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

* These authors contributed equally to this work.
† To whom correspondence should be addressed. E-mail: slehmann@iq.harvard.edu.

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] groups of related nodes that correspond to functional subunits such as protein complexes[5–7] 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.[5,11–13] 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 biologically relevant link communities in protein-protein interaction[6,7,14] and metabolic networks[15] and show that a large social network[10,16] 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 that 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.[17] A popular measure of community quality, modularity, is defined by comparing the the number of connections within a community with the expected number of connections within the community under randomisation of the network.[18] However, these standard definitions of community structure break down when overlap is pervasive. In many real networks, nodes typically possess multiple roles.[6,7,9,10,14,15] Pervasive overlap in real networks is distinct from 'fuzzy' community overlap with relaxed interfaces,[19–21] because overlap can exist for each and every node (Fig. 1a,b). When overlap is pervasive, counterintuitively, each community has many more external than internal connections. This 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. 1a-c).

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. 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 interconnected. Each link is defined in a single context, allowing for a unique hierarchical tree (where each leaf is a link from the original network) to be constructed (see Methods). The result is a dendrogram whose branches represent link communities. In this dendrogram, links occupy unique positions and nodes naturally occupy multiple positions, due to their links. Agglomerating links leads to a dendrogram containing clearer and richer information than those of traditional methods. Extracting communities by cutting this dendrogram at various thresholds reveals the overlapping communities at multiple levels.

By clustering links we can now formulate overlapping community discovery as a well-posed optimisation problem, embracing overlap at every node without penalising that nodes participate in multiple communities. For this purpose, we in-

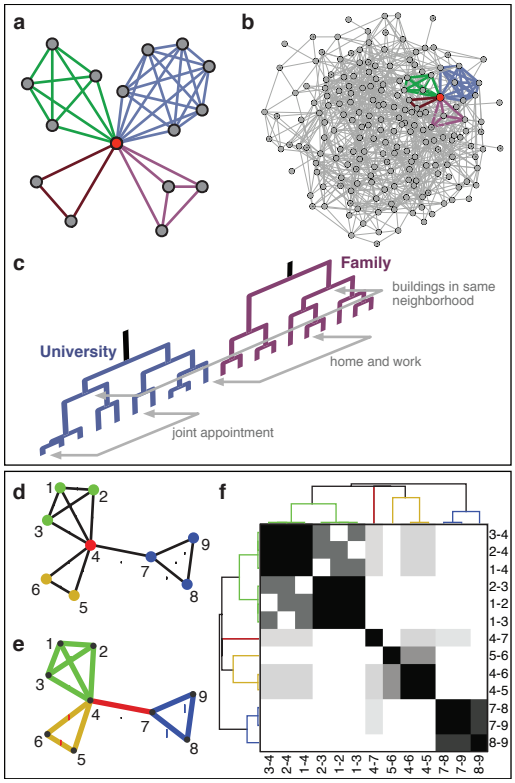

Figure 1: Overlapping communities lead to dense networks and prevent the discovery of a single node hierarchy. (**a**) Locally, structure in social networks is simple: an individual node sees the communities it belongs to. (**b**) Complex global structure emerges when *every* node is in the situation displayed in (**a**). (**c**) Pervasive overlap hinders the discovery of hierarchical organisation since nodes exist simultaneously in many leaves throughout the dendrogram, preventing a single tree from encoding the full hierarchy. Bottom Panel, an example network with (**d**) node communities and (**e**) link communities. (**f**) The link similarity matrix (darker matrix elements show more similar pairs of links) and resulting dendrogram. See SI for additional examples.

troduce a natural objective function, the *partition density* D, based on the link density (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. 2); one can also optimise D directly.

To investigate multi-scale structural complexity in real networks, we study link communities in a social network derived from the anonymised billing records of a mobile phone company (with a total of 8 million subscribers), representing the call patterns and locations of each user.[10,16,22] We generated a network of reciprocal calls between the users who make at least one call during a 30-week period within a particular 350 km by 80 km region which contains several large cities (Fig. 2a). We partition the link dendrogram at the threshold with maximum partition density (see also Fig. 4a). The three largest communities, at this optimum, are spatially correlated in the regions surrounding a major city (Fig. 2b). 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. 2c). Near the leaves, however, we find smaller, tightly clustered groups located inside densely populated regions (Fig. 2c). In Fig. 2e, we plot the network topology of the largest community from Fig. 2c, showing the multi-scale complexity of the underlying social group. Finally, Fig. 2f shows the highly overlapping structure in the largest sub-community. The dendrogram for this subgroup explicitly shows significant hierarchical structure alongside pervasive overlap. Additional validation of the discovered structure is presented in the Supplementary Information (SI).

We analyse recently published protein-protein interaction (PPI) networks of *Saccharomyces cerevisiae*, compiled into three genome-scale networks:[14] yeast two-hybrid (Y2H), affinity purification followed by mass spectrometry (AP/MS), and literature curated (LC). We also study the metabolic network reconstruction of *E. coli* K-12 MG1655 strain (iAF1260), one of the most elaborate reconstructions currently available.[15] We contrast link- and node communities (see Methods) on this test set, which covers networks from sparse (Y2H, $\langle k \rangle \sim 3$) to dense (*E. coli*, $\langle k \rangle \sim 17$), and from networks that are highly modular (AP/MS, LC) to networks with no visually apparent modular structure (*E. coli*). Figure 3 shows that, based on GO-terms and pathway annotations, link communities have more biological relevance, across all types of networks. Several specific example communities, as well as lists of all discovered communities with their most enriched annotations, are contained in the SI and Supplementary Table 1,2.

Detailed statistics for the metabolic and phone networks are presented in Fig. 4a which contains coverage, the ratio of second largest to largest community sizes s_2/s_1, and partition density D, as a function of the clustering threshold. The community size distribution at the optimum D is heavy tailed for both networks (Fig. 4b). The number of communities per node distinguishes the two networks (Fig. 4b insets): We can identify currency metabolites (water, ATP, etc.) by the high number of communities they participate in. Meanwhile, mobile phone users are limited to a smaller range of community memberships, most likely due to social and time constraints.

In summary, we have studied hierarchical organisation in the presence of pervasive community overlap. To incorporate both overlap and hierarchy, we developed a general approach based on hierarchical link communities. In most networks, it is a realistic assumption that links, rather than nodes, are characterised by a single attribute, such as community assignment. From this simple initial assumption we have resolved a major conflict in complex network research: how to combine community overlap with hierarchical structure. Many current techniques for analysing network structure,[12,13] identify hierarchical structures well, but are unable to correctly analyse networks with pervasive community overlap. One community detection method, clique percolation,[9] successfully accounts for strong overlap, but suffers from problems due to sparsity and is unable to describe the large-scale hierarchical structure of real

