques are said to be adjacent
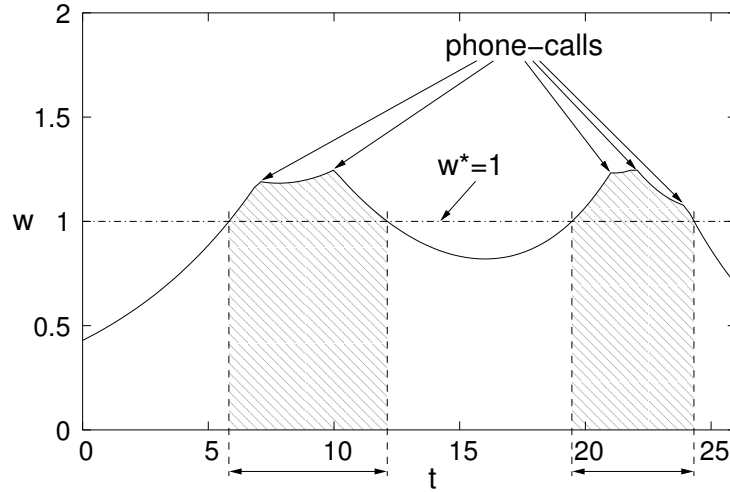
Figure 2: The link-weight as a function of time for a connection in the phone-call network. If a weight threshold of $w^* = 1$ is introduced, the link is absent outside the shaded intervals.

if they share $k - 1$ nodes. The k-clique percolation clusters can be best visualised with the help of k-clique templates, that are objects isomorphic to a complete graph of k vertices. As shown in Fig.3., such a template can be placed onto any k-clique in the graph, and rolled to an adjacent k-clique by relocating one of its vertices and keeping its other $k - 1$ vertices fixed. Thus, the k-clique percolation clusters (k-clique communities) of a graph are all those subgraphs that can be fully explored by rolling a k-clique template in them but cannot be left by this template.
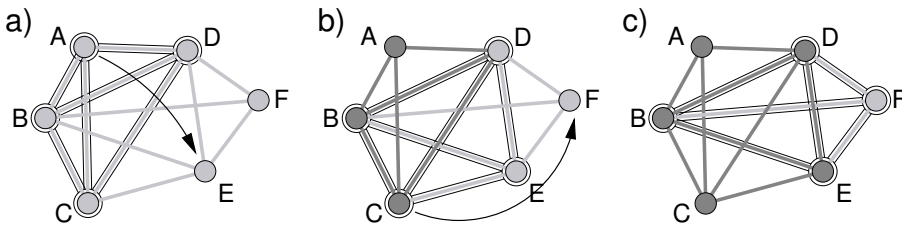

Figure 3: Illustration of k-clique template rolling for $k = 4$. a) In the first step, the template is placed on the k-clique $ABCD$. b) In the next step it is "rolled" to $BCDE$. c) Finally, the template reaches $BDEF$.

When applied to weighted networks, the CPM method has two parameters: the k-clique size k, and a weight threshold w^* (links weaker than w^* are ignored). By increasing k or w^*, the communities start to shrink and fall apart, but at the same time they become also more cohesive. In the opposite case, at low k there is a critical w^*, under which a giant community appears in the system that smears out the details of the community structure by merging (and making invisible) many smaller communities. The criterion used to fix these parameters is based on finding a community structure as highly structured as possible: at the highest k value for which a giant community may emerge, the w^* is decreased just below the critical point. The actual values of these parameters in our studies were $k = 3, w^* = 0.1$ in case of the co-authorship network, and $k = 4, w^* = 1.0$ in case of the phone-call network.

The main advantages of the CPM are that its community definition *is not too restrictive, it is local, it is based on the density of the links and it allows overlaps between the communities*: a node can be part of several k-clique percolation clusters at the same time. The significance of community overlaps can be seen in *e.g.,* Fig.1b in the main text, showing the local community structure in the vicinity of a

randomly selected node from the co-authorship network. Previous studies revealed, that the occurrence of overlapping and nested communities is common in other real networks as well [3].

