LETTERS

Link communities reveal multiscale complexity in networks

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

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 hierarchical organization, 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 organizing 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 organization. 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 organized community structures spanning inner-city to regional scales while maintaining pervasive overlap. Our results imply that link communities are fundamental building blocks that reveal overlap and hierarchical organization 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[19–24]. Counterintuitively, highly overlapping communities can have many more external than internal connections (Fig. 1a, b). Because pervasive overlap breaks even this fundamental assumption, a new approach is needed.

The discovery of hierarchy and community organization has always been considered a problem of determining the correct membership (or memberships) of each node. Notice that, whereas nodes belong to multiple groups (individuals have families, co-workers and friends; Fig. 1c), links often exist for one dominant reason (two people are in the same family, work together or have common interests). 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 closely interrelated links.

Placing each link in a single context allows us to reveal hierarchical and overlapping relationships simultaneously. We use hierarchical clustering with a similarity between links to build a dendrogram where each leaf is a link from the original network and branches represent link communities (Fig. 1d, e and Methods). In this dendrogram, links occupy unique positions whereas nodes naturally occupy multiple positions, owing to their links. We extract link communities at multiple levels by cutting this dendrogram at various thresholds. Each node inherits all memberships of its links and can thus belong to multiple, overlapping communities. Even though we assign only a single membership per link, link communities can 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 it is necessary to determine the best level at which to cut the tree. For this purpose, we introduce 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[25] (Methods). Computing D at each level of the link dendrogram allows us to pick the best level to cut (although meaningful structure exists above and below that threshold). It is also possible to optimize D directly. We can now formulate overlapping community discovery as a well-posed optimization problem, accounting for overlap at every node without penalizing that nodes participate in multiple communities.

As an illustrative example, Fig. 1f shows link communities around the word 'Newton' in a network of commonly associated English words. (See Supplementary Information, section 6, for details on networks used throughout the text.) The 'clever, wit' community is correctly identified inside the 'smart/intellect' community. The words 'Newton' and 'Gravity' both belong to the 'smart/intellect', 'weight' and 'apple' communities, illustrating that link communities capture multiple relationships between nodes. See Supplementary Information, section 3.6, for further visualizations.

Having unified hierarchy and overlap, we provide quantitative, real-world evidence that a link-based approach is superior to existing, node-based approaches. Using data-driven performance measures, we analyse link communities found at the maximum partition density in real-world networks, compared with node communities found by three widely used and successful methods: clique percolation[9], greedy modularity optimization[26] and Infomap[21]. Clique percolation is the most prominent overlapping community algorithm, greedy modularity optimization is the most popular modularity-based[20] technique and Infomap is often considered the most accurate method available[27].

We compiled a test group of 11 networks covering many domains of active research and representing the wide body of available data (Supplementary Table 2). These networks vary from small to large, from sparse to dense, and from those with modular structure to those with highly overlapping structure. We highlight a few data sets of particular scientific importance: The mobile phone network is the

[1]Center for Complex Network Research, Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA. [2]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Harvard University, Boston, Massachusetts 02215, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138, USA. [4]College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

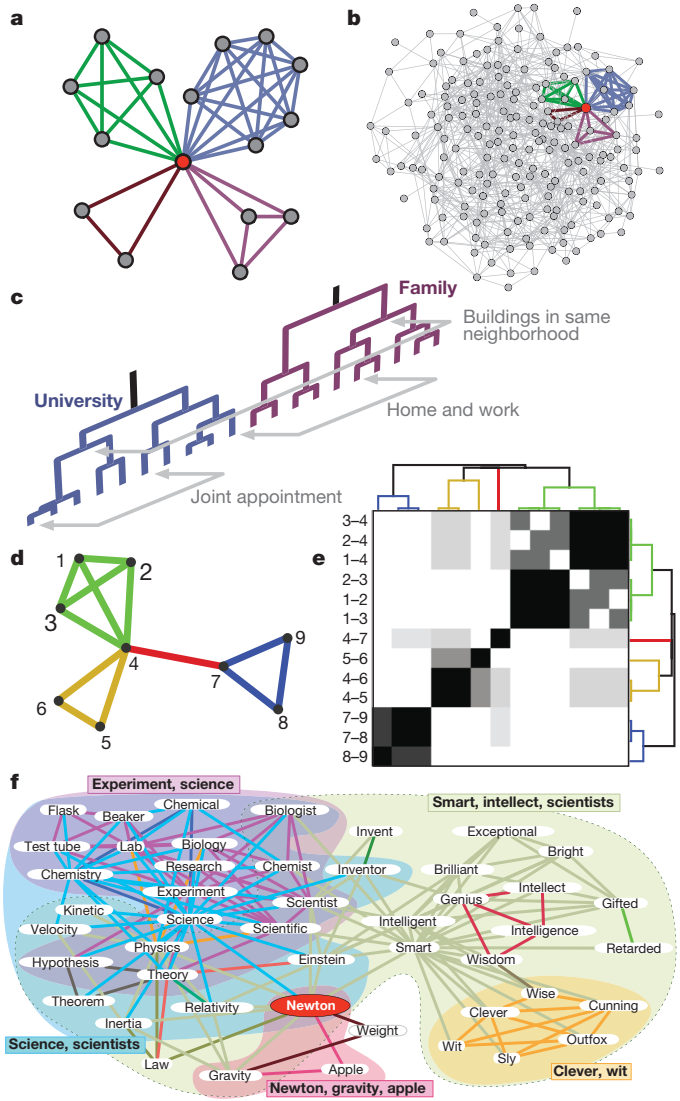

Figure 1 | Overlapping communities lead to dense networks and prevent the discovery of a single node hierarchy. **a**, Local structure in many networks is simple: an individual node sees the communities it belongs to. **b**, Complex global structure emerges when every node is in the situation displayed in **a**. **c**, Pervasive overlap hinders the discovery of hierarchical organization because nodes cannot occupy multiple leaves of a node dendrogram, preventing a single tree from encoding the full hierarchy. **d**, **e**, An example showing link communities (colours in **d**), the link similarity matrix (**e**; darker entries show more similar pairs of links) and the link dendrogram (**e**). **f**, 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 capture concepts related to science and allow substantial overlap. Note that the words were produced by experiment participants during free word associations.

most comprehensive proxy of a large-scale social network currently in existence[17,18]; the metabolic network iAF1260, from *Escherichia coli* K-12 MG1655 strain, is one of the most elaborate reconstructions currently available[16]; and the three protein–protein interaction networks of *Saccharomyces cerevisiae* are the most recent and complete protein–protein interaction data yet published[14].

These networks possess rich metadata that allow us to describe the structural and functional roles of each node. For example, the biological roles of each protein in the protein–protein interaction network can be described by a controlled vocabulary (Gene Ontology terms[28]). By calculating metadata-based similarity measures between nodes (Methods and Supplementary Information, section 5), 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 are expected to belong to more communities than metabolites that participate in fewer pathways. Some methods may find high-quality communities but only for a small fraction of the network; coverage measures describe how much of the network was classified by each algorithm ('community coverage') and how much overlap was discovered ('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 normalized such that the best method attains a value of one. 'Composite performance' is the sum of these four normalized measures, such that the maximum achievable score is four. Full details are in Methods and Supplementary Information, sections 5 and 6.

Figure 2 displays the results of this quantitative comparison, showing 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, which are expected to have pervasively overlapping structure.

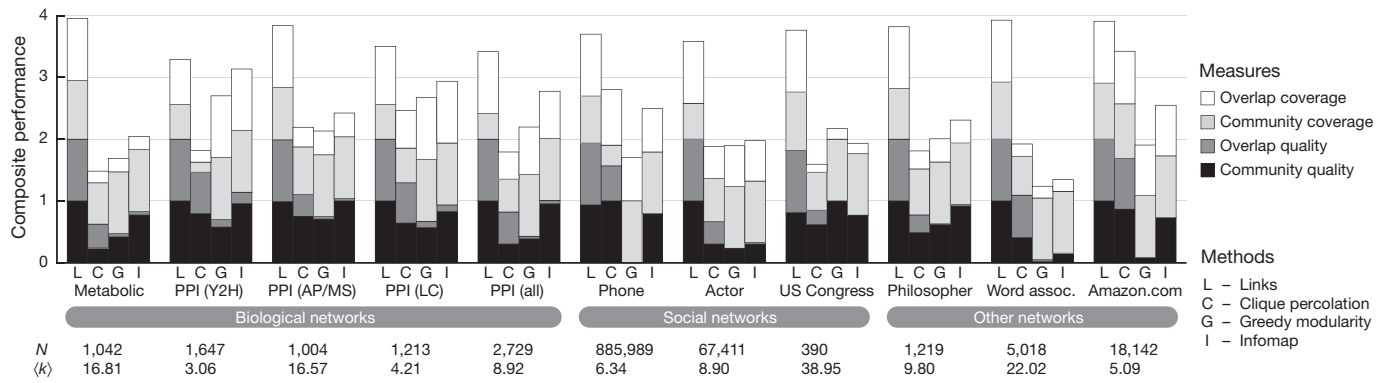

Figure 2 | Assessing the relevance of link communities using real-world networks. Composite performance (Methods and Supplementary Information) is a data-driven measure of the quality (relevance of discovered memberships) and coverage (fraction of network classified) of community and overlap. Tested algorithms are link clustering, introduced here; clique percolation[9]; greedy modularity optimization[26]; and Infomap[21]. Test networks were chosen for their varied sizes and topologies and to represent the different domains where network analysis is used. Shown for each are the number of nodes, N, and the average number of neighbours per node, $\langle k \rangle$. Link clustering finds the most relevant community structure in real-world networks. AP/MS, affinity-purification/mass spectrometry; LC, literature curated; PPI, protein–protein interaction; Y2H, yeast two-hybrid.

It is instructive to examine further the statistics of link communities in the metabolic and mobile phone networks (Fig. 3). The community size distribution at the optimum value of D is heavy tailed for both networks, whereas the number of communities per node distinguishes them (Fig. 3, insets): Mobile phone users are limited to a smaller range of community memberships, most likely as a result of social and time constraints. Meanwhile, the membership distribution of the metabolic network displays the universality of currency metabolites (water, ATP and so on) through the large number of communities they participate in. Notable previous work[11,15] removed currency metabolites before identifying meaningful community structure. The statistics presented here match current knowledge about the two systems, further confirming the communities' relevance.

Having established that link communities at the maximal partition density are meaningful and relevant, we now show that the link dendrogram reveals meaningful communities at different scales. Figure 4a–c shows that mobile phone users in a community are spatially co-located. Figure 4a maps the most likely geographic locations of all users in the network; several cities are present. In Fig. 4b, we show (insets) several communities at different cuts above the optimum threshold, revealing small, intra-city communities. Below the optimum threshold, larger, yet still spatially correlated, communities exist (Fig. 4c). Because we expect a tight-knit community to have only small geographical dispersion, the clustered structures on the map indicate that the communities are meaningful. The geographical correlation of each community does not suddenly break down, but is sustained over a wide range of thresholds. In Fig. 4d, we look more closely at the social network of the largest community in Fig. 4c, extracting the structure of its largest subcommunity along with its remaining hierarchy and revealing the small-scale structures encoded in the link dendrogram. This example provides evidence for the presence of spatial, hierarchical organization at a societal scale. To validate the hierarchical organization of communities quantitatively throughout the dendrogram, we use a randomized control dendrogram that quantifies how community quality would evolve if there were no hierarchical organization beyond a certain point. Figure 4e shows that the quality of the actual communities decays much more slowly than the control, indicating that real link dendrograms possess a large range of high quality community structures. The quantitative results of Fig. 4 are typical for the full test group, implying that rich, meaningful community structure is contained within the link dendrogram. Additional results supporting these conclusions are presented in Supplementary Information, section 7.

Many cutting-edge networks are far from complete. For example, an ambitious project to map all protein–protein interactions in yeast is currently estimated to detect approximately 20% of connections[14]. As the rate of data collection continues to increase, networks become

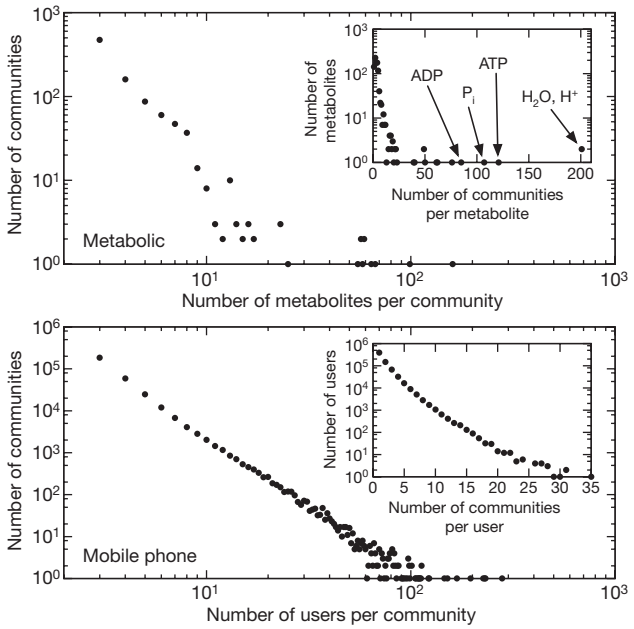

Figure 3 | Community and membership distributions for the metabolic and mobile phone networks. The distribution of community sizes and node memberships (insets). Community size shows a heavy tail. The number of memberships per node is reasonable for both networks: we do not observe phone users that belong to large numbers of communities and we correctly identify currency metabolites, such as water, ATP and inorganic phosphate (P_i), 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].

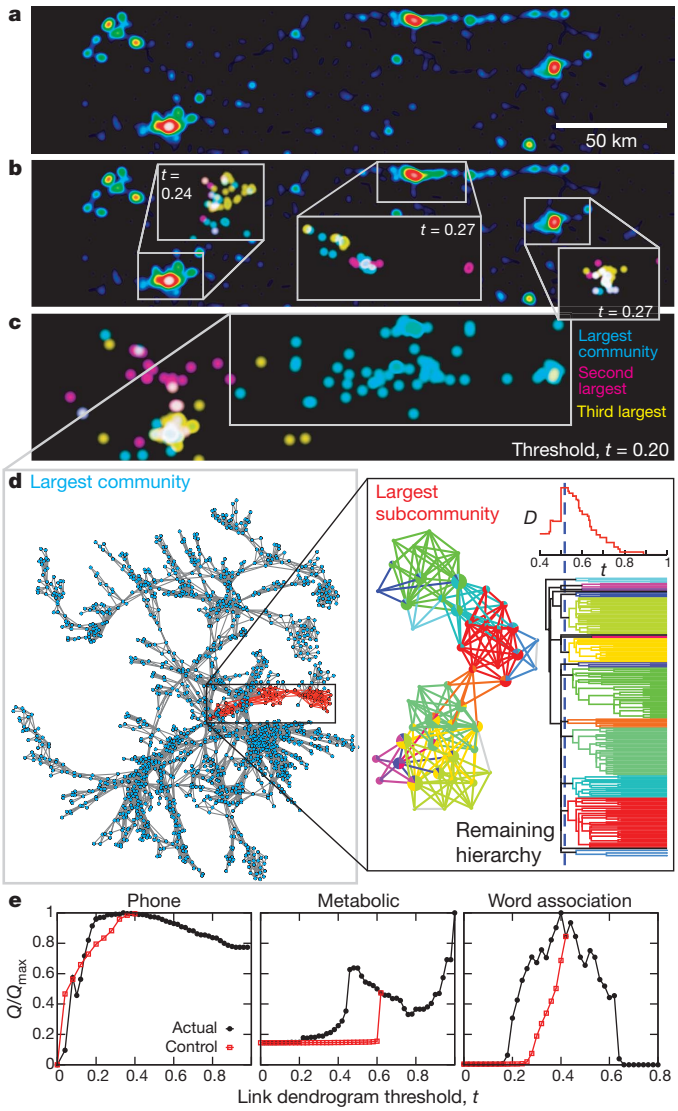

Figure 4 | Meaningful communities at multiple levels of the link dendrogram. **a**–**c**, The social network of mobile phone users displays co-located, overlapping communities on multiple scales. **a**, Heat map of the most likely locations of all users in the region, showing several cities. **b**, Cutting the dendrogram above the optimum threshold yields small, intra-city communities (insets). **c**, Below the optimum threshold, the largest communities become spatially extended but still show correlation. **d**, The social network within the largest community in **c**, with its largest subcommunity highlighted. The highlighted subcommunity is shown along with its link dendrogram and partition density, D, as a function of threshold, t. Link colours correspond to dendrogram branches. **e**, Community quality, Q, as a function of dendrogram level, compared with random control (Methods).

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 steps towards understanding the consequences of a link-based approach, but its full potential remains unexplored. Our work has primarily focused on the highly overlapping community structure of complex networks, but, as we have shown, the hierarchy that organizes these overlapping communities holds great promise for further study.

While finalizing this manuscript, we have been made aware of a similar approach developed independently by T. S. Evans and R. Lambiotte[29,30].

METHODS SUMMARY

Link communities. We denote the set of node i and its neighbours as $n_+(i)$. For link pairs that share a node, the similarity between links e_{ik} and e_{jk} is $S(e_{ik},e_{jk})=|n_+(i)\cap n_+(j)|/|n_+(i)\cup n_+(j)|$. Single-linkage hierarchical clustering then builds a link dendrogram (agglomerate ties in S simultaneously). Cutting this dendrogram at some threshold yields link communities. See Supplementary Information for details, generalizations to multipartite and weighted graphs, and other algorithms.

Partition density. For a network with M links, $\{P_1, \ldots, P_C\}$ is a partition of the links into C subsets. Subset P_c has $m_c=|P_c|$ links and $n_c=\left|\cup_{e_{ij}\in P_c}\{i,j\}\right|$ nodes. Then we define

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

This is m_c normalized by the minimum and maximum numbers of links possible between n_c connected nodes. (We assume that $D_c=0$ if $n_c=2$.) The partition density, D, is the average of D_c, weighted by the fraction of present links:

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

Equation (1) does not possess a resolution limit[25] because 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 relative to all nodes (global baseline). Overlap quality is the mutual information between the number of nontrivial memberships and the overlap metadata (Supplementary Table 2). Community coverage is the fraction of nodes belonging to 1+ nontrivial communities. Overlap coverage, because methods with equal community coverage can extract different amounts of overlap, is the average number of nontrivial memberships per node. See Supplementary Information for full details.

Control dendrogram. To study the hierarchy beyond some threshold, t_*, we begin hierarchical clustering, merging all edge pairs with $S\geq t_*$ and thus fixing the community structure at threshold $t=t_*$. Then we randomly shuffle similarities amongst the remaining edge pairs with $S<t_*$, and continue the merging process. Full details are in Supplementary Information, section 7.4.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 October 2009; accepted 13 May 2010.
Published online 20 June 2010.

1. Newman, M. E. J., Barabási, A.-L. & Watts, D. J. *The Structure and Dynamics of Networks* (Princeton Univ. Press, 2006).
2. Caldarelli, G. *Scale-Free Networks: Complex Webs in Nature and Technology* (Oxford Univ. Press, 2007).
3. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. Critical phenomena in complex networks. *Rev. Mod. Phys.* **80,** 1275–1335 (2008).
4. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99,** 7821–7826 (2002).
5. Fortunato, S. Community detection in graphs. *Phys. Rep.* **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 Univ. 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,** 814–818 (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. *Proc. Natl Acad. Sci. USA* **104,** 15224–15229 (2007).
13. Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453,** 98–101 (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. *Mol. Syst. Biol.* **3,** 121 (2007).
17. Onnela, J.-P. *et al.* Structure and tie strengths in mobile communication networks. *Proc. Natl Acad. Sci. USA* **104,** 7332–7336 (2007).
18. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453,** 779–782 (2008).
19. 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).
20. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69,** 026113 (2004).
21. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA* **105,** 1118–1123 (2008).
22. Reichardt, J. & Bornholdt, S. Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.* **93,** 218701 (2004).
23. Li, D. *et al.* Synchronization interfaces and overlapping communities in complex networks. *Phys. Rev. Lett.* **101,** 168701 (2008).
24. Lancichinetti, A., Fortunato, S. & Kertesz, J. Detecting the overlapping and hierarchical community structure in complex networks. *N. J. Phys.* **11,** 033015 (2009).
25. Fortunato, S. & Barthélemy, M. Resolution limit in community detection. *Proc. Natl Acad. Sci. USA* **104,** 36–41 (2007).
26. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70,** 066111 (2004).
27. Lancichinetti, A. & Fortunato, S. Community detection algorithms: a comparative analysis. *Phys. Rev. E* **80,** 056117 (2009).
28. The Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Res.* **36,** D440–D444 (2008).
29. Evans, T. S. & Lambiotte, R. Line graphs, link partitions and overlapping communities. *Phys. Rev. E* **80,** 016105 (2009).
30. Evans, T. S. & Lambiotte, R. Edge partitions and overlapping communities in complex networks. Preprint at ⟨http://arxiv.org/abs/0912.4389⟩ (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank A.-L. Barabási, S. Ahnert, J. Park, D.-S. Lee, P.-J. Kim, N. Blumm, D. Wang, M. A. Yildirim and H. Yu. The authors acknowledge the Center for Complex Network Research, supported by the James S. McDonnell Foundation 21st Century Initiative in Studying Complex Systems; the NSF-DDDAS (CNS-0540348), NSF-ITR (DMR-0426737) and NSF-IIS-0513650 programmes; US ONR Award N00014-07-C; the NIH (U01 A1070499-01/Sub #:111620-2); the DTRA (BRBAA07-J-2-0035); the NS-CTA sponsored by US ARL (W911NF-09-2-0053); and NKTH NAP (KCKHA005). S.L. acknowledges support from the Danish Natural Science Research Council.

Author Contributions Y.-Y.A., J.P.B. and S.L. designed and performed the research and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.L. (sune.lehmann@gmail.com).

METHODS

Link communities. For an undirected, unweighted network, we 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 pairs, we find the similarity, S, between links e_{ik} and e_{jk} to be

$$S(e_{ik},e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \quad (2)$$

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 equation (2) (ties in S are agglomerated simultaneously). Cutting this dendrogram at some clustering threshold—for example the threshold with maximum partition density (see below)—yields link communities. See Supplementary Information for details, generalizations 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 P_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

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

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

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

Equation (3) does not possess a resolution limit[25] because 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 relative to all nodes (global baseline). Overlap quality is the mutual information between the number of nontrivial memberships and the overlap metadata (Supplementary Table 2). Community coverage is the fraction of nodes belonging to 1+ nontrivial communities. Overlap coverage, because 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). See Supplementary Information for details.

Control dendrogram. To test whether the hierarchical structure is valid beyond some threshold, t_*, we introduce the following control. First we compute the similarities $S(e_{ik}, e_{jk})$ for all connected edge pairs (e_{ik}, e_{jk}), as normal. We then perform our standard single-linkage hierarchical clustering, merging all edge pairs in descending order of S for $S \geq t_*$, fixing the community structure up to $t = t_*$. Below t_*, we randomly shuffle similarities among the remaining edge pairs with $S < t_*$, then proceed with the merging process as before. This randomization only alters the merging order, and ensures that the rate of edge pair merging is preserved, because the same similarities are clustered. This strictly controls not only the merging rate but also the similarity distributions and the high-quality community structure found at t_*. This procedure ensures that the dendrogram is properly randomized while other salient features are conserved. Full details are in Supplementary Information, section 7.4.