Link communities reveal multi-scale complexity in networks

Yong-Yeol Ahn[1,2], James P. Bagrow[1,2] & Sune Lehmann[3,4]*

[1]Center for Complex Network Research, Department of Physics, Northeastern University, Boston, MA 02115.

[2]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Harvard University, Boston, MA 02215.

[3]Institute for Quantitative Social Science, Harvard University, Cambridge MA, 02138.

[4]College of Computer and Information Science, Northeastern University, Boston MA, 02115.

February 15, 2010

Networks have become a key approach to understanding systems of interacting objects, unifying the study of diverse phenomena including biological organisms and human society.[1–3] One crucial step when studying the structure and dynamics of networks is to identify communities;[4,5] groups of related nodes that correspond to functional subunits such as protein complexes[6,7] or social spheres.[8–10] Communities in networks often overlap[9,10] such that nodes simultaneously belong to several groups. Meanwhile, many networks are known to possess multi-scale, hierarchical organisation, where communities are recursively grouped into a hierarchical structure.[11–13] However, the fact that many real networks have communities with pervasive overlap, where each and every node belongs to more than one group, has the consequence that a global hierarchy of nodes cannot capture the relationships between overlapping groups. Here we reinvent communities as groups of links rather than nodes and show that this unorthodox approach successfully reconciles the antagonistic organising principles of overlapping communities and hierarchy. In contrast to the existing literature, which has entirely focused on grouping nodes, link communities naturally incorporate overlap while revealing hierarchical organisation. We find relevant link communities in many networks, including major biological networks such as protein-protein interaction[6,7,14] and metabolic networks,[11,15,16] and show that a large social network[10,17,18] contains hierarchically organised community structures spanning inner-city to regional scales while maintaining pervasive overlap. Our results imply that link communities are fundamental building blocks which reveal overlap and multi-scale hierarchical organisation in networks to be two aspects of the same phenomenon.

Although no common definition has been agreed upon, it is widely accepted that a community should have more internal than external connections.[19] A popular measure of community quality, modularity, is defined by comparing the number of connections within a community with the expected number of connections within the community under randomisation of the network.[20] However, these standard definitions of community structure break down when overlap is pervasive; in many real networks, nodes typically possess multiple roles[6,7,9,10,14,16] (Fig. 1a). Pervasive overlap in real networks is distinct from 'fuzzy' community overlap with relaxed interfaces[21–23] because overlap can exist for each and every node (Fig. 1b). When overlap is pervasive, counterintuitively, each community has many more external than internal connections. Overlap creates another serious problem: a single dendrogram cannot fully encode the hierarchy, since this dendrogram assumes disjoint community structure and prohibits nodes from simultaneously belonging to multiple, overlapping groups (Fig. 1c).

Although the discovery of hierarchy and community organisation has always been considered a problem of determining the correct membership(s) of each node, notice that, while *nodes* belong to multiple groups (individuals have families, coworkers *and* friends), *links* often exist for one dominant reason (two people are in the same family, work together *or* have common interests). Thus, in contrast to nodes, link membership typically is uniquely defined, even when nodes belong to multiple, diverse communities (Fig. 1d,e). Instead of assuming that a community is a set of nodes with many links between them, we consider a community to be a set of links that are densely interrelated.

With each link defined in a single context, we can now build a hierarchical tree where each leaf is a link from the original network (see Methods and Fig. 1d-f). Hierarchical clustering, based on the similarity of pairs of links, results in a dendrogram whose branches represent link communities (Fig. 1f). In this dendrogram, links occupy unique positions and nodes naturally occupy multiple positions, due to their links. Agglomerating links leads to a dendrogram containing richer information than those of traditional methods. We extract link communities at multiple levels by cutting this dendrogram at various thresholds, which can then be translated into overlapping node communities. Each node inherits all of the memberships of its links and can thus belong to multiple, overlapping communities. Even though we assign only a single membership per link, link communities also capture multiple relationships between nodes, because multiple nodes can simultaneously belong to several communities together.