4 Basic statistics of the community structure

One of the most basic properties characterising the partitioning of a network is the overall coverage of the community structure, *i.e.* the ratio of nodes contained in at least one community. In case of the co-authorship network the average value of this ratio was above 59%, which is a reasonable coverage for the CPM. In contrast, we could only achieve a significantly smaller ratio for the phone-call network. At such a large system size, in order to be able to match the communities at subsequent time steps in reasonable time we had to decrease the number of communities by choosing a higher k and w^* parameter ($k = 4$ and $w^* = 1.0$), and keeping only the communities having a size larger or equal to $s = 6$. Therefore, in the end the ratio of nodes contained in at least one community was reduced to 11%. However, this still means more than 400000 customers in the communities on average, providing a representative sampling of the system. By lowering the k to $k = 3$, the fraction of nodes included in the communities is raised to 43%. Furthermore, a significant number of additional nodes can be also classified into the discovered communities. For example, if a node not yet classified has link(s) only to a single community (and, if it has no links connecting to nodes in any other community) it can be safely added to that community. Carrying out this process iteratively, the fraction of nodes that can be classified into communities increases to 72% for the k=3 co-authorship network, and to 72% (61%) for the k=3 (k=4) mobile phone network, which, in principle, allows us to classify over 2.4 million users into communities.

   Another important statistics describing the community system is the community size distribution. In Fig.4a we show the community size distribution in the phone-call network at different time steps. They all resemble to a power-law with a high exponent. In case of $t = 0$, the largest communities are somewhat smaller than in the later time steps. This is due to the fact that the events before the actual time step cannot contribute to the link-weights in case of $t = 0$, whereas they can if $t > 0$. In Fig.4b we can follow the time evolution of the community size distribution in the co-authorship network. In this case $t = 0$ corresponds to the birth of the system itself as well (whereas in case of the phone-calls it does not), therefore the network and the communities in the network are small in the first few time steps. Later on, the system is enlarged, and the community size distribution is stabilised close to a power-law. In Figs.4c-d we show the number of communities as a function of the community size at different time steps in the examined systems. For the phone-call network (Fig.4c), this distribution is more ore less constant in time. In contrast, (due to the growth of the underlying network) we can see an overall growth in the number of communities with time in the co-authorship network (Fig.4d). Since the number of communities drops down to only a few at large community sizes in both systems, we used size binning when calculating the statistics shown in the manuscript.

5 Further properties of the phone-call communities

According to Fig.1c in the main text, the zip-code and the age of people in a phone-call community is on average more similar than in a random set of the same size drawn uniformly from the available users. However, the difference in the homogeneity of the age is less pronounced than in case of the zip-code. A plausible reason for this effect is that due to the strong social relation between parents and children, many communities contain members coming from different generations. This is supported by the distribution of the age difference in communities, shown in Fig.5.: there is a major peak at zero corresponding to members with the same age, however there is also another peak at 25, corresponding to the typical age difference between parents and children.
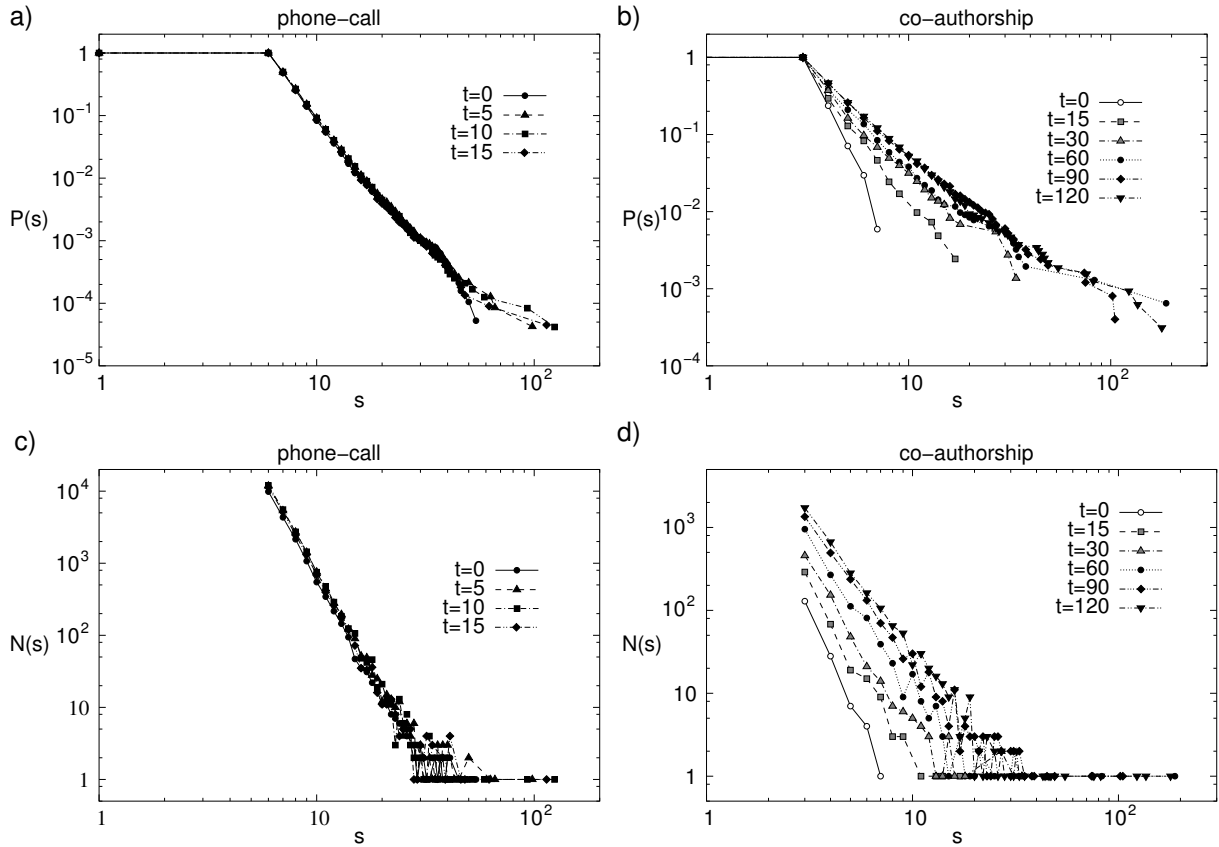
Figure 4: a) The cumulative community size distribution in the phone-call network at different time steps. b) The time evolution of the cumulative community size distribution in the co-authorship network. c) The number of communities of a given size at different time steps in the phone-call network. d) The time evolution of the number of communities with a given size in the co-authorship network.

