nature

Vol 435|9 June 2005|doi:10.1038/nature03607

LETTERS

Uncovering the overlapping community structure of complex networks in nature and society

Gergely Palla[1,2], Imre Derényi[2], Illés Farkas[1] & Tamás Vicsek[1,2]

Many complex systems in nature and society can be described in terms of networks capturing the intricate web of connections among the units they are made of[1–4]. A key question is how to interpret the global organization of such networks as the coexistence of their structural subunits (communities) associated with more highly interconnected parts. Identifying these *a priori* unknown building blocks (such as functionally related proteins[5,6], industrial sectors[7] and groups of people[8,9]) is crucial to the understanding of the structural and functional properties of networks. The existing deterministic methods used for large networks find separated communities, whereas most of the actual networks are made of highly overlapping cohesive groups of nodes. Here we introduce an approach to analysing the main statistical features of the interwoven sets of overlapping communities that makes a step towards uncovering the modular structure of complex systems. After defining a set of new characteristic quantities for the statistics of communities, we apply an efficient technique for exploring overlapping communities on a large scale. We find that overlaps are significant, and the distributions we introduce reveal universal features of networks. Our studies of collaboration, word-association and protein interaction graphs show that the web of communities has non-trivial correlations and specific scaling properties.

Most real networks typically contain parts in which the nodes (units) are more highly connected to each other than to the rest of the network. The sets of such nodes are usually called clusters, communities, cohesive groups or modules[8,10,11–13]; they have no widely accepted, unique definition. In spite of this ambiguity, the presence of communities in networks is a signature of the hierarchical nature of complex systems[5,14]. The existing methods for finding communities in large networks are useful if the community structure is such that it can be interpreted in terms of separated sets of communities (see Fig. 1b and refs 10, 15, 16–18). However, most real networks are characterized by well-defined statistics of overlapping and nested communities. This can be illustrated by the numerous communities that each of us belongs to, including those related to our scientific activities or personal life (school, hobby, family) and so on, as shown in Fig. 1a. Furthermore, members of our communities have their own communities, resulting in an extremely complicated web of the communities themselves. This has long been understood by sociologists[19] but has never been studied systematically for large networks. Another, biological, example is that a large fraction of proteins belong to several protein complexes simultaneously[20].

In general, each node i of a network can be characterized by a membership number m_i, which is the number of communities that the node belongs to. In turn, any two communities α and β can share $s^{\rm ov}_{\alpha,\beta}$ nodes, which we define as the overlap size between these communities. Naturally, the communities also constitute a network, with the overlaps being their links. The number of such links of community α can be called its community degree, $d^{\rm com}_{\alpha}$. Finally, the size $s^{\rm com}_{\alpha}$ of any community α can most naturally be defined as the number of its nodes. To characterize the community structure of a large network we introduce the distributions of these four basic quantities. In particular we focus on their cumulative distribution

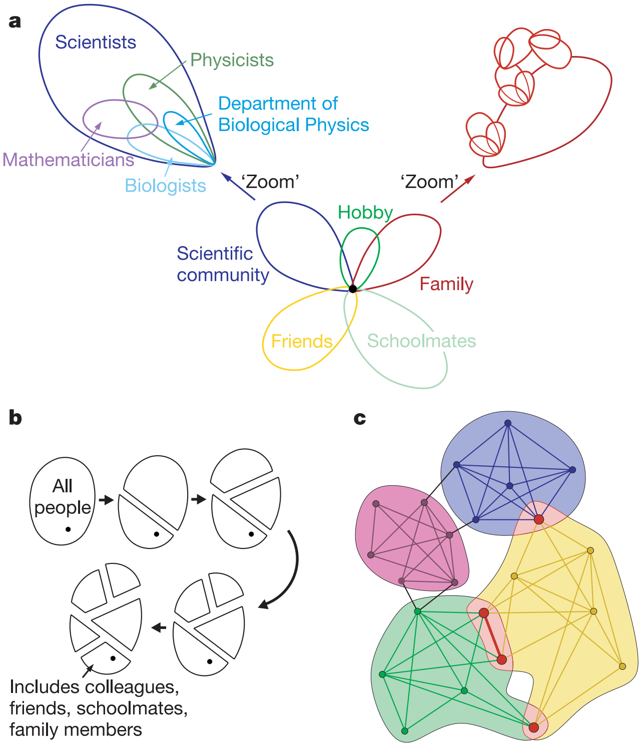

Figure 1 | Illustration of the concept of overlapping communities. a, The black dot in the middle represents either of the authors of this paper, with several of his communities around. Zooming in on the scientific community demonstrates the nested and overlapping structure of the communities, and depicting the cascades of communities starting from some members exemplifies the interwoven structure of the network of communities. **b**, Divisive and agglomerative methods grossly fail to identify the communities when overlaps are significant. **c**, An example of overlapping k-clique communities at $k = 4$. The yellow community overlaps the blue one in a single node, whereas it shares two nodes and a link with the green one. These overlapping regions are emphasized in red. Notice that any k-clique (complete subgraph of size k) can be reached only from the k-cliques of the same community through a series of adjacent k-cliques. Two k-cliques are adjacent if they share $k - 1$ nodes.

[1]Biological Physics Research Group of the Hungarian Academy of Sciences, Pázmány P. stny. 1A, H-1117 Budapest, Hungary. [2]Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary.

functions denoted by $P(m)$, $P(s^{\mathrm{ov}})$, $P(d^{\mathrm{com}})$ and $P(s^{\mathrm{com}})$. For the overlap size, for example, $P(s^{\mathrm{ov}})$ means the proportion of those overlaps that are larger than s^{ov}. Further relevant statistical features will be introduced later.

The basic observation on which our community definition relies is that a typical community consists of several complete (fully connected) subgraphs that tend to share many of their nodes. Thus, we define a community, or more precisely a k-clique 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)[21–23]. This definition seeks to represent the fact that it is an essential feature of a community that its members can be reached through well-connected subsets of nodes. There are other parts of the whole network that are not reachable from a particular k-clique, but they potentially contain further k-clique communities. In turn, a single node can belong to several communities. All these can be explored systematically and can result in many overlapping communities (illustrated in Fig. 1c). In most cases, relaxing this definition (for example, by allowing incomplete k-cliques) is practically equivalent to decreasing k. For finding meaningful communities, the way in which they are identified is expected to satisfy several basic requirements: it cannot be too restrictive, it should be based on the density of links, it is required to be local, it should not yield any cut-node or cut-link (whose removal would disjoin the community) and, of course, it should allow overlaps. We employ the community definition specified above, because none of the others in the literature satisfy all these requirements simultaneously[21,24].

Although the numerical determination of the full set of k-clique communities is a polynomial problem, we use an algorithm (which can be downloaded from http://angel.elte.hu/clustering/) that is exponential, because it is significantly more efficient for the graphs corresponding to real data. This method is based on first locating all cliques (maximal complete subgraphs) of the network and then identifying the communities by carrying out a standard component analysis of the clique–clique overlap matrix[21]. More details about the method and its speed are given in Supplementary Information.

We use our method for binary networks (that is, with undirected and unweighted links). An arbitrary network can always be transformed into a binary one by ignoring any directionality in the links and keeping only those that are stronger than a threshold weight w^*. Changing the threshold is like changing the resolution (as in a microscope) with which the community structure is investigated: by increasing w^* the communities start to shrink and fall apart. A similar effect can be observed by changing the value of k as well: increasing k makes the communities smaller and more disintegrated but also at the same time more cohesive.

When we are interested in the community structure around a particular node, it is advisable to scan through some ranges of k and w^* and monitor how its communities change. As an illustration, in Fig. 2 we show diagrams of the communities of three selected nodes of three large networks: the social network of scientific collaborators[25] (Fig. 2a), the network of word associations[26] related to cognitive sciences (Fig. 2b) and the molecular-biological network of protein–protein interactions[27] (Fig. 2c). These pictures can serve as tests or validations of the efficiency of our algorithm. In particular,

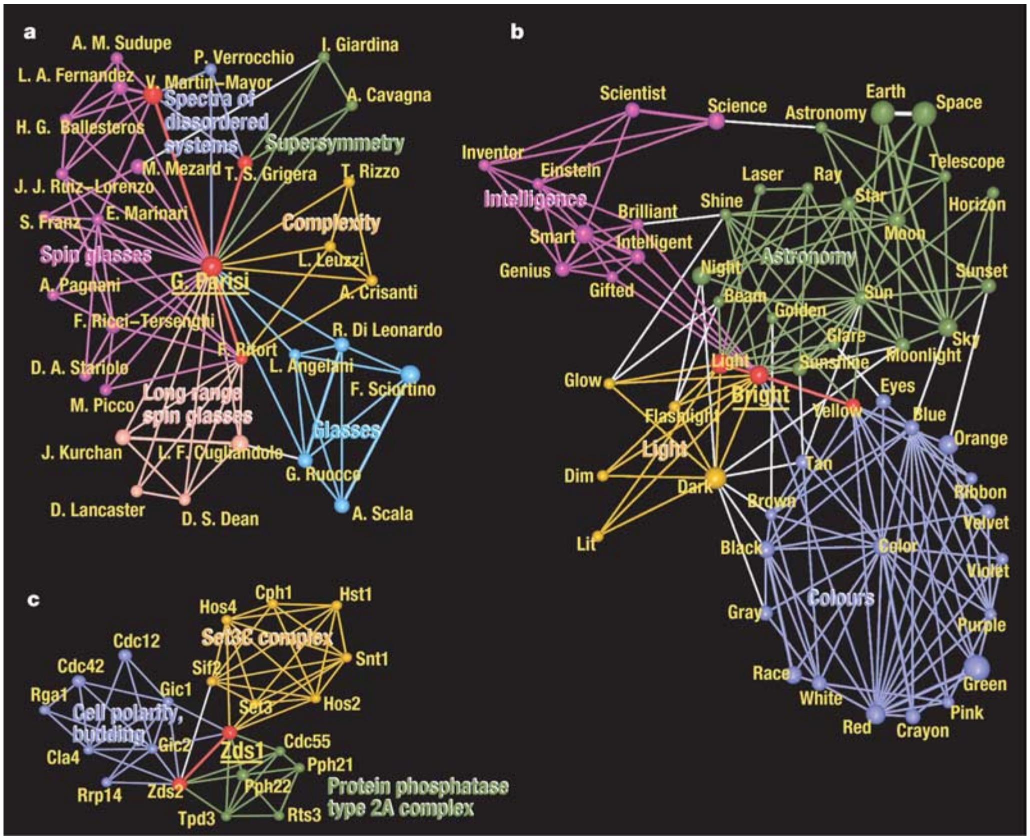

Figure 2 | **The community structure around a particular node in three different networks.** The communities are colour coded, the overlapping nodes and links between them are emphasized in red, and the volume of the balls and the width of the links are proportional to the total number of communities they belong to. For each network the value of k has been set to 4. **a**, The communities of G. Parisi in the co-authorship network of the Los Alamos Condensed Matter archive (for threshold weight $w^* = 0.75$) can be associated with his fields of interest. **b**, The communities of the word 'bright' in the South Florida Free Association norms list (for $w^* = 0.025$) represent the different meanings of this word. **c**, The communities of the protein Zds1 in the DIP core list of the protein–protein interactions of *S. cerevisiae* can be associated with either protein complexes or certain functions.

the communities of G. Parisi (whose contributions in different fields of physics are well known) shown in Fig. 2a are associated with his fields of interest, as can be deduced from the titles of the papers involved. The four-clique communities of the word 'bright' (Fig. 2b) correspond to the various meanings of this word. An important biological application is finding the communities of proteins, based on their interactions. Indeed, most proteins in the communities shown in Figs 2c and 3 can be associated with either protein complexes or certain functions, as can be looked up by using the GO-TermFinder package[28] and the online tools of the *Saccharomyces* Genome Database (SGD)[29]. For some proteins no function is yet available. Thus, the fact that they show up in our approach as members of communities can be interpreted as a prediction of their functions. One such example can be seen in the enlarged portion of Fig. 3. For the protein Ycr072c, which is required for the viability of the cell and appears in the dark green community on the right, SGD provides no biological process (function). By far the most significant GO term for the biological process of this community is 'ribosome biogenesis/assembly'. We can therefore infer that Ycr072c is likely to be involved in this process. In addition, new cellular processes can be predicted if as yet unknown communities are found with our method.

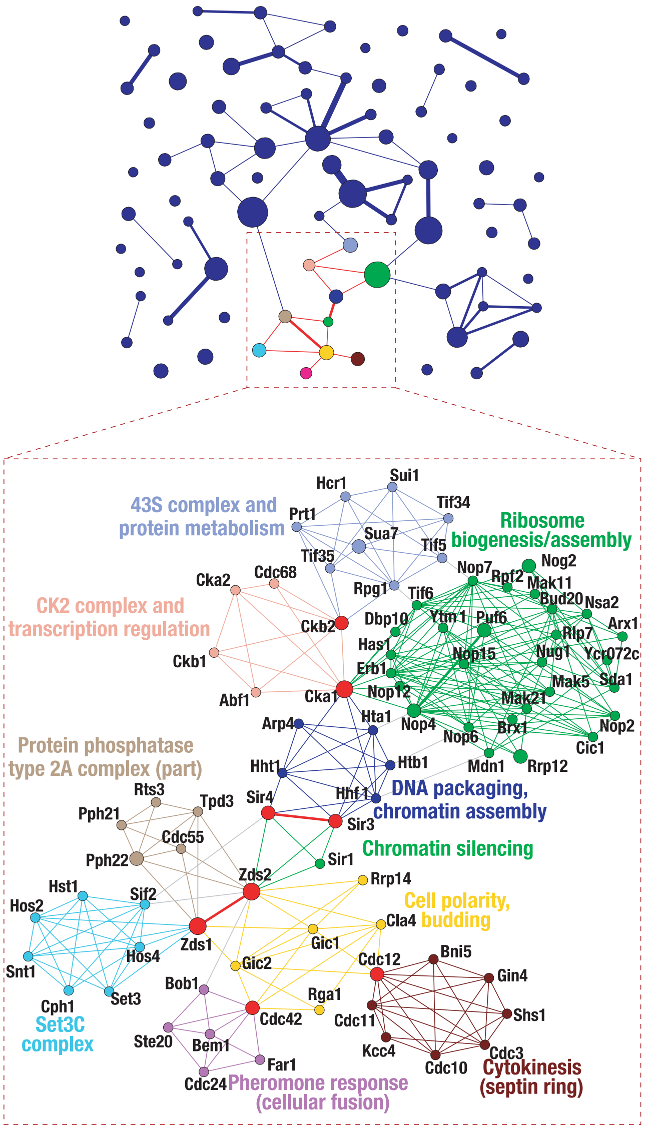

Figure 3 | Network of the 82 communities in the DIP core list of the protein–protein interactions of *S. cerevisiae* for $k = 4$. The areas of the circles and the widths of the links are proportional to the size of the corresponding communities ($s_{\alpha}^{\mathrm{com}}$) and to the size of the overlaps ($s_{\alpha,\beta}^{\mathrm{ov}}$), respectively. The coloured communities (top) are cut out and magnified to reveal their internal structure (bottom): the nodes and links of the original network have the same colour as their communities, those that are shared by more than one community are emphasized in red, and the grey links are not part of these communities. The areas of the circles and the widths of the links are proportional to the total number of communities they belong to.

These examples (and further examples included in Supplementary Information) show the advantages of our approach over the existing divisive and agglomerative methods recently used for large real networks. Divisive methods cut the network into smaller and smaller pieces, and each node is forced to remain in only one community and be separated from its other communities, most of which then necessarily fall apart and disappear. This happens, for example, with the word 'bright' when we apply the method described in ref. 16: it tends to stay together mostly with the words of the community related to 'light', while most of its other communities (for example, those related to 'colours'; see Fig. 2b) completely disintegrate ('green' becomes associated with the vegetables, 'orange' with the fruits, and so on). Agglomerative methods do the same, but in the reverse direction. For example, when we applied the agglomerative method of ref. 18, at some point 'bright', as a single word, joined a 'community' of 890 other words. In addition, such methods inevitably lead to a tree-like hierarchical rendering of the communities, whereas our approach allows the construction of an unconstrained network of communities.

The networks chosen above have been constructed in the following ways. In the co-authorship network of the Los Alamos e-print archives[25] each article contributes a value $1/(n - 1)$ to the weight of the link between every pair of its n authors. In the South Florida Free Association norms list[26] the weight of a directed link from one word to another indicates the frequency with which the people in the survey associated the end point of the link with its starting point. For our purposes these directed links have been replaced by undirected ones with a weight equal to the sum of the weights of the corresponding two oppositely directed links. In the Database of Interacting Proteins (DIP) core list of the protein–protein interactions of *Saccharomyces cerevisiae*[27] each interaction represents an unweighted link between the interacting proteins. These networks are very large, consisting of 30,739, 10,617 and 2,609 nodes and 136,065, 63,788 and 6,355 links, respectively.

Although different values of k and w^* might be optimal for the local community structure around different nodes, we should set some global criterion to fix their values if we wish to analyse the statistical properties of the community structure of the entire network. The criterion we use is based on finding a community structure that is as highly structured as possible. In the related percolation phenomena[23] a giant component appears when the number of links is increased above some critical point. Therefore, to approach this critical point from below, for each selected value of k (typically between 3 and 6) we lower the threshold w^* until the largest community becomes twice as big as the second largest one. In this way we ensure that we find as many communities as possible, without the negative effect of having a giant community that would smear out the details of the community structure by merging many smaller communities. We denote by f^* the fraction of links stronger than w^*,

Table 1 | Statistical properties of the network of communities

Network	N^{com}	$\langle d^{\mathrm{com}} \rangle$	$\langle C^{\mathrm{com}} \rangle$	$\langle r \rangle$
Co-authorship	2,450	12.10	0.44	0.58
Word association	670	11.33	0.56	0.72
Protein interaction	82	1.54	0.17	0.26

N^{com} is the number of communities, $\langle d^{\mathrm{com}} \rangle$ is the average community degree, $\langle C^{\mathrm{com}} \rangle$ is the average clustering coefficient of the network of communities, and $\langle r \rangle$ is the average fraction of shared nodes in the communities.

and use only those values of k for which f^* is not too small (not smaller than 0.5). This has led us to $k = 6$ and $k = 5$ with $f^* = 0.93$ and 0.75, respectively, for the collaboration network, and $k = 4$ with $f^* = 0.67$ for the word-association network. For the former network both sets of parameters result in very similar communities (see Supplementary Information). Because for unweighted networks no threshold weight can be set, for these we simply select the smallest value of k for which no giant community appears. For the protein interaction network this gives $k = 4$, resulting in 82 communities. Because of this relatively low number, we can depict the entire network of protein communities as in Fig. 3.

The four distributions characterizing the global community structure of these networks are shown in Fig. 4. Although the scaling of the size of non-overlapping communities has already been shown for social networks[17,18], it is striking to observe how this aspect of large real networks is preserved even when a more complete picture (allowing overlaps) is investigated. In Fig. 4a the power-law dependence $P(s^{\mathrm{com}}) \propto (s^{\mathrm{com}})^{-\tau}$ with an exponent ranging between $\tau = 1$ and $\tau = 1.6$ is well pronounced and is valid over nearly the entire range of community sizes.

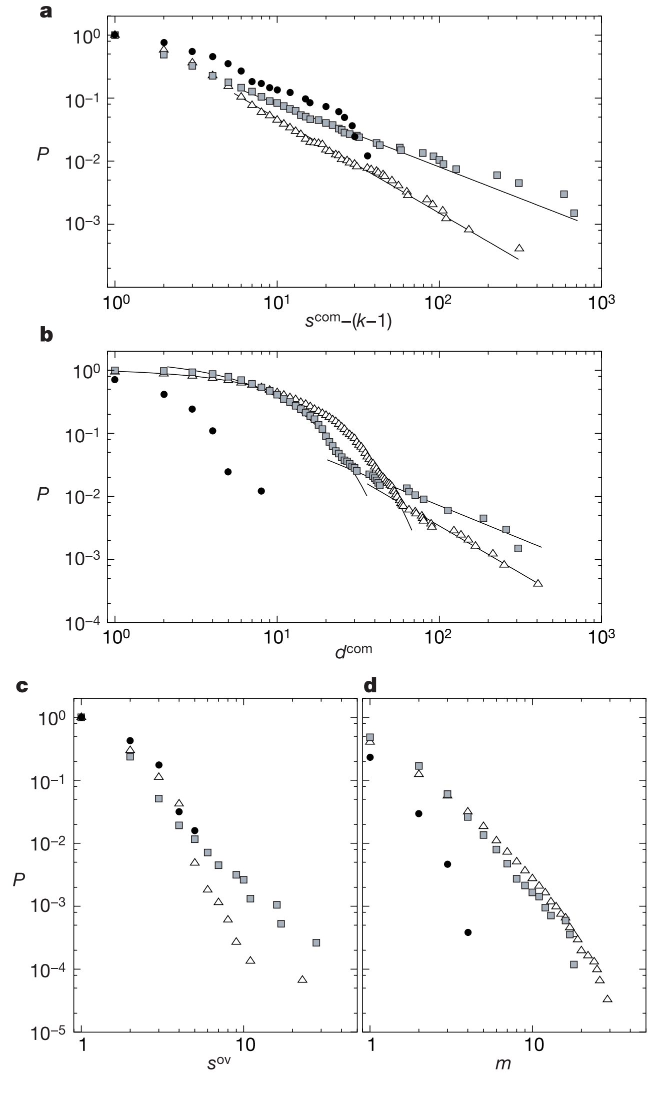

Figure 4 | Statistics of the *k*-clique communities for three large networks. The networks are the co-authorship network of the Los Alamos Condensed Matter archive (triangles, $k = 6$, $f^* = 0.93$), the word-association network of the South Florida Free Association norms (squares, $k = 4$, $f^* = 0.67$), and the protein interaction network of the yeast *S. cerevisiae* from the DIP database (circles, $k = 4$). **a**, The cumulative distribution function of the community size follows a power law with exponents between -1 (upper line) and -1.6 (lower line). **b**, The cumulative distribution of the community degree starts exponentially and then crosses over to a power law (with the same exponent as for the community size distribution). **c**, The cumulative distribution of the overlap size. **d**, The cumulative distribution of the membership number.

It is well known[2–4] that the nodes of large real networks have a power-law degree distribution. Will the same kind of distribution hold when we move to the next level of organization and consider the degrees of the communities? We find that it is not so. The community degrees (Fig. 4b) have a unique distribution, consisting of two distinct parts: an exponential decay $P(d^{\mathrm{com}}) \propto \exp(-d^{\mathrm{com}}/d_0^{\mathrm{com}})$ with a characteristic community degree d_0^{com} (which is of the order of $\langle d^{\mathrm{com}} \rangle$ shown in Table 1), followed by a power-law tail proportional to $(d^{\mathrm{com}})^{-\tau}$. This new kind of behaviour is consistent with the community size distribution if we assume that, on average, each node of a community has a contribution δ to the community degree. The tail of the community degree distribution is therefore simply proportional to that of the community size distribution. At the first part of $P(d^{\mathrm{com}})$, in contrast, a characteristic scale $d_0^{\mathrm{com}} \approx k\delta$ appears, because most of the communities have a size of the order of k (see Fig. 4a) and their distribution around d_0^{com} dominates this part of the curve. Thus, the degree to which $P(d^{\mathrm{com}})$ deviates from a simple scaling depends on k or, in other words, on the prescribed minimum cohesiveness of the communities.

The extent to which different communities overlap is also a relevant property of a network. Although the range of overlap sizes is limited, the behaviour of the cumulative overlap size distribution $P(s^{\mathrm{ov}})$, shown in Fig. 4c, is close to a power law for each network, with a rather large exponent. We can conclude that there is no characteristic overlap size in the networks. Finally, in Fig. 4d we display the cumulative distribution of the membership number $P(m)$. These plots demonstrate that a node can belong to several communities. In the collaboration and word-association networks there seems to be no characteristic value for the membership number: the data are close to a power-law dependence, with a large exponent. However, in the protein interaction network the largest membership number is only 4, which is consistent with the also rather short distribution of its community degree. To show that the communities we find are not due to an artefact of our method, we have also determined the above distributions for 'randomized' graphs with parameters (size, degree sequence, k and f^*) the same as in our three examples but with links stochastically redistributed between the nodes. We have found that the distributions are indeed extremely truncated, signifying a complete lack of the rich community structure determined for the original data.

In Table 1 we have collected a few statistical properties of the network of communities. It should be pointed out that the average clustering coefficients $\langle C^{\mathrm{com}} \rangle$ are relatively high, indicating that two communities overlapping with a given community are likely to overlap with each other as well, mostly because they all share the same overlapping region. The high fraction of shared nodes is yet another indication of the importance of overlaps between the communities.

The specific scaling of the community degree distribution is a hitherto undescribed signature of the hierarchical nature of the systems we study. We find that if we consider the network of communities instead of the nodes themselves, we still observe a degree distribution with a fat tail, but a characteristic scale appears, below which the distribution is exponential. This is consistent with our understanding of a complex system having different levels of organization with units specific to each level. In the present case the principle of organization (scaling) is preserved (with some specific modifications) when going to the next level, in good agreement with the recent finding of the self-similarity of many complex networks[30].

With recent technological advances, huge sets of data are accumulating at a tremendous pace in various fields of human activity (including telecommunications, the Internet and stock markets) and in many areas of life and social sciences (such as biomolecular assays, genetic maps and groups of World Wide Web users). Understanding both the universal and specific features of the networks associated with these data has become a significant task. The knowledge of the community structure enables the prediction of some essential features of the systems under investigation. For example, because with our approach it is possible to 'zoom' in on a single unit in a network and uncover its communities (and the communities connected to these, and so on), we provide a tool with which to interpret the local organization of large networks and can predict how the modular structure of the network changes if a unit is removed (for example, in a gene knockout experiment). A unique feature of our method is that we can simultaneously look at the network at a higher level of organization and locate the communities that have a key role within the web of communities. Among the many possible applications is a more sophisticated approach to the spreading of infections (for example, real or computer viruses) or information in highly modular complex systems.

Received 17 January; accepted 7 April 2005.

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. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74,** 47–97 (2002).
4. Mendes, J. F. F. & Dorogovtsev, S. N. *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ. Press, Oxford, 2003).
5. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297,** 1551–1555 (2002).
6. Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA* **100,** 12123–12128 (2003).
7. Onnela, J.-P., Chakraborti, A., Kaski, K., Kertész, J. & Kanto, A. Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E* **68,** 056110 (2003).
8. Scott, J. *Social Network Analysis: A Handbook* 2nd edn (Sage, London, 2000).
9. Watts, D. J., Dodds, P. S. & Newman, M. E. J. Identity and search in social networks. *Science* **296,** 1302–1305 (2002).
10. Shiffrin, R. M. & Börner, K. Mapping knowledge domains. *Proc. Natl Acad. Sci. USA* **101,** 5183–5185 (2004).
11. Everitt, B. S. *Cluster Analysis* 3rd edn (Edward Arnold, London, 1993).
12. Knudsen, S. *A Guide to Analysis of DNA Microarray Data* 2nd edn (Wiley-Liss, New York, 2004).
13. Newman, M. E. J. Detecting community structure in networks. *Eur. Phys. J. B* **38,** 321–330 (2004).
14. Vicsek, T. The bigger picture. *Nature* **418,** 131 (2002).
15. Blatt, M., Wiseman, S. & Domany, E. Super-paramagnetic clustering of data. *Phys. Rev. Lett.* **76,** 3251–3254 (1996).
16. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99,** 7821–7826 (2002).
17. 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).
18. Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69,** 066133 (2004).
19. Faust, K. in *Models and Methods in Social Network Analysis* (eds Carrington, P., Scott, J. & Wasserman, S.) 117–147 (Cambridge Univ. Press, New York, 2005).
20. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
21. Everett, M. G. & Borgatti, S. P. Analyzing clique overlap. *Connections* **21,** 49–61 (1998).
22. Batagelj, V. & Zaversnik, M. Short cycles connectivity. *arXiv* cs.DS/0308011 ⟨http://arxiv.org/abs/cs/0308011⟩ (2003).
23. Derényi, I., Palla, G. & Vicsek, T. Clique percolation in random networks. *Phys. Rev. Lett.* **94,** 160202 (2005).
24. Kosub, S. in *Network Analysis* (eds Brandes, U. & Erlebach, T.) 112–142 (Lecture Notes in Computer Science 3418, Springer, Berlin, 2005).
25. Warner, S. E-prints and the Open Archives Initiative. *Library Hi Tech* **21,** 151–158 (2003).
26. Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. The University of South Florida word association, rhyme, and word fragment norms. ⟨http://www.usf.edu/FreeAssociation/⟩.
27. Xenarios, I. *et al.* DIP: the Database of Interacting Proteins. *Nucleic Acids Res.* **28,** 289–291 (2000).
28. Boyle, E. I. *et al.* GO:TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20,** 3710–3715 (2004).
29. Cherry, J. M. *et al.* Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387,** 67S–73S (1997).
30. Song, C., Havlin, S. & Makse, H. A. Self-similarity of complex networks. *Nature* **433,** 392–395 (2005).

Supplementary Information accompanies the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A.-L. Barabási and P. Pollner for discussions, and B. Kovács and G. Szabó for help with visualization and software support. This research was supported by the Hungarian Research Grant Foundation (OTKA).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.V. (vicsek@angel.elte.hu).