The link dendrogram provides a rich hierarchy of structure but to obtain the most relevant communities one needs to determine the best level at which to cut the tree. For this purpose, we introduce

[*] Correspondence and requests for materials should be addressed to S.L. (email: *slehmann@iq.harvard.edu*).

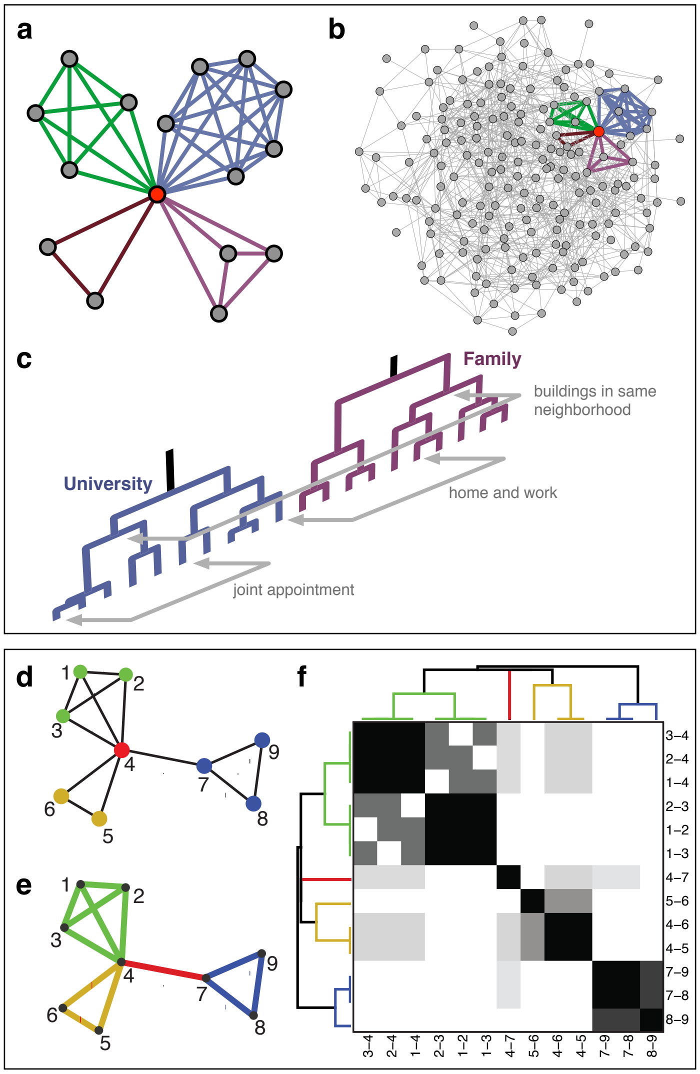

Figure 1: Overlapping communities lead to dense networks and prevent the discovery of a single node hierarchy. (**a**) Local structure in social networks is simple: an individual node sees the communities it belongs to. (**b**) Complex global structure emerges when *every* node is in the situation displayed in (a). (**c**) This pervasive overlap hinders the discovery of hierarchical organisation since nodes cannot occupy multiple leaves of a node dendrogram, preventing a single tree from encoding the full hierarchy. Bottom Panel, an example network with (**d**) node communities and (**e**) link communities; the non-overlapping link communities in (e) naturally incorporate the overlapping node in (d). (**f**) The link similarity matrix (darker matrix elements show more similar pairs of links) and resulting link dendrogram. See SI for additional examples.

a natural objective function, the *partition density D*, based on link density inside communities; unlike modularity,[20] D does not suffer from a resolution limit[24] (see Methods). Computing D at each level of the link dendrogram allows us to pick the best level to cut, though structure exists above and below that threshold (Fig. 1f, see also Fig. 4). One can also optimise D directly. Link communities allows us to formulate overlapping community discovery as a well-posed optimisation problem, embracing overlap at every node without penalising that nodes participate in multiple communities.

While link communities naturally unify hierarchy and overlap, this improvement becomes far more valuable if we can provide quantitative, real-world evidence that a link-based approach is superior to existing, node-based approaches. To do so, we compare our link community approach to a set of widely-used and successful methods on a variety of diverse network datasets using data-driven performance measures. Finally, we explore the new hierarchical, overlapping structure discovered using our link-based approach.

We analyse link communities found at the optimum partition density (see Methods) in real-world networks, compared with node communities found by three other methods: clique percolation,[9] greedy modularity optimisation,[25] and Infomap.[26] Clique percolation is the most prominent overlapping community algorithm, greedy modularity optimisation is the most popular modularity-based[20] technique, and Infomap is considered the most accurate method available.[27]

We compiled a test corpus containing 11 diverse networks (see Table 1 and SI Table S2). These networks vary from small to large; from sparse to dense; from networks with highly modular, non-overlapping structure to networks with non-modular, highly entangled structure. The test networks cover many domains of active research and represent the wide body of data currently available. We highlight a few datasets of particular scientific importance: The mobile phone network is the most comprehensive proxy of a large-scale social network in existence;[17,18] the metabolic network, from the reconstruction of *E. coli* K-12 MG1655 strain (iAF1260), is one of the most elaborate reconstructions currently available;[16] and the three protein-protein interaction (PPI) networks of *Saccharomyces cerevisiae* are the most recent and complete PPI data yet published.[14]

All the test networks possess rich *metadata* which allow us to describe the structural and functional roles of each node. For example, the biological roles of each protein in PPI network can be described by a set of controlled vocabulary (Gene Ontology (GO) terms[28]); analogous metadata exist for other networks. By calculating metadata-based similarity measures between nodes (see Methods and SI Sec. S5), we can determine the quality of communities by the similarity of the nodes they contain (*community quality*). Likewise, we can use metadata to estimate the expected amount of overlap around a node, testing the quality of the discovered overlap according to the metadata (*overlap quality*). For example, metabolites that participate in more metabolic pathways should belong to more communities than metabolites that belong to few pathways. Some methods may find high quality communities but only for a small fraction of the network; coverage measures detail how much of the network was classified by each algorithm (*community coverage*), and how much overlap was discov-

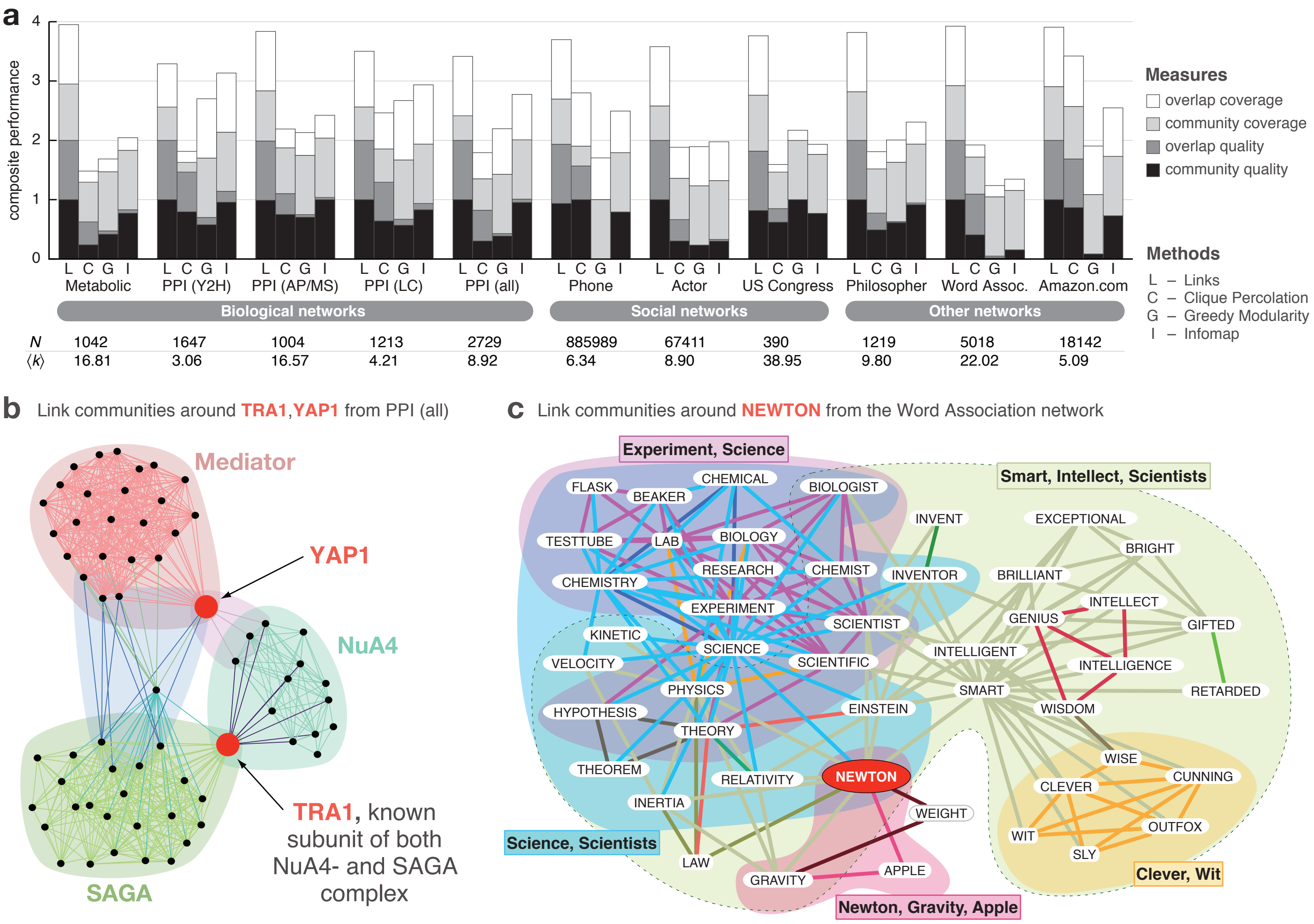

Figure 2: Assessing the relevance of link communities and other methods using real world networks. (**a**) Composite performance (see Methods and SI) is a data-driven measure of community/overlap quality (relevance of discovered memberships) and coverage (fraction of network classified). Tested algorithms over various networks are: link clustering, introduced here; clique percolation;[9] greedy modularity optimisation;[25] and Infomap.[26] These algorithms compare link clustering to a broad array of available overlapping and non-overlapping community tools, some of which are very complex. The test networks were chosen for their vastly different sizes and topologies and to represent the different domains where network analysis is used. Shown for each network is the number of nodes N and the average number of neighbours per node $\langle k \rangle$. This chart shows that link clustering finds the most relevant community structure in real world networks. (**b**) Example link communities from the full PPI network around TRA1 and YAP1. Coloured names represent corresponding protein complexes. (**c**) Example link communities from the full word association network around the word NEWTON. Link colours represent communities and filled regions provide a guide for the eye. Link communities correctly capture various concepts related to science and allow substantial overlap between communities.

Table 1: **Test networks and metadata.** See SI for full information and dataset references.

	metadata	
network	community	overlap
PPI[a]	Protein functions (GO terms)	No. of GO terms
Metabolic	Pathway annotations (KEGG)	No. of KEGG annotations
Phone	Most likely user location	Call activity
Actor	IMDb keywords for actor's films	Seniority (year of first role)
US Congress	Political ideology	Seniority (no. of terms)
Philosopher	Wikipedia hyperlinks	No. of subject categories
Word Assoc.	WordNet definitions	No. of definitions
Amazon.com	Product tags	No. of product categories

[a] There are four PPI networks. One is the union of the other three.

ered (*overlap coverage*). Each community algorithm is tested by comparing its output with the metadata, to determine how well the discovered community structure reflects the metadata, according to the four measures. Each measure is normalised such that the best method attains a value of one. *Composite performance* is the sum of these four normalised measures, such that the maximum achievable score is four. Details about the metadata and quality measures are explained in Methods and SI Sec. S5 and S6.

Figure 2a displays the results of this quantitative comparison. The quality measures show that link communities reveal more about every network's metadata than other tested methods. Not only is our approach the overall leader in every network, it is also the winner in most individual aspects of the composite performance for all networks, particularly the quality measures. The performance of link communities stands out for dense networks, such as the metabolic and word association networks, because other methods cannot interpret their pervasively overlapping structure. In addition to this quantitative evidence, we also display illustrative link communities found within the full PPI and word association networks (Figs. 2b,c; see also SI Sec. S3.4 and Supplementary Tables 1-3). Link communities in Fig. 2b capture known protein complexes and correctly identify the multiple known memberships of TRA1. A more complex community structure, around the word NEWTON in Fig. 2c, illustrates the rich structure discovered. This example also shows that multiple relationships between nodes are captured by link communities: NEWTON and GRAVITY both belong to the 'science', 'weight', and 'apple' communities. The 'clever/wit' community is correctly identified within the 'smart/intellect' community. Taken together, Fig. 2 provides strong evidence that link communities are a fruitful and valid approach to uncovering network structure.

It is particularly instructive to further examine the link communities of the metabolic and mobile phone networks, presented in Fig. 3. Here we show community coverage, the ratio of the number of links within the second largest to largest communities s_2/s_1, and partition density D, as a function of the dendrogram cut threshold (Fig. 3a). That maxima in D coincide with $s_2/s_1 \to 1/2$ indicates that discovered link communities are well structured.[9,29] Likewise, the community size distribution at the optimum D is heavy tailed for both networks (Fig. 3b). These properties show that the optimum D is related to a critical point where the link communities are neither fragmented nor gelated. The number of communities per node distinguishes the two networks (Fig. 3b insets). Mobile phone users are limited to a smaller range of community memberships, most likely due to social and time constraints. Meanwhile, the membership distribution of the metabolic network clearly displays the universality of *currency metabolites* (water, ATP, etc.) by the large number of communities they participate in. This further reinforces the relevance of link communities, since notable previous work[11,15] removed currency metabolites prior to identifying meaningful, community structure.

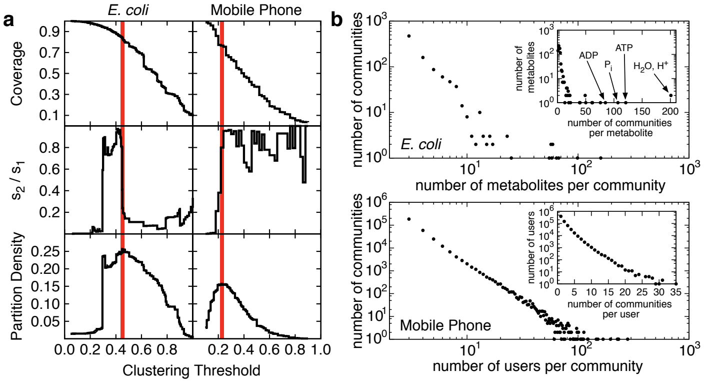

Figure 3: Statistics for the *E. coli* metabolic and mobile phone networks. (**a**) Community coverage, the ratio of the number of links in the two largest communities, and the partition density D, respectively. In both networks, peaks in D align with $s_2/s_1 \to 1/2$, implying that the maximum of D corresponds to the percolation transition point where community size exhibits a power-law distribution[29] (see also SI Figs. S16, S17). (**b**) The distribution of community sizes and node memberships (insets). The distribution of community size shows a heavy tail. The number of memberships per node is reasonable for both networks: we do not observe phone users that belong to large numbers of communities and we correctly identify *currency* metabolites, such as water and ATP, that are prevalently used throughout metabolism. The appearance of currency metabolites in many metabolic reactions is naturally incorporated into link communities, whereas their presence hindered community identification in previous work.[11,15]

Having established that link communities at the optimum partition density are meaningful and relevant, it remains to be seen how useful the link dendrogram's hierarchy can be. For this purpose, we study the mobile phone communication network in Fig. 4. Using the geographic coordinates of each user, encoded in their most likely location (Fig. 4a), we study the geography of communities throughout the hierarchy. We first build the link dendrogram and cut it at the threshold with maximum partition density (see also Fig. 3a). The three largest communities, at this optimum, are spatially correlated in the regions surrounding a major city (Fig. 4b). By partitioning the dendrogram above and below the optimum, we uncover larger, region-spanning groups and smaller, intra-city communities, respectively. Specifically, as we approach the root of the dendrogram, we see large, spatially extended communities (Fig. 4c). Near the leaves, however, we find smaller, tightly clustered groups located inside densely populated regions (Fig. 4d). In Fig. 4e, we plot the network topology of the largest community from Fig. 4c, revealing the underlying social network. Finally, Fig. 4f shows the highly overlapping structure in the largest subcommunity shown in Fig. 4e. The dendrogram for this subgroup (Fig. 4g) shows significant hierarchical structure alongside pervasive overlap. These results imply that the link dendrogram encodes useful information on multiple scales throughout its structure, not just at the optimum threshold, and provide the first evidence for the presence of a spatial, hierarchical organisation at a societal scale.

The technology to gather network datasets is constantly improving and many cutting-edge networks are far from complete. For example, an ambitious project to map all protein-protein interactions in yeast is estimated to detect approximately 20% of connections.[14] As data collection improves, networks become denser and denser, overlap becomes increasingly pervasive, and approaches specifically designed to untangle complex, highly overlapping structure become essential. More generally, the shift in perspective from nodes to links represents a fundamentally new way to study complex systems. Here we have taken the first steps towards understanding the consequences of a link-based approach, but the full potential remains unexplored. Our work has primarily focused on the highly overlapping community structure of complex networks, but the study of the hierarchy that organizes these overlapping communities has great potential for further exploration. Knowledge of a complex system's hierarchical organization has many possible applications including a more nuanced approach to analysing complex diseases as well as the dynamics of society and economic systems.

Methods

Link communities

For an undirected, unweighted network, denote the set of node i and its neighbours as $n_+(i)$. Limiting ourselves to link pairs that share a node, expected to be more similar than disconnected link pairs, the similarity S between links e_{ik} and e_{jk} is given by:

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}. \tag{1}$$

Shared node k does not appear in S because it provides no additional information and introduces bias. Single-linkage hierarchical clustering builds a link dendrogram from Eq. (1) (ties in S are agglomerated simultaneously). Cutting this dendrogram at some clustering threshold—e.g. the threshold with maximum partition density (see below)—yields link communities. See SI for details, generalisations to multipartite and weighted graphs, and the usage of other algorithms.

Partition density

For a network with M links and N nodes, $P = \{P_1, \ldots, P_C\}$ is a partition of the links into C subsets. The number of links in subset c is $m_c = |P_c|$. The number of *induced nodes*, all nodes that those links touch, is $n_c = \left|\cup_{e_{ij} \in P_c} \{i, j\}\right|$. Note that $\sum_c m_c = M$ and $\sum_c n_c \geq N$ (assuming no unconnected nodes). The link density D_c of community c is

$$D_c = \frac{m_c - (n_c - 1)}{\frac{n_c(n_c-1)}{2} - (n_c - 1)}. \tag{2}$$

This is the number of links in c, normalised by the minimum and maximum number of links possible between those nodes, assuming they remain connected. (We assume $D_c = 0$ if $n_c = 2$.) The *partition density* D is the average of D_c, weighted by the fraction of present links:

$$D = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}. \tag{3}$$

Equation 3 does not possess a resolution limit[24] since each term is local in c.

Community validation

Nontrivial communities possess 3+ nodes. We use metadata "enrichment" to assess *community quality*, comparing how similar nodes are within nontrivial communities versus all nodes (global baseline). *Overlap quality* is the mutual information between the number of nontrivial memberships and the overlap metadata (Table 1). *Community coverage* is the fraction of nodes belonging to 1+ nontrivial communities. *Overlap coverage*, since methods with equal community coverage can extract different amounts of overlap, is the average number of nontrivial memberships per node (equivalent to community coverage for non-overlapping methods). Measures are normalised so the best method has a value of 1 and their sum is shown in Fig. 2. See SI for details.

References

[1] Newman, M. E. J., Barabási, A.-L. & Watts, D. J. *The Structure and Dynamics of Networks:* (Princeton University Press, 2006), 1 edn.

[2] Caldarelli, G. *Scale-Free Networks: Complex Webs in Nature and Technology* (Oxford University Press, USA, 2007).

[3] Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. Critical phenomena in complex networks. *Reviews of Modern Physics* **80**, 1275–61 (2008).

[4] Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**, 7821–7826 (2002).

[5] Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).

[6] Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature* **440**, 637–643 (2006).

[7] Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).

[8] Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications*. Structural analysis in the social sciences (Cambridge University Press, 1994).

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

[10] Palla, G., Barabási, A. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007).

[11] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).

[12] Sales-Pardo, M., Guimera, R., Moreira, A. & Amaral, L. Extracting the hierarchical organization of complex systems. *PNAS* **104**, 15224–15229 (2007).

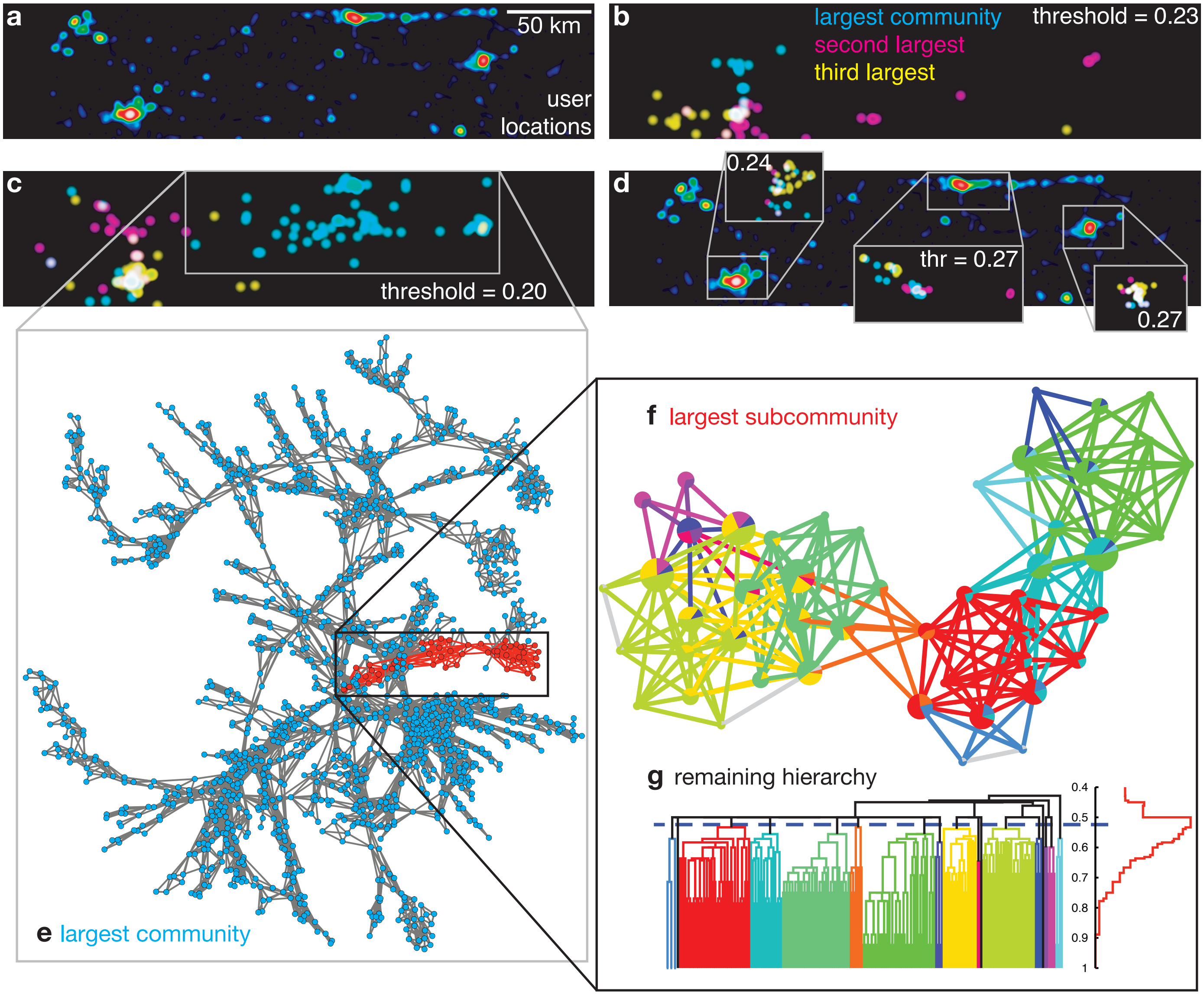

Figure 4: The social network of mobile phone users displays geographically correlated, overlapping communities at multiple scales. (**a**) A heatmap of the most likely locations of all users in the region, showing several cities. Panels (b-d) cover the same geographic region. (**b**) A map displaying only the three largest communities at the dendrogram's optimum threshold. Each community is geographically distinct. (**c**) Cutting the dendrogram at a lower threshold, the largest communities become spatially extended but still show correlation. (**d**) Meanwhile, higher thresholds yield smaller, intra-city communities. These small communities, along with the corresponding threshold value, are shown above the heatmap presented in (a). (**e**) The social network within the largest community in (c) with its largest sub-community highlighted. (**f–g**) The highlighted sub-community in (e), along with its link dendrogram and partition density as a function of threshold. The link colors in (f) correspond to the dendrogram's branch colours in (g).

[13] Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98 (2008).

[14] Yu, H. *et al.* High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science* **322**, 104–110 (2008).

[15] Guimerà, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).

[16] Feist, A. M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular Systems Biology* **3**, 1 (2007).

[17] Onnela, J.-P. *et al.* Structure and tie strengths in mobile communication networks. *PNAS* **104**, 7332 (2007).

[18] Gonzalez, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 479 (2008).

[19] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2658–2663 (2004).

[20] Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Physical Review E* **69**, 026113 (2004).

[21] Reichardt, J. & Bornholdt, S. Detecting fuzzy community structures in complex networks with a potts model. *Phys. Rev. Lett.* **93**, 218701 (2004).

[22] Li, D. *et al.* Synchronization interfaces and overlapping communities in complex networks. *Phys. Rev. Lett.* **101**, 168701 (2008).

[23] Lancichinetti, A., Fortunato, S. & Kertesz, J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* **11**, 033015 (2009).

[24] Fortunato, S. & Barthélemy, M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences* **104**, 36–41 (2007).

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

[26] Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105**, 1118–1123 (2008).

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

[28] Gene Ontology Consortium. *Nucleic Acids Res.* **36**, D440 (2008).

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

[30] Evans, T. S. & Lambiotte, R. Line graphs, link partitions and overlapping communities. *Phys. Rev. E* **80**, 016105 (2009).

Acknowledgements

While finalising this manuscript, we have been made aware that a similar approach has been developed independently by T.S. Evans and R. Lambiotte[30] (see also arXiv:0912.4389). The authors thank A.-L. Barabási, S. Ahnert, J. Park, D.-S. Lee, P.-J. Kim, N. Blumm, D. Wang, and M. A. Yildirim for invaluable discussions and ideas; and H. Yu for providing the GO pairwise similarity data for proteins. J.P.B. acknowledges support from DTRA grant BRBAA07-J-2-0035. S.L. acknowledges support by the Danish Natural Science Research Council and James S. McDonnell Foundation 21st Century Initiative in Studying Complex Systems, the National Science Foundation within the DDDAS (CNS-0540348), ITR (DMR-0426737) and IIS-0513650 programs, as well as by the U.S. Office of Naval Research Award N00014-07-C and the NAP Project sponsored by the National Office for Research and Technology (KCKHA005).

Competing Interests The authors declare that they have no competing financial interests.

Author Contributions All authors contributed equally to this work.

Correspondence Correspondence and requests for materials should be addressed to S.L.
(email: slehmann@iq.harvard.edu).