Beside the zip-code and the age, the statistics of the *service usage* of the customers supports the validity of the communities as well. In our primary data, the number of times people have used a certain service in one of the two weeks long periods was also available. (There were altogether 34 available services for the customers). However, for most services, the probability for a randomly selected customer using the service at all is very low. For this reason, instead of comparing the average number of members using the same service in communities and random sets, we compare the $N_{\mathrm{u}}^{\mathrm{com}}(n_{\mathrm{u}})$ number of communities having n_{u} members using the same service to the same quantity in random sets, denoted by $N_{\mathrm{u}}^{\mathrm{rand}}(n_{\mathrm{u}})$. For each service, random sets with the same size distribution as the communities were constructed 10000 times, and $N_{\mathrm{u}}^{\mathrm{rand}}(n_{\mathrm{u}})$ was averaged over the samples. As it can be seen from Fig.6., for 13 services the $N_{\mathrm{u}}^{\mathrm{com}}(n_u)$ number of communities having n_u members using the service is significantly larger than in case of random sets. In fact, the $N_{\mathrm{u}}^{\mathrm{com}}(n_{\mathrm{u}})/N_{\mathrm{u}}^{\mathrm{rand}}(n_{\mathrm{u}})$ ratio in some cases reaches infinity, indicating that there were no random sets at all containing such high number of service users as some communities.

6 Matching communities at subsequent time steps

After the communities have been extracted at each time step separately, the set of communities at succeeding time steps have to be matched with each other. This raises the problem of finding the pre-image
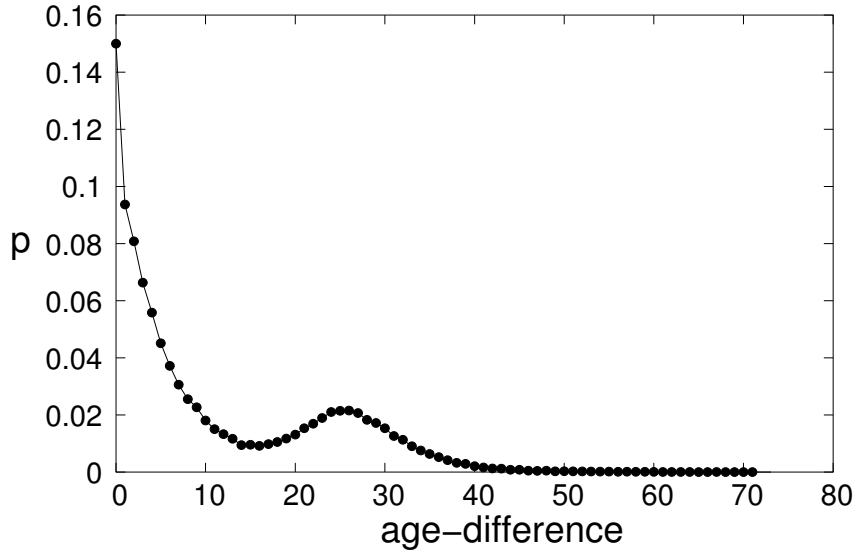
Figure 5: The probability distribution of the age difference between community members in the phone-call network. The most probable values are close to zero, indicating that a pair of members from a community are most likely to be of the same age. However, there is a smaller peak around 25 as well, therefore, if the ages of two members deviate from each other significantly, then they are most likely to be a generation apart.

of a given community at t amongst the communities at $t-1$, and its next state amongst the communities at $t+1$. The communities obtained by the CPM method can have overlaps, therefore this question is far from trivial. A simple approach would be to match communities from consecutive time steps in descending order of their relative node overlap. The relative node overlap between communities A and B can be defined as

$$C(A, B) \equiv \frac{|A \cap B|}{|A \cup B|},$$ (2)

where $|A \cap B|$ is the number of common nodes in A and B, and $|A \cup B|$ is the number of nodes in the union of the two communities. However, the nodes shared between the communities can undermine this type of community conjugation between consecutive time steps: In case a small community A is inflated by large magnitude between time steps t and $t+1$, and at $t+1$ it overlaps with another community B_{t+1}, then the relative overlap (2) between A_{t+1} and B_t can be larger than the relative overlap between A_{t+1} and A_t. In Fig.7. we show a situation of this type at $k=3$. At t we have the community A containing three nodes (blue) and the community B with five nodes (green) and no overlaps between the two communities. Then A is enlarged to $s=12$ at $t+1$, whereas B loses a node and gains a new member instead. Furthermore, four nodes in B becomes a member of A as well. Note that A_{t+1} and B_{t+1} are still separate communities as a k-clique template cannot be rolled out from B to A. The relative node overlaps read the following:

$$C(A_t, A_{t+1}) = \frac{1}{4}, \qquad C(B_t, A_{t+1}) = \frac{4}{13},$$ (3)

therefore $C(B_t, A_{t+1}) > C(A_t, A_{t+1})$. Since a community may overlap with several other communities at the same time, which them self can overlap with many more other communities as well, the matching of the communities based on solely relative node overlap becomes a complicated process in the light of the example shown above.
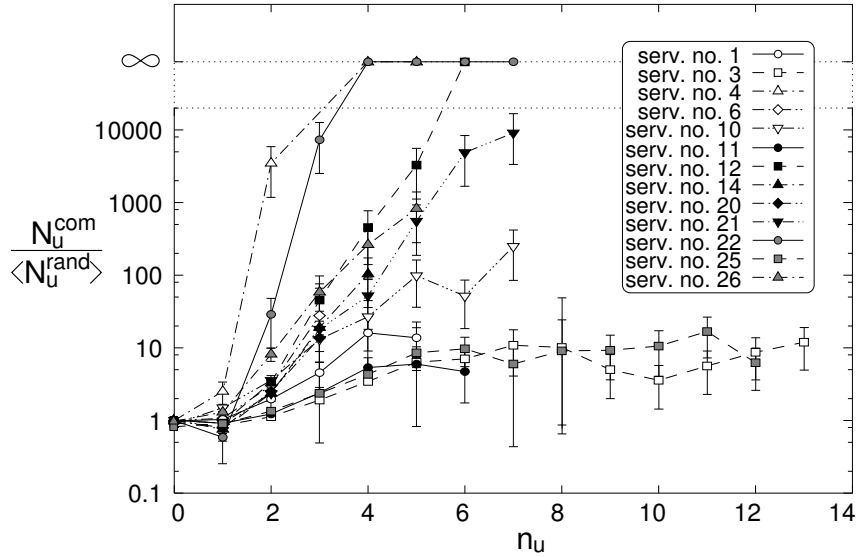
Figure 6: The number of communities divided by the average number of random sets containing the same n_u number of people using a given service. Each sample of the random sets was prepared with size distribution of the communities determined for the phone-call network. Altogether 10000 samples were prepared, the error-bars correspond to $N_u^{com}/(\langle N_u^{rand}\rangle + \sigma_u^{rand})$ and $N_u^{com}/(\langle N_u^{rand}\rangle - \sigma_u^{rand})$, where σ_u^{rand} denotes the standard deviation in case of the random sets.

To overcome this difficulty, we refine the identification of communities as shown in Fig.1f in the main text and in Fig.8. 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. When new links are introduced in a network, the CPM communities may remain unchanged, they may grow, or a group of CPM communities may become joined into a single community, however no CPM community may decay by losing members. From this it follows that if we merge two networks, any CPM community in any of the original networks will be contained in exactly one community in the joined network. Let us denote the set of communities from t by $\mathbf{D}$, the set of communities from $t+1$ by $\mathbf{E}$, and the set of communities from the joint network by $\mathbf{V}$. For any community $D_i \in \mathbf{D}$ or $E_j \in \mathbf{E}$ we can find exactly one community $V_k \in \mathbf{V}$ containing it. A very important point is that when checking whether D_i or E_j is contained or not in V_k, we compare the *links* in the corresponding communities. The CPM permits a community to contain even all nodes of another community, (*e.g.* if
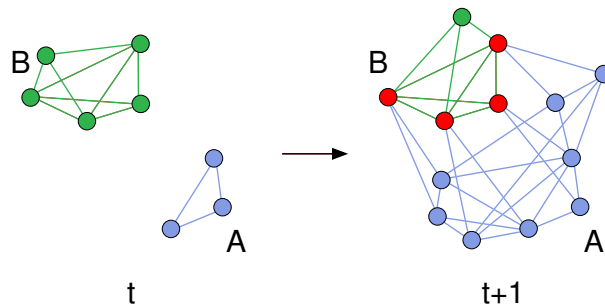

Figure 7: A situation where simple relative node overlap between the communities is not enough to decide the matching between t and $t+1$. (The k-clique size is $k = 3$.)

in Fig.7. the community B would not gain a new member, then all nodes in B_{t+1} would be contained in A_{t+1} as well), therefore distinction between communities has to be made based on links rather then members. (In the studied systems an example where a smaller community can be formed "on top" of a larger community in this way could be a small group of friends working for the same large company, or a small group of scientists collaborating in an additional field of interest as well beside the main research area of their larger community).

Turning back to our matching algorithm: When matching the communities in $\mathbf{D}$ and in $\mathbf{E}$, first for every community $V_k \in \mathbf{V}$ in the joint system we extract the list of communities $D_i^k \in \mathbf{D}$ and $E_j^k \in \mathbf{E}$ that are contained in V_k (this means $D_i^k \subseteq V_k$ and $E_j^k \subseteq V_k$). (Note that either of the lists may be empty). Then the relative overlap between every possible (D_i^k, E_j^k) pairs can be obtained as

$$C_{ij}^k = \frac{\left| D_i^k \cap E_j^k \right|}{\left| D_i^k \cup E_j^k \right|}, \tag{4}$$

and we match the pairs of communities in descending order of their relative overlap. As an illustration of the above process, in Fig.8 we show three simple scenarios occurring in the community evolution of the phone-call network. In Fig.8a both lists D_i^k and E_j^k consist of only a single community, therefore these can be matched right away. However, in Fig.8b the D_i^k list contains two elements, let us denote the light blue community of size $s = 6$ by D_1^k and the dark blue community consisting of nine nodes by D_2^k. The corresponding E_j^k list contains a single community E_1^k having 15 members. The relative overlaps between the communities are given as $C_{1,1}^k = 2/5$ and $C_{2,1}^k = 3/5$. Since the $C_{2,1}^k$ relative overlap of the yellow E_1^k community with the dark blue D_2^k community is larger than the $C_{1,1}^k$ relative overlap with the light blue D_1^k, we assign E_1^k to D_2^k. As a consequence the light blue D_1^k community comes to the end of its life at t, and it is swallowed by D_2^k. The opposite process is shown in Fig.8c: in this case the D_i^k list consists of a single community D_1^k of size $s = 15$, whereas the E_j^k list has two elements, the yellow community labelled E_1^k with six members, and the orange community labelled E_2^k containing ten nodes. The relative overlaps are $C_{1,1}^k = 2/5$ and $C_{1,2}^k = 2/3$, therefore the D_1^k is matched to E_2^k, and E_1^k is treated as a new born community. In general, whenever the community V_k contains more communities from $\mathbf{D}$ than from $\mathbf{E}$, the communities D_i^k left with no counterpart from E_j^k finish their life's at t, and when V_k contains more communities from $\mathbf{E}$ than from $\mathbf{D}$, the communities E_j^k left with no counterpart from D_i^k are considered as new born communities.

In some cases we can observe that although a community was disintegrated, after a few steps it suddenly reappears in the network. Our conjecture is that this is more likely to be the consequence of a temporally lower publishing-rate/calling-rate of the people in question than of the real disassembly and re-assembly of the corresponding social community between the people. Therefore, whenever a newborn community includes a formerly disintegrated one, then the last state of the old community is elongated to fill the gap before the reappearance, and the newborn community is treated as the continuation of the old one, as shown in Fig.9.

7 Merging of communities

During the time evolution, a pair (or a larger group) of initially distinct communities can join together to form a single community. A very interesting question connected to this is that can we find a simple relation between the size of a community and the likelihood that if will take part in such process? To investigate this issue we carried out measurements similar to those in Ref.[5]. The basic idea is that if the merging process is uniform with respect to the size of the communities s, then communities with
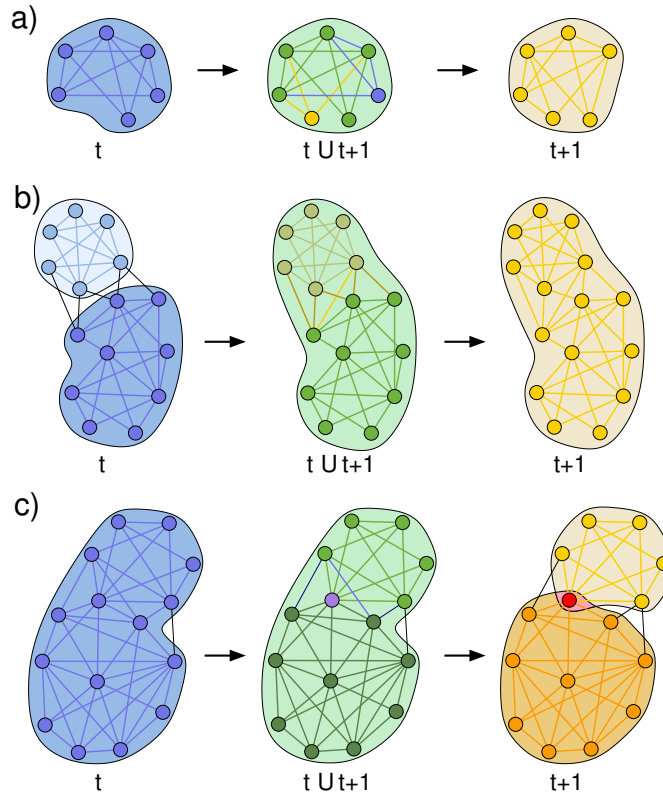
Figure 8: Simple scenarios in the community evolution of the phone-call network for $k = 4$. The communities at t are coloured blue, the communities at $t + 1$ are coloured yellow, and the communities in the joint network are coloured green. a) a community simply 'propagates', b) the dark blue community swallows the light blue, c) the yellow community is detached from the orange one.

a given s are chosen at a rate given by the size distribution of the available communities. However, if the merging mechanism prefers large (or small) sizes, then communities with large (or small) s are chosen with a higher rate compared to the size distribution of the available communities. To monitor this enhancement, at each time step t the cumulative size-pair distribution $P_t(s_1, s_2)$ was recorded. Simultaneously, the un-normalised cumulative size-pair distribution of the communities merging between t and $t + 1$ was constructed; we shall denote this distribution by $w_{t \to t+1}(s_1, s_2)$. The value of this rate-like variable $w_{t \to t+1}(s_1^*, s_2^*)$ at a given value of s_1^* and s_2^* is equal to the number of pairs of communities that merged between t and $t + 1$ and had sizes $s_1 > s_1^*$ and $s_2 > s_2^*$. To detect deviations from uniform merging probabilities, the ratio of $w_{t \to t+1}(s_1, s_2)$ and $P_t(s_1, s_2)$ was accumulated during the time evolution resulting in

$$W(s_1, s_2) \equiv \sum_{t=0}^{t_{\max}-1} \frac{w_{t \to t+1}(s_1, s_2)}{P_t(s_1, s_2)}. \tag{5}$$

When the merging process is uniform with respect to the community size the $W(s_1, s_2)$ becomes a flat function: on average we see pairs of communities merging with sizes s_1 and s_2 at a rate equal to the probability of finding a pair of communities of these sizes. However, if the merging process prefers large (or small) communities, than pairs with large (or small) sizes merge at a higher rate than the probability of finding such pairs, and $W(s_1, s_2)$ becomes increasing (or decreasing) with the size.

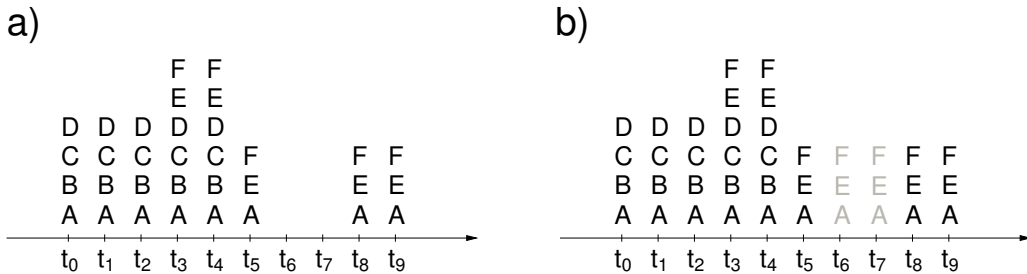The reason for using un-normalised $w_{t \to t+1}(s_1, s_2)$ distributions is that in this way each merging

a)

```
              F F
              E E
D D D D D
C C C C C F       F F
B B B B B E       E E
A A A A A A       A A
────────────────────────────▶
t₀ t₁ t₂ t₃ t₄ t₅ t₆ t₇ t₈ t₉
```

b)

```
              F F
              E E
D D D D D
C C C C C F F F F F
B B B B B E E E E E
A A A A A A A A A A
────────────────────────────▶
t₀ t₁ t₂ t₃ t₄ t₅ t₆ t₇ t₈ t₉
```

Figure 9: a) A community is disintegrated after step t_5, and it is reborn at step t_8. b) We treat the community as if it was alive at steps t_6 and t_7 too, with the same nodes as at step t_5.

event contributes to $W(s_1, s_2)$ with equal weight, and the time steps with a lot of merging events count more than those with only a few events. In the opposite case, (when $w_{t \to t+1}(s_1, s_2)$ is normalised for each pairs of subsequent time steps $t, t + 1$), the merging events occurring between time steps with a lot of other merging events are suppressed compared to the events with only a few other parallel events, as each pairs of consecutive time steps $t, t + 1$ contribute to the $W(s_1, s_2)$ function with equal weights. This difference between normalised and un-normalised $w_{t \to t+1}(s_1, s_2)$ becomes important in case of the co-authorship network, where in the beginning the system is small and merging is rare, and later on as the system is developing, merging between communities becomes a regular event.

We note that the ratio of the un-normalised size-pair distribution of the communities merging between t and $t + 1$ (denoted by $\hat{w}_{t \to t+1}(s_1, s_2)$) and the size-pair distribution of the communities $p(s_1, s_2)$ could be used to detect deviation from uniform merging probability in the same way as the cumulative distributions. In this approach the value of the rate-like variable $\hat{w}_{t \to t+1}(s_1^*, s_2^*)$ at a given value of s_1^* and s_2^* is equal to the number of pairs of communities that merged between t and $t + 1$ and had sizes $s_1 = s_1^*$ and $s_2 = s_2^*$. The emerging

$$\hat{W}(s_1, s_2) \equiv \sum_{t=0}^{t_{\max}-1} \frac{\hat{w}_{t \to t+1}(s_1, s_2)}{p_t(s_1, s_2)} \tag{6}$$

has the same properties as $W(s_1, s_2)$ regarding the attachment probabilities: it is increasing when large sizes are preferred and decreasing when small sizes are preferred. However, in our studies we choose to use the cumulative distributions because the resulting $W(s_1, s_2)$ is much smoother than $\hat{W}(s_1, s_2)$ obtained from (6).

In Fig.10. we show $W(s_1, s_2)$ for both networks, and the picture suggests that large sizes are preferred in the merging process. This is consistent with our findings that the content of large communities is changing at a faster rate compared to the small ones. Swallowing other communities is an efficient way to bring numerous new members into the community in just one step, therefore taking part in merging is beneficial for large communities following a survival strategy based on constantly changing their members. Another interesting aspect of the results shown in Fig.10. is that they are analogous to the attachment mechanism of links between already existing nodes in collaboration networks [6]: the probability for a new link to appear between two nodes with degree d_1 and d_2 is roughly proportional to $d_1 \times s_2$. Similarly, the probability that two communities of sizes s_1 and s_2 will merge is proportional to to $s_1 \times s_2$, therefore the large communities attract each other in a similar manner to hubs in collaboration networks.
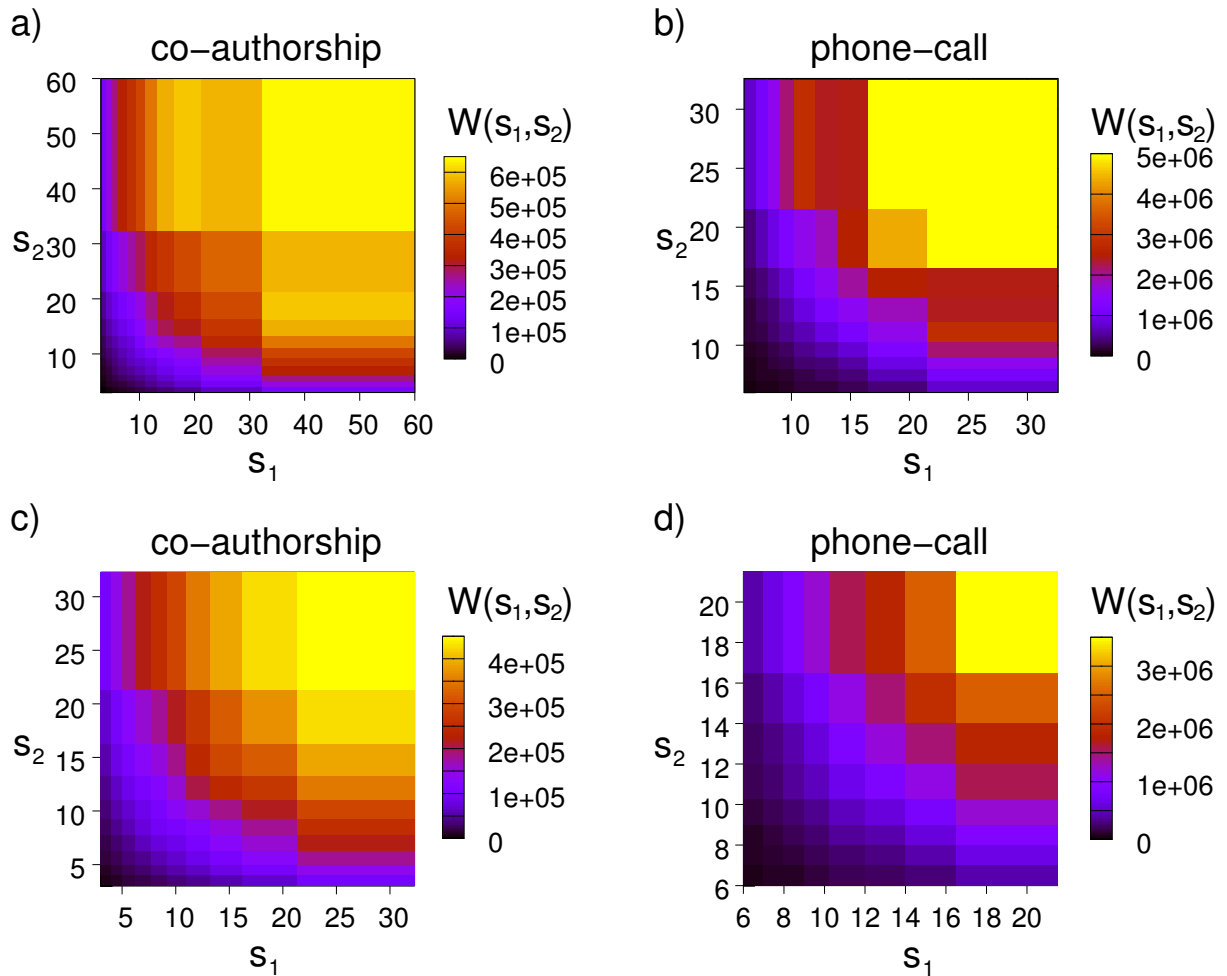
11

Figure 10: The merging of communities. a) the $W(s_1, s_2)$ function for the co-authorship network, b) the $W(s_1, s_2)$ function for the phone-call network, c) the region with smaller $W(s_1, s_2)$ in (a) enlarged, d) the region with smaller $W(s_1, s_2)$ in (b) enlarged.

References

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

[2] Ramasco, J. J., & Morris, S. A. Social inertia in collaboration networks. *Phys. Rev. E* **73**, 016122 (2006).

[3] 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).

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

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

[6] 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).