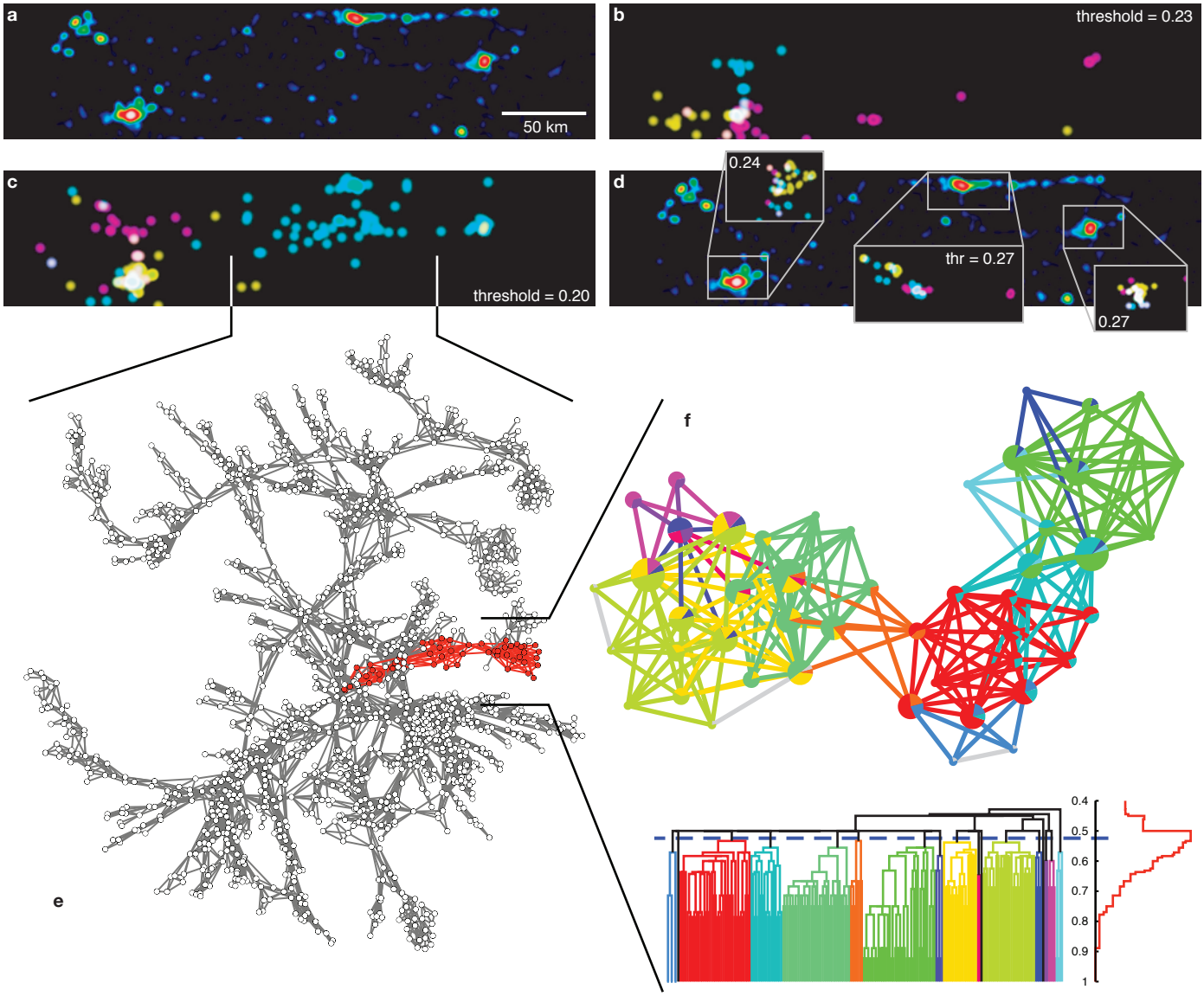

Figure 2: Spatial and nested structures are found at many levels in a mobile phone network. (**a**) Total population density. (**b**) The three largest communities at the optimum threshold cluster around a single city. (**c**) At a lower threshold, the largest communities become spatially extended, but still show correlation. (**d**) High thresholds yield smaller, intra-city communities. (**e**) The largest community in (**c**) with largest sub-community highlighted. (**f**) The highlighted sub-community in (**e**), along with the link dendrogram and Partition Density as a function of clustering threshold.

networks. Link communities incorporate both aspects simultaneously and stands out when compared to node clustering. Our link-centric viewpoint addresses the long-standing question of formulating overlapping community detection as an optimisation problem by introducing a new objective function, the partition density. Not only are strong overlap and hierarchical organisation not mutually exclusive, real networks possess both elements simultaneously.

Methods

Link similarity measure Define the *inclusive* neighbours $n_+(i)$ as the neighbours of node i, and node i itself. Limiting ourselves to link pairs that share a node, which are expected to be more similar than disconnected link pairs, the similarity S between links e_{ik} and e_{jk}, sharing node k can now be given by, e.g., the Jaccard index:

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

The shared node k does not appear in S because it provides no additional information and introduces bias. The SI contains a detailed discussion of this measure as well as generalisations to multipartite and weighted graphs.

Hierarchical clustering Each link is initially assigned to its own community; then, at each step, the pair of links with

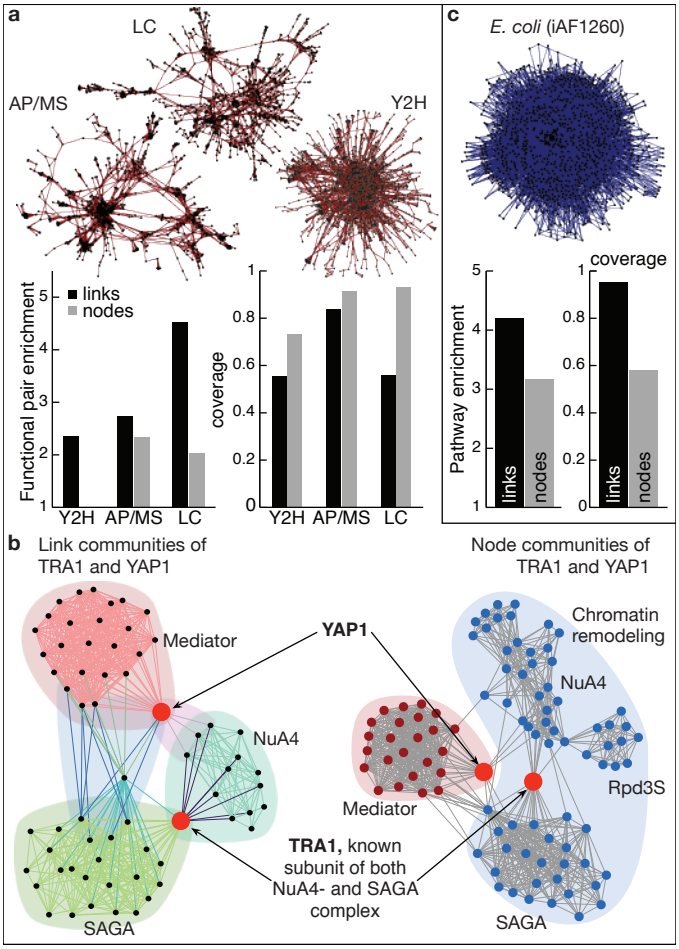

Figure 3: Link and node communities in biological networks. (**a**) The PPI networks of *S. cerevisiae* (Y2H, AP/MS, and LC) are displayed along with enrichment of functionally similar pairs[14] and coverage, the fraction of nodes placed into communities (see Methods). Link clustering consistently finds more relevant communities than node clustering in all of these networks. (**b**) An example from the compendium PPI network of *S. cerevisiae*, showing the communities around protein TRA1, illustrating the importance of overlap and the rich information contained within link communities. Node clustering groups the distinct functional complexes of TRA1 into a single community while link clustering correctly identifies complexes. (**c**) Similarly, we show the *E. coli* metabolic network (iAF1260), which lacks observable global modular structure, in contrast to the networks used in previous studies,[11,23] along with pathway enrichment and coverage. In this denser network, with more pervasive overlap, link clustering outperforms node clustering at both enrichment and coverage. See SI for details regarding measures, algorithms, and more examples.

the largest similarity is chosen and their respective communities are merged (single-linkage). Ties are agglomerated simultaneously. This process is repeated until all links have been agglomerated into a single cluster.

Partition density For a network with M total links and N total nodes, define $P = \{P_1, \ldots, P_C\}$ as a partition of its 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). We define the link density D_c of subset c as

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

In other words, this quantity is the number of links in community c, normalised by the minimum and maximum number of links possible between those same nodes, assuming they remain connected. We now define the *partition density* D as the average of D_c over all communities, weighted by the fraction of links present in each:

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

Node communities As with link similarity, similarity between node i and j can be defined as $S(i,j) = |n_+(i) \cap n_+(j)| \,/\, |n_+(i) \cup n_+(j)|$. Node communities (used in Fig. 3) are generated using the same single-linkage hierarchical clustering; the node dendrogram is cut at maximum modularity.[18] This approach is closely related to the method used in Ravasz et al.[11] This method was chosen to be as similar to link clustering as possible in order to be a fair control.

Enrichment and coverage To compare sensitivity, we introduce a coverage measure, defined as the fraction of nodes that belong to at least one community of three or more nodes, which is the smallest size for a non-trivial grouping of nodes. To test specificity, we use functional similarity for proteins, and pathway similarity for metabolites (see SI for details).

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.[25] The authors thank A.-L. Barabási, S. Ahnert, J. Park, D.-S. Lee, P.-J. Kim 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).

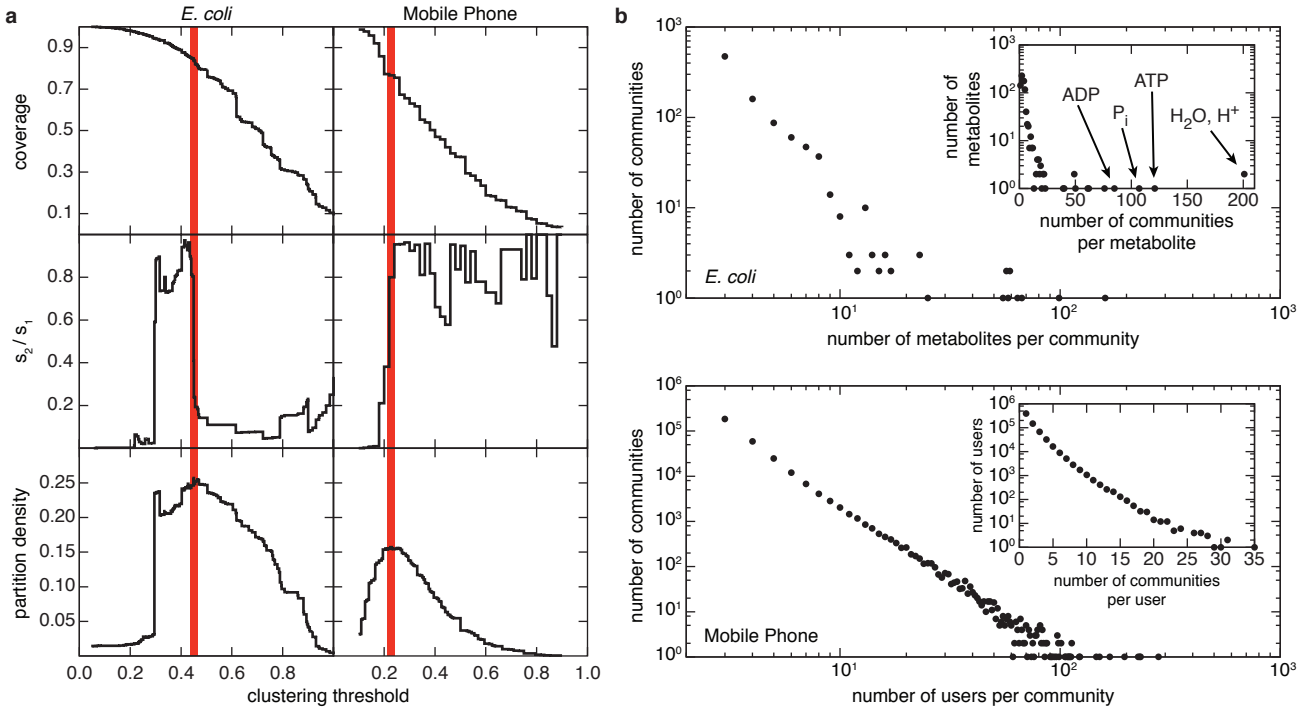

Figure 4: Statistics for the *E. coli* metabolic and mobile phone networks. (**a**) Coverage, the ratio of the number of edges in the two largest communities, and the partition density D, respectively. The denser metabolic network requires a higher threshold to separate compared to the mobile phone data. In both networks, peaks in D correspond to s_2/s_1 nearing 1/2, a possible transition point.[24] (**b**) The distribution of community sizes and node memberships (insets). Currency metabolites, such as water, belong to many communities, as expected. See SI for protein-protein interaction networks.

Supplementary Information

Link Communities Reveal Multi-Scale Complexity in Networks
by Yong-Yeol Ahn, James P. Bagrow, Sune Lehmann

S1 Network Datasets

Here we discuss the biological and social datasets used throughout this work.

S1.1 Biological networks

We analyzed the protein-protein interaction (PPI) network of *Saccharomyces cerevisiae* and the metabolic network of *Eschericia coli*. We use a recently published dataset of PPI networks compiled into three genome-scale networks: yeast two-hybrid (Y2H), affinity purification followed by mass spectrometry (AP/MS), and literature curated (LC).[14] Various statistics for the PPI networks are shown in Fig. S5. To validate the biological relevance of the communities discovered by Hierarchical link clustering (HLC), in Sec. S1.1.2 we compile these three network into a compendium of protein interactions; otherwise the three networks are kept separate.

We also use a metabolic network reconstruction of *E. coli* K-12 MG1655 strain (iAF1260), one of the most elaborate metabolic network reconstructions currently available.[15] From this reconstruction, we retain only cellular reactions, ignore information regarding the compartments (cytoplasm and periplasm), and project the network into metabolite space (two metabolites are connected if they share a reaction). For instance, if an enzyme catalyzes the metabolites A and B into C and D, the resulting network would contain a clique of A, B, C, and D.

This set of biological networks covers a wide range of network topologies, from sparse (Y2H, $\langle k \rangle \sim 3$) to dense (the metabolic network of *E. coli*, $\langle k \rangle \sim 17$), and from networks that are highly modular (AP/MS, LC) to networks with no visually apparent modular structure (*E. coli*). These networks are shown in Figs. 3A–C and 3E of the main text.

S1.1.1 Global statistics

To compare each community detection method's **sensitivity**, we use a coverage measure, defined as the fraction of nodes that belong to at least one community with three or more nodes. This size threshold is introduced since clique percolation (CPM) can only find communities of size three or more, by definition. (HLC and most modularity-based methods assign every node/edge into at least one community.) In order to test **specificity**, we use a functional similarity measure for proteins, and a pathway similarity measure for metabolites:

Proteins We adopt the same measure as the paper that published the datasets.[14] The enrichment of functionally similar pairs of proteins for a community c is defined by $\frac{N_{cs}}{N_c} / \frac{N_{as}}{N_a}$, where N_c is the number of possible pairs of proteins within the community boundary regardless of the existence of links between them, N_{cs} is the number of functionally similar pairs among N_c pairs based on their Gene Ontology (GO) Biological Process annotations ,[26] $N_a = N(N-1)/2$ is the total number of possible pairs in the network, and N_{as} is the number of functionally similar pairs among all N_a pairs. Functional similarity is determined by the total ancestry measure with a p-value cutoff of 10^{-3}.[27]

Metabolites We use two measures for the metabolic network. The first is defined in the same sense as the PPI's functional enrichment measure, as J_c/J_a, where J_a is the average Jaccard overlap of pathways between every pair of metabolites in the network and J_c is the average Jaccard overlap between every possible pair of metabolites within community c. The Jaccard overlap between a pair of metabolites a, b is calculated by $J(a,b) = |P_a \cap P_b| \,/\, |P_a \cup P_b|$, where P_m is the set of pathways that contain metabolite m. The second is defined by $\sum_i N_i^{\max}/N_i$, where N_i represents the number of metabolites that have at least one pathway annotation in a community i, and $N_i^{\max}$ is the number of metabolites in the largest subset of community i which share the same pathway annotation.

S1.1.2 Biological relevance of detected communities

In analyzing each community's biological relevance, we use two networks: a single compendium of the Y2H, AP/MS, and LC datasets; and the metabolic network of *E. coli*. We evaluate the resulting communities using biological annotations. For the PPI network, we perform GO-term enrichment analysis to identify each community's biological role(s) or correspondence to existing protein complexes. For the metabolic network, we use the pathway annotation of each compound to identify the probable role(s) of each community in the metabolism.

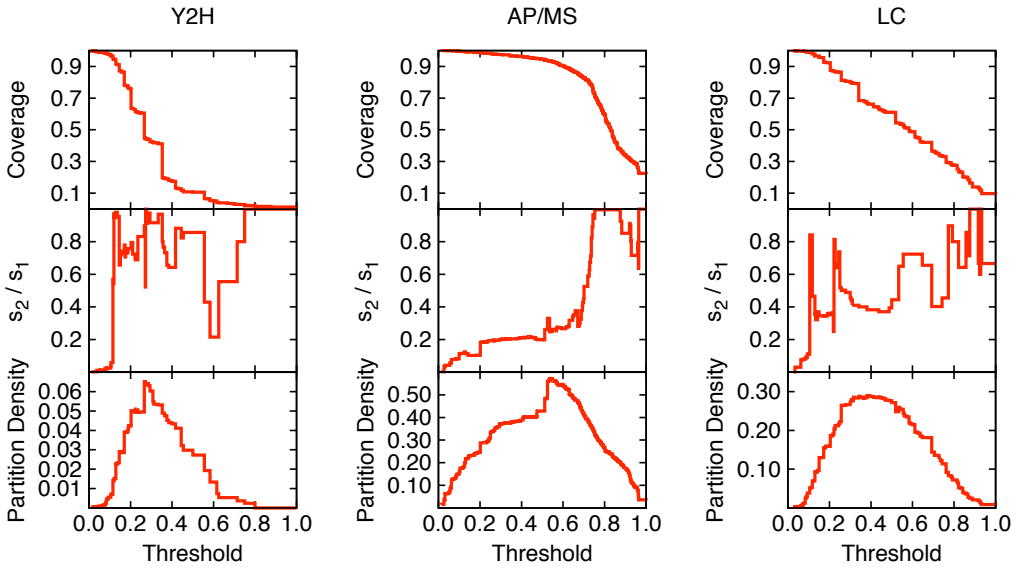

Figure S5: Several statistics for the protein-protein interaction networks. Compare with Fig. 4 in the main text.

We use GO-TermFinder software[28] version 0.82 to find enriched GO terms and estimate the p-values for each GO term. First, we find all GO terms with p-value less than 0.05, then we pick up only the most significant term for each aspect (biological process, cellular component, molecular function). These terms and p-values are listed along with the community members in Supplementary Table 1. This table shows that more than 80% of communities have at least one enriched GO-term with p-value lower than 0.0001 and more than 30% of communities have at least one enriched GO-term with p-value lower than 10^{-10}.

For the metabolic network, we first filter out communities where less than three members possess pathway annotations. Then, we calculate the enriched pathway annotations shared by the largest number of community members. We compile this information in Supplementary Table 2.

S1.1.3 Examples of community structure

Fig. S6 shows the community structure around protein YML007W. There are three major communities, all three are related to the transcription process, identified as the mediator complex, NuA4 HAT complex, and SAGA complex,[29–31] respectively. Note the overlapping membership of protein YHR099W, which is already known as a subunit of both NuA4 complex and SAGA complex.[32–34] Figure S7 shows three major communities around the protein YBL041W, which belongs to the core of the proteasome complex.[35] We can directly observe that the proteasome consists of two parts: the core and the regulatory particles, and HLC finds two corresponding communities plus a community connecting the two. As expected from the structure of the proteasome, the core is less exposed to other communities, while the regulatory particles have several connected communities. Likewise, Fig. S8 shows the community structure around Acetyl-CoA, illustrating several roles that Acetyl-CoA plays in the metabolic network.

S1.2 Mobile phone network

This dataset catalogs approximately 8 million users, all calls among these users, and the locations of users when they initiate a phone call (the tower from which the call originated). Self-reported demographic information such as age and gender is also available for most users. We generate the network by constraining the location to a 350 km by 80 km region and two nodes in the region are connected only if they each call the other person at least once during a 30- week period. We assign to each user a single location, that of the tower they most frequently used. The final network contains approximately 600 thousand nodes and 2.8 million edges.

Applying HLC, the partition density and coverage, as a function of the threshold, are shown in Fig. S9. This shows that HLC achieves much better coverage than clique percolation at its preferred value of $k = 4$.[36]

As in biological networks, coverage is not the only important aspect of community detection. In the case of the mobile phone network, we can also use external information, the age and geographic location of the users, to qualify the accuracy of the discovered partition. First, we compute the age difference between pairs of nodes across the network and then for pairs within the same community. In a similar manner, we can look at the spatial "spread" of each community by making the assumption that each node is located at the cell tower it most frequently uses and computing the standard deviation σ in the distances between nodes in the same community. When compared to randomized communities we again find strong spatial clustering. See Fig. S10.

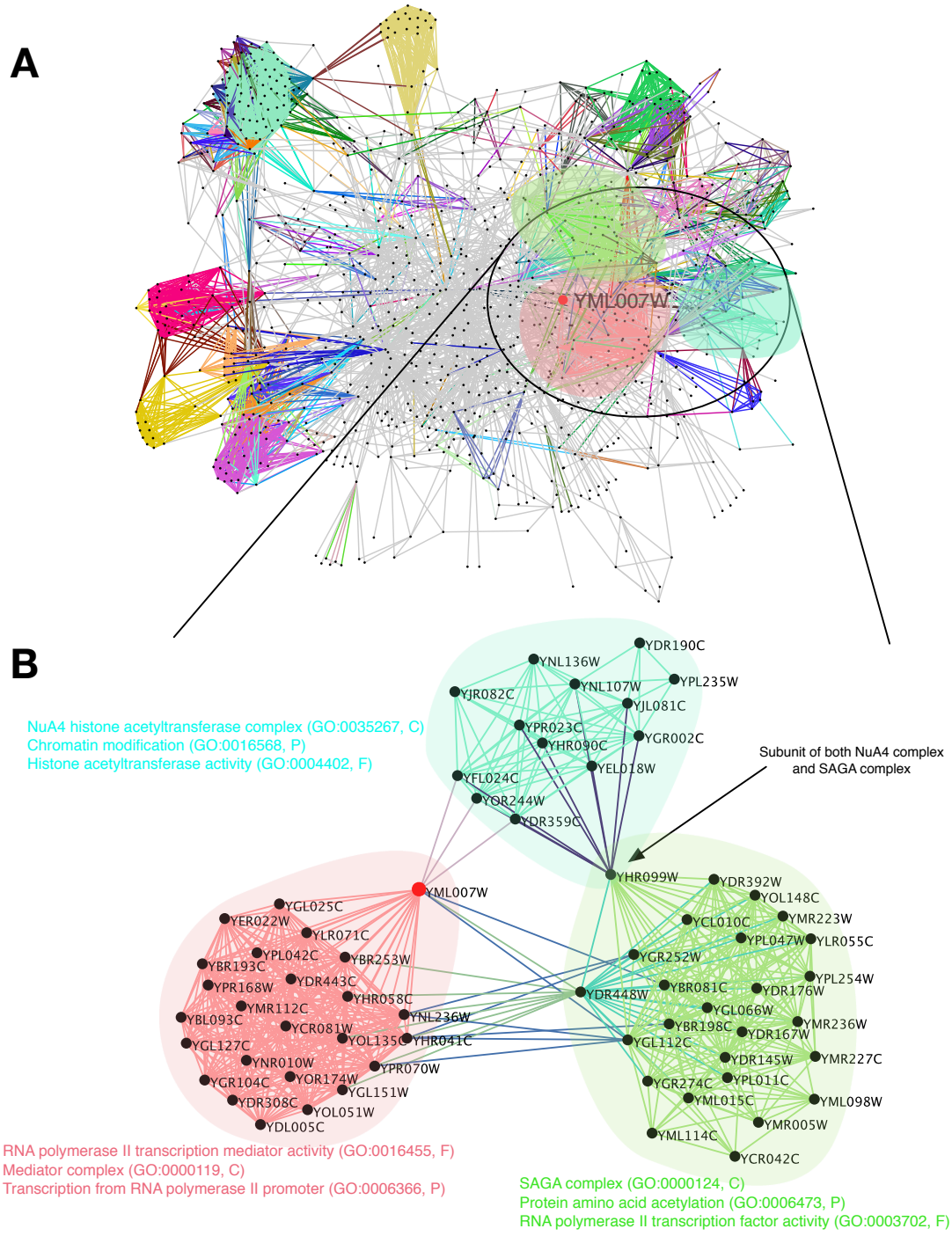

Figure S6: An example of overlapping community structure in the PPI compendium network. (**A**) The subnetwork surrounding protein YML007W (snowball sampled out to three steps). (**B**) The communities around YML007W. Only GO terms with p-value smaller than 10^{-10} are displayed (with colors corresponding to their communities).

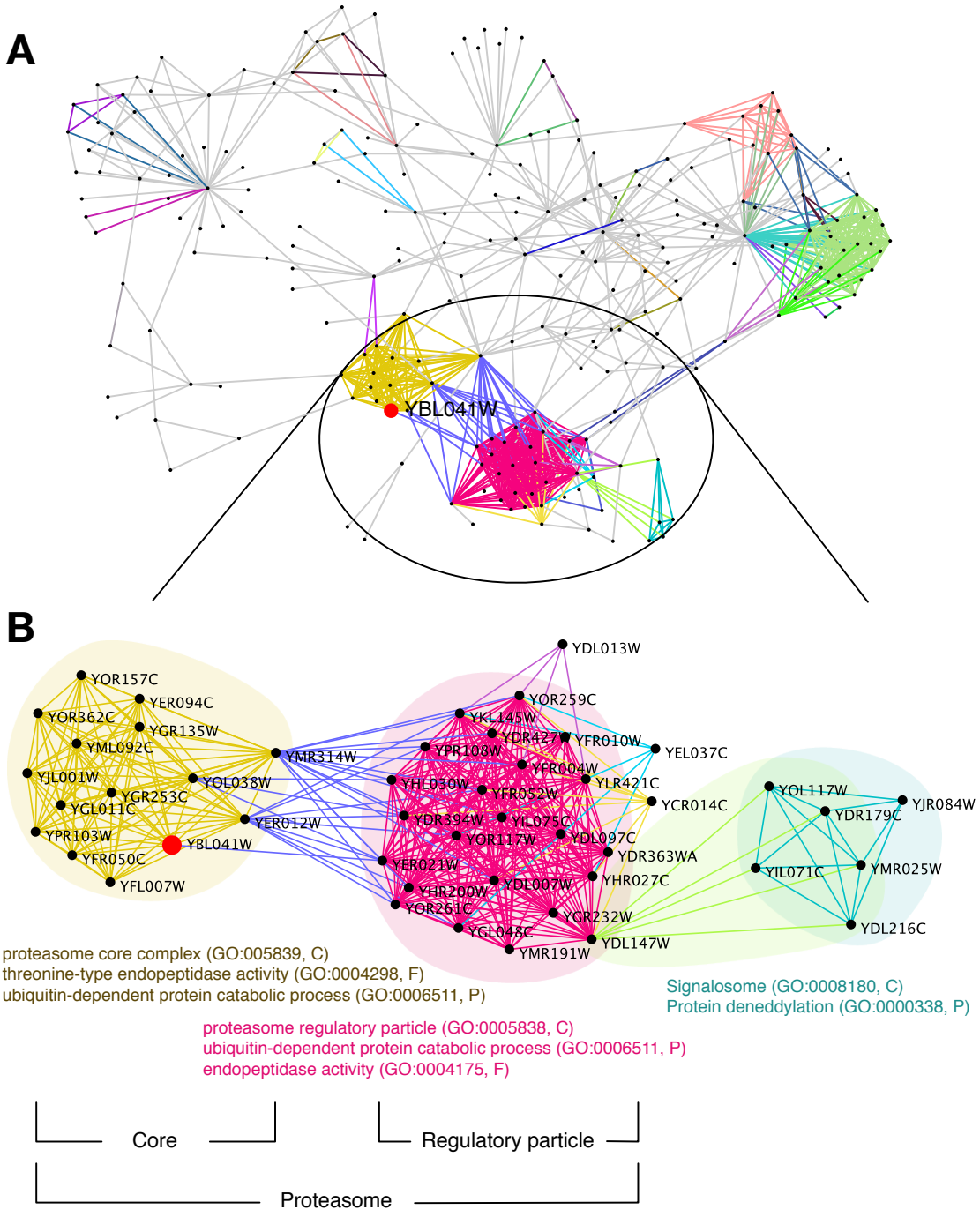

Figure S7: Another example of overlapping community structure. (**A**) The subnetwork surrounding protein YBL041W (snowball sampled out to three steps). (**B**) The communities surrounding YBL041W. Only GO terms with p-value smaller than 10^{-10} are displayed (with colors corresponding to their communities). These communities correspond to the core and the regulatory particles of the proteasome complex and a community connecting the two.

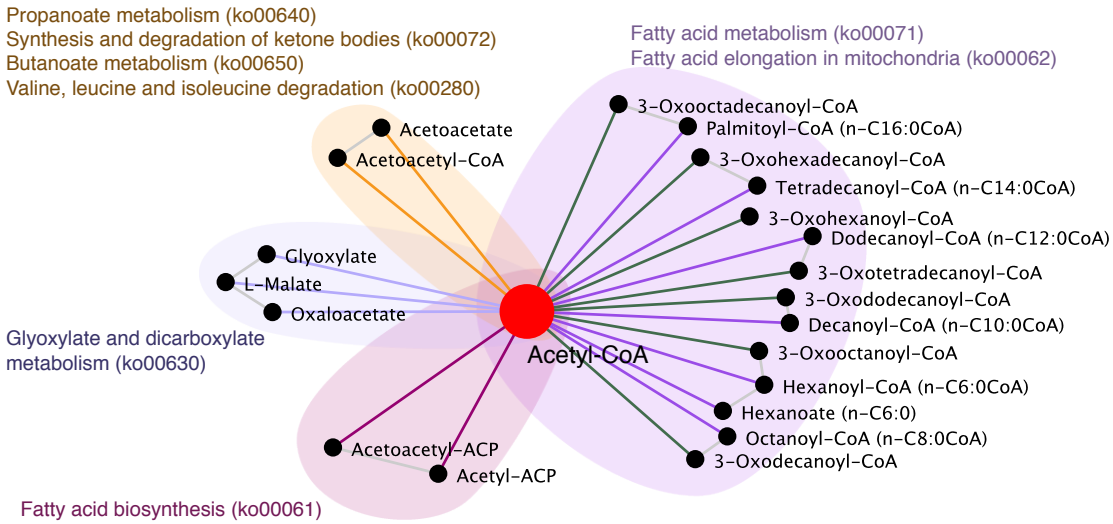

Figure S8: Overlapping community structure around Acetyl-CoA in the *E. coli* metabolic network. Acetyl-CoA plays several different and important roles in metabolism. Shown are only communities with homogeneity score equal to 1 (all compounds inside each community share at least one pathway annotation); all other links, including those that contribute to community structure, are omitted. Pathway annotations shared by all community members are displayed with corresponding colors. The two communities to the right of Acetyl-CoA are grouped since they share the same exact pathway annotations.

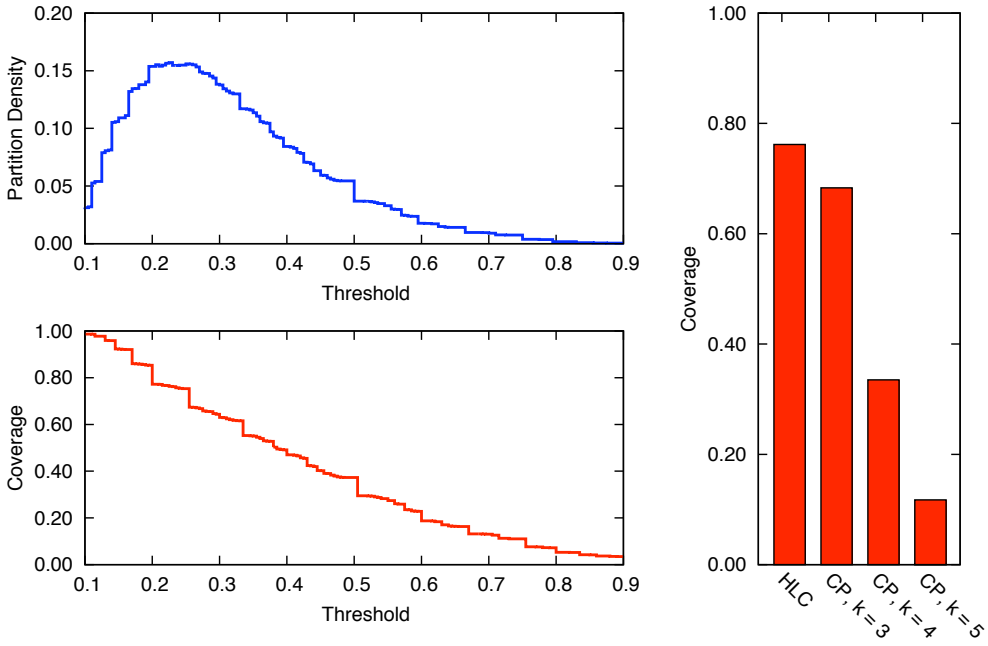

Figure S9: (**Left**) The partition density and coverage as a function of the clustering threshold for the mobile phone network. (**Right**) The coverage (defined in Sec. S1.1) at maximum partition density for HLC compared to that of clique percolation. At the optimum threshold of 0.23, HLC achieves better coverage, more than twice that of clique percolation (the authors in[36] use $k = 4$ exclusively).

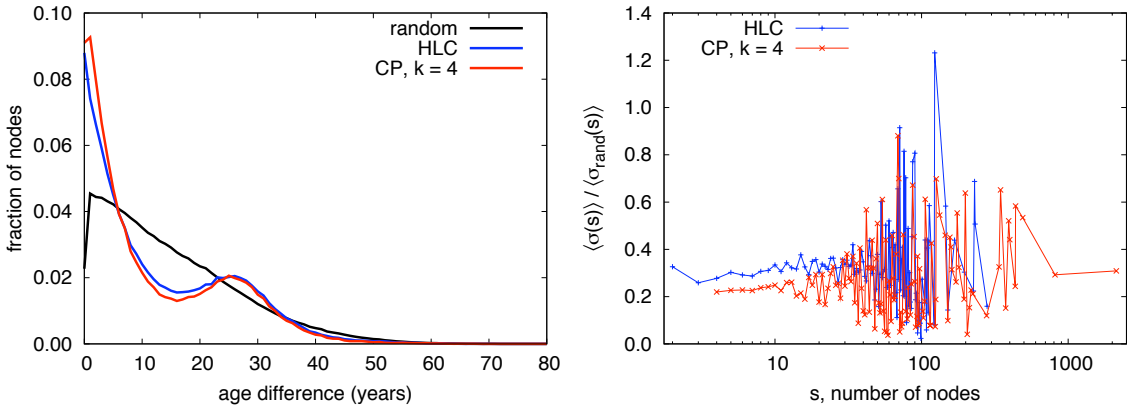

Figure S10: Using demographic information to qualify communities in the mobile phone network. Note that HLC achieves twice the coverage of CPM. (**Left**) The age difference for random pairs of nodes chosen from the entire network and chosen from within discovered communities. An average age for new parents of ~ 27 years is immediately evident from just cell phone records. (**Right**) A comparison of the geographic 'dispersion' of nodes inside communities. Shown is the standard deviation $\langle \sigma(s) \rangle$ of the geographic locations (most probable towers) of nodes within the same community, averaged over communities with the same number of nodes s versus $\langle \sigma_{\mathrm{rand}}(s) \rangle$, the same quantity but from randomly chosen sets of nodes of size s. The plot confirms that both methods find significant, spatially correlated structure. CPM finds especially good structures, slightly outperforming HLC, albeit with much less coverage overall.

These results emphasize the point that CPM does well when detecting tightly knit communities. However, in the case of HLC, we are able to vary the clustering threshold to obtain fine- grained control, tuning for larger or smaller communities and observing hierarchical community structure spanning from the level of small communities consisting of only a few nodes, to large groups signifying much broader societal structures. For instance, Fig. 2 in the main text shows loosely connected large-scale communities (more than 500 people), which span the scale of cities (and are geographically distinct).

S1.3 Abundance of overlap

The abundance of overlap between communities is evident in social networks as pointed out in .[9] Intuitively this finding makes immediate sense: individuals belong to several distinct communities corresponding to friends, family, etc. The same concept also applies to biological networks. To underscore this point, Fig. S11 shows the distribution of functions per protein and pathways per metabolite. Although the number of functional categories and the number of pathways from databases does not directly correspond to the exact number of protein complexes or to that of metabolic pathways, it clearly shows that the overlap cannot be ignored in finding modules in biological systems. In the case of proteins, approximately two thirds of all proteins currently belong to more than one functional category. Although the metabolic network seems to exhibit less overlap than PPI, it is obvious that the currency metabolites such as water, proton, or ATP participate in a broad spectrum of pathways. According to the KEGG database,[37] the number of pathways assigned to water is only 5. However, HLC puts water into more than 200 communities (Fig. 4 in the main text), which correctly captures the abundant nature of currency metabolites .[38]

Finally, the dense network shown in Fig. 1B is meant to be an illustration of the consequences of strong community overlap. However, it was constructed using an existing social network model,[39] which suggested that social networks can be modeled by probabilistically projecting a bipartite network that consists of people and communities.

S2 Methods

S2.1 Link clustering

S2.1.1 Constructing a dendrogram

The main text has introduced the HLC method to classify links into topologically related groups. Here we provide further motivation for the suggested pair-wise link similarity measure. For simplicity we limit ourselves to only *connected* pairs of links (i.e. sharing a node) since it is unlikely that a pair of disjoint links are more similar to each other than a pair of links that share a node; at the same time this choice is much more efficient. For a connected pair of links e_{ik} and e_{jk}, we call the shared node k a *keystone* node and i and j *impost* nodes.

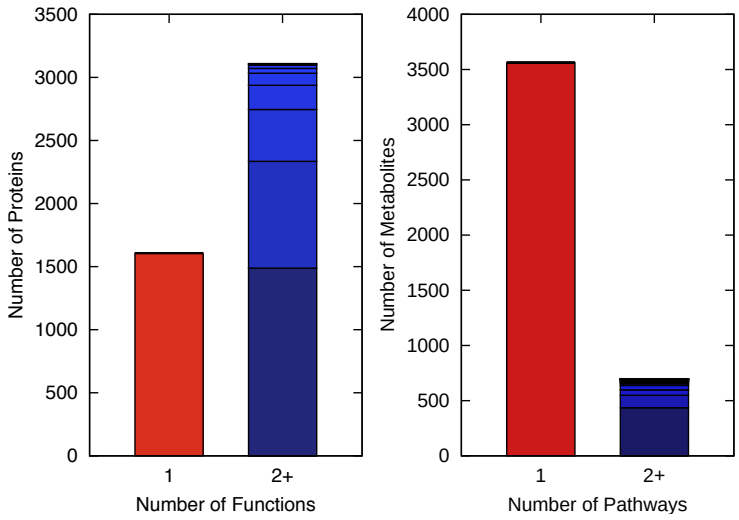

Figure S11: (**Left**) The number of functional categories per protein. Each box in the right bar indicates, from bottom, proteins with 2, 3, ... functions, respectively. We consider the highest hierarchy (26 categories) of the Functional Catalog (FunCat) of the Munich Information center for Protein Sequence (MIPS) database as the protein functions. According to the catalog, nearly two thirds of all proteins have multiple functions. (**Right**) The number of pathways per metabolites. We use Kyoto Encyclopedia of Genes and Genomes[37] (KEGG) database for pathway annotation.

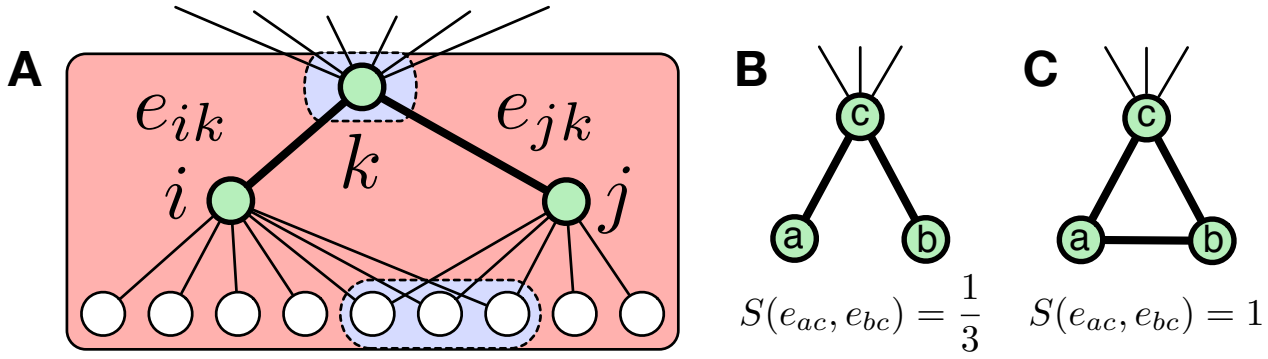

Figure S12: (**A**) The similarity measure $S(e_{ik}, e_{jk})$ between edges e_{ik} and e_{jk} sharing node k. For this example, $|n_+(i) \cup n_+(j)| = 12$ and $|n_+(i) \cap n_+(j)| = 4$, giving $S = 1/3$. Two simple cases: (**B**) an isolated ($k_a = k_b = 1$), connected triple (a,c,b) has $S = 1/3$, while (**C**) an isolated triangle has $S = 1$.

If the only available information is the network topology, the most fundamental characteristic of a node is its neighbors. Since a link consists of two nodes, it is natural to use the neighbor information of the two nodes when we define a similarity between two links. However, since the links we are considering already share the keystone node, the neighbors of the keystone node provide no useful information. Moreover, if the keystone node is a hub, then the similarity is likely to be dominated by the keystone node's neighbors. For instance, if the hub's degree increases the similarity between the links connected to the hub also increases. This bias due to the keystone node's degree also prohibits us from applying traditional methods directly to the *line graph* of the original graph, which is constructed by mapping the links into nodes. (Since a hub of degree k becomes a fully connected subgraph of size k in the line graph, the community structure can become radically different.) Thus, we neglect the neighbors of the keystone. We first define the *inclusive* neighbors of a node i as:

$$n_+(i) \equiv \{x \mid d(i,x) \leq 1\} \tag{S4}$$

where $d(i,x)$ is the length of the shortest path between nodes i and x. The set simply contains the node itself and its neighbors. From this, the similarity S between links can be given by, e.g., the Jaccard index:[40]

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \tag{S5}$$

An example illustration of this similarity measure is shown in Fig. S12 (See Sec. S3.1 for generalizations of the similarity).

With this similarity, we use single-linkage hierarchical clustering to find hierarchical community structures. We use single-linkage mainly due to simplicity and efficiency, which enables us to apply HLC to large-scale networks. However, it is also possible to use other options such as complete-linkage or average-linkage clustering. Each link is initially assigned to its own community; then, at each time step, the pair of links with the largest similarity are chosen and their respective communities are merged. Ties, which are common, are agglomerated simultaneously. This process is repeated until all links belong to a single cluster. The history of the clustering process is then stored in a dendrogram, which contains all the information of the hierarchical community organization.

The similarity value at which two clusters merge is considered as the strength of the merged community, and is encoded as the height of the relevant dendrogram branch to provide additional information.

S2.1.2 Partitioning the dendrogram

Hierarchical clustering methods repeatedly merge groups until *all* elements are members of a single cluster. This eventually forces highly disparate regions of the network into single clusters. To find meaningful communities rather than just the hierarchical organization pattern of communities, it is crucial to know where to partition the dendrogram. Modularity has been widely used for similar purposes in node-hierarchies,[18,41] but is not easily defined for overlapping communities.[1] Thus, we introduced a new quantity, the *partition density* D, that measures the quality of a link partition. The partition density has a single global maximum along the dendrogram in almost all cases, because the value is just the average density at the top of the dendrogram (a single giant community with every link and node) and it is very small at the bottom of the dendrogram (most communities consists of a single link). This process is illustrated in Fig. S14.

S2.1.3 Link dendrograms and node hierarchy

A link dendrogram can be very different from a node dendrogram, see Fig. S13. As an (admittedly extreme) example, consider the graph shown in Fig. S15. Here we have constructed a simple network with two levels of hierarchy, consisting of four very dense communities, loosely connected into pairs which are then more loosely connected. At the lower level of the link dendrogram, we find six communities, not the expected four. The reason is that HLC has correctly identified the two sets of cross-community links.

S2.2 Node clustering

As a control, we compare HLC to a Hierarchical Node Clustering (HNC) method. The HNC method is used in Fig. 3 of the main text. HNC is closely related to the method introduced in Ravasz et al.[11] There are many ways to define a similarity between two nodes. We tried four different variations of the node similarity. The four versions are following:

- $S(i,j) = |n(i) \cap n(j)|/|n(i) \cup n(j)|$,
- $S(i,j) = |n(i) \cap n(j)|/\min(k_i, k_j)$,
- $S(i,j) = |n_+(i) \cap n_+(j)|/|n_+(i) \cup n_+(j)|$,
- $S(i,j) = |n_+(i) \cap n_+(j)|/\min(k_i, k_j)$,

where $n(i)$ means the neighbors, not inclusive neighbors, of the node i. Among those, we use the version in Eq. (S6) since it finds more relevant communities across most networks we used. In addition, it is the definition most similar to link similarity. Thus, the node similarity is chosen to be

$$S(i,j) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}, \tag{S6}$$

where, as in the main text, $n_+(i)$ are the inclusive neighbors of node i. To determine the node dendrogram, we use the same single linkage hierarchical clustering as we used for clustering links. This node dendrogram is cut at the point of maximum modularity.[18]

S2.3 Other methods

In order to evaluate its performance, we have compared HLC to existing, popular community detection methods. We chose two representative algorithms: the Clique Percolation Method (CPM)[9] and a modularity-based agglomerative clustering algorithm.[45] We apply all three frameworks (HLC, CPM and modularity) to the biological networks studied in Fig. 3 of the main text, PPI (Y2H, AP/MS, LC) and the *E. Coli.* metabolic network. The results are displayed in Fig. S16.

The modularity method[45,46] by definition identifies the membership of all nodes, but, as a consequence, the resulting communities are the least relevant in most cases. These results also highlight limitations of CPM's more rigid community definition. In the metabolic network, CPM's coverage is largely due to one giant community containing most nodes, leading to a miniscule enrichment value. Removing the giant community increases the enrichment value to close to 8, but only 12 small communities ($\sim 5\%$ of nodes) remain. This situation is hardly changed by increasing clique size. For Y2H, however, the problem is sparsity: there are not enough cliques to find structure. When the network is too dense, the network becomes super-critical in the sense of clique percolation and leads to giant clique communities. In contrast, when the network is too sparse, the network is sub-critical and there are not enough connected cliques to find.

[1] Several modifications of modularity that allow for "fuzzy" communities with relaxed interfaces (or overlapping nodes) to exist[19–21,42,43] have been suggested. However, in order to avoid the trivial optimum, where all nodes are part of all communities, each of these methods *penalize* overlap, and are therefore not suitable for networks with pervasive overlap. (See Fig. 1b of the main text)

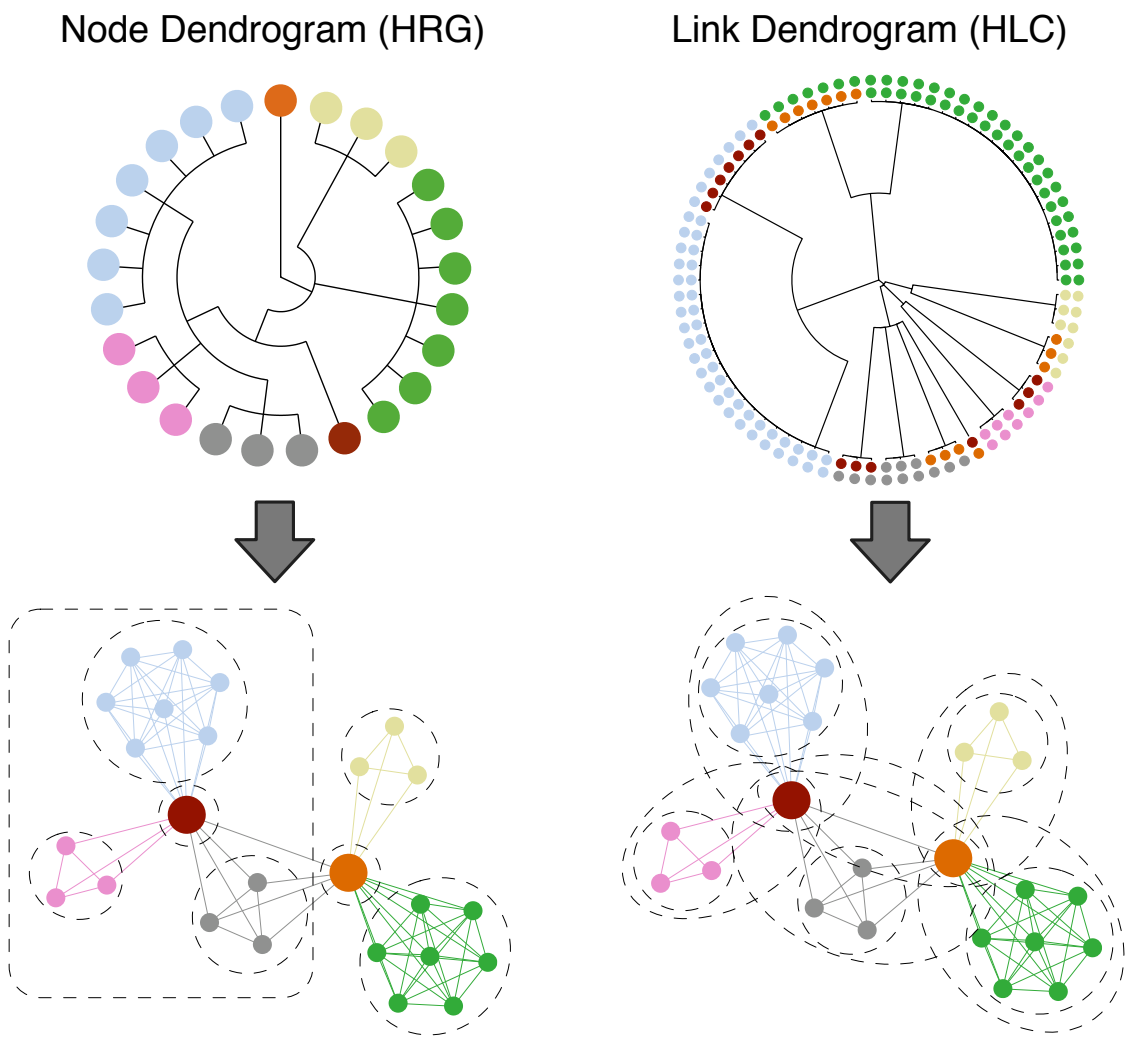

Figure S13: Comparison of a node dendrogram and link dendrogram in the presence of overlap. The node dendrogram is obtained by using the HRG method (consensus dendrogram) ,[13] and the link dendrogram is obtained from HLC. Nodes are colored to distinguish each node or clique and dotted lines represent several hierarchies in the dendrogram. In the link dendrogram, two colored circles at each leaf represent the link between the nodes with the given colors. Note that HRG scatters the red, orange, and gray nodes in the dendrogram, even though they belong to the same clique. One cannot retrieve the clique community that consists of red, orange, and gray nodes. In contrast, the link dendrogram captures every clique while at the same time constructing a reasonable hierarchical tree. Note that the links of the red node are placed in appropriate branches of the dendrogram according to their context. Also note the internal hierarchical structures found inside each clique. Finally, real networks possess significantly more overlap than this example.

Network	Y2H	AP/MS	LC	*E. coli* (iAF1260)
Modularity Q	0.715801	0.679996	0.836293	0.335831

Table S1: The modularity values for the four biological networks studied in the main text, found using modularity optimization.[46]

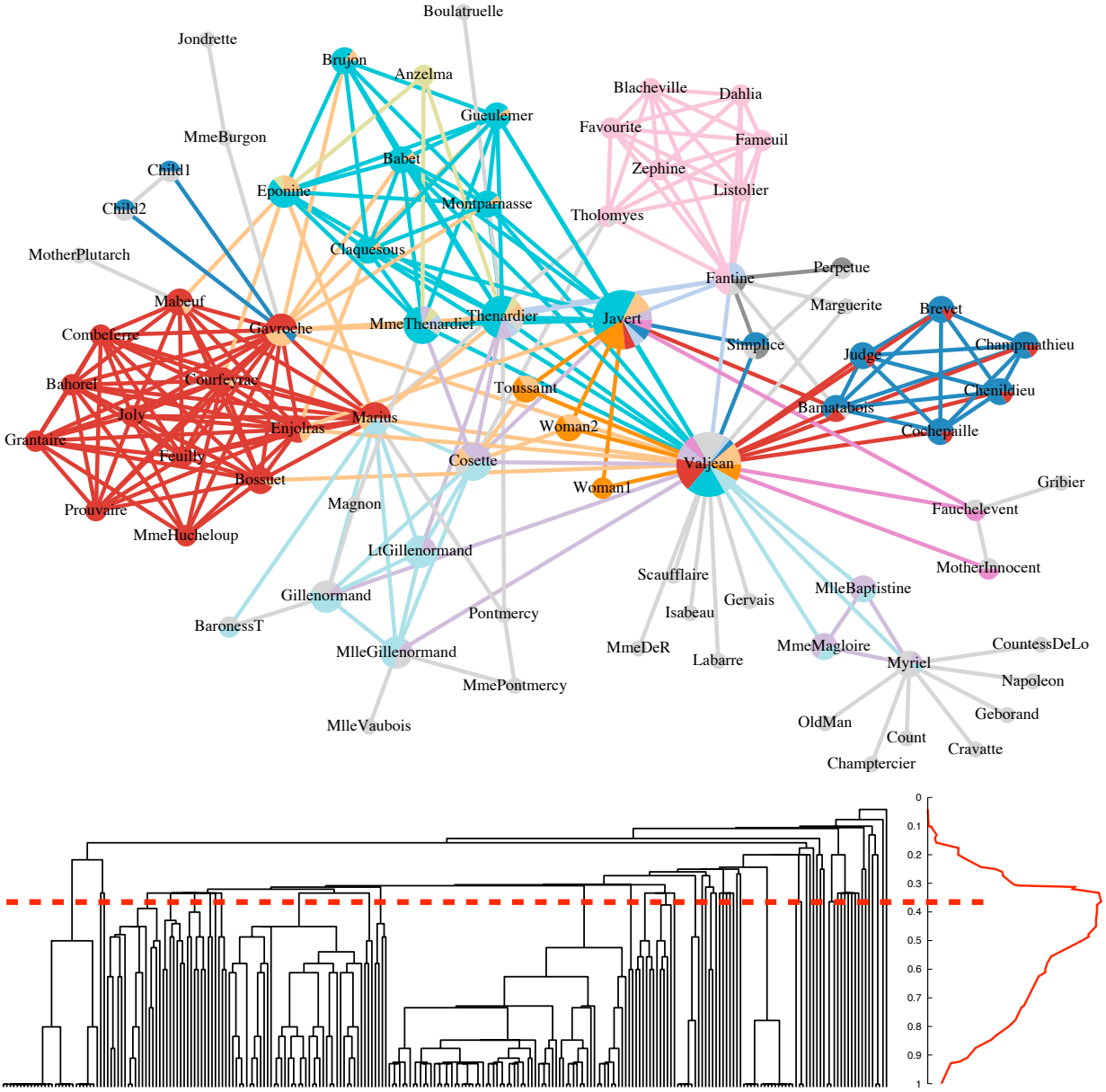

Figure S14: An example of link clustering for the coappearance network of characters in the novel *Les Misérables*.[44] (**Top**) the network with link colors indicating the clustering, with grey indicating single-link clusters. Each node is depicted as a pie-chart representing its membership distribution. The main characters have more diverse community membership. (**Bottom**) the full link dendrogram and partition density. Note the internal blue community in the large blue and red clique containing Valjean. HLC discovers hierarchical structure even inside of cliques.

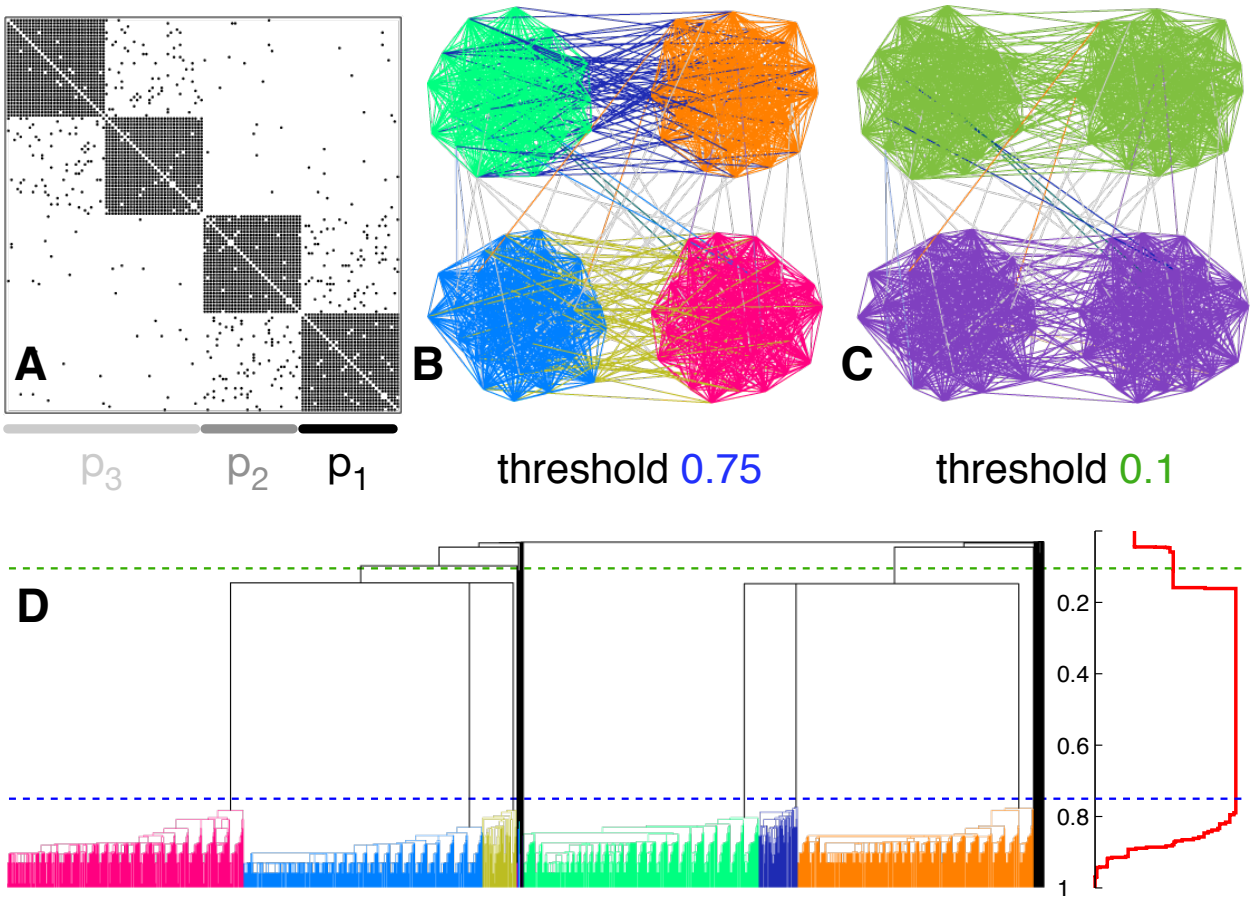

Figure S15: Building link dendrogram intuition. Shown is an example illustrating how hierarchy can be captured at multiple levels of the link dendrogram. (**A**) The 128×128 adjacency matrix for a network of four densely connected communities (each possible link exists with probability p_1), each connected to another community (p_2), and finally the two pairs are weakly connected (p_3). For this example, $p_i = \frac{1-\epsilon}{12^{i-1}}, \epsilon = 0.02$. The communities at a high (**B**) and low (**C**) threshold, and the full dendrogram (**D**) are shown. The chosen values of p_i lead to a very "stretched" dendrogram and partition density, as expected. While one expects to identify four communities at the higher threshold, six are actually found, since the inter-community edges are identified by HLC.

S2.3.1 Modularity optimization

Although the particular modularity algorithm used here is the most popular one, more accurate methods exist, based on simulated annealing, extremal optimization, and more. (See[41] for additional details.) However, the modularity values we found are quite high, so the lack of accuracy in our comparison is more likely due to neglecting overlap rather than failing to find good partitions. The particular values found for the four biological networks are shown in Table S1. Note that visibly modular networks such as AP/MS and LC show high modularity values.

S2.3.2 Hierarchy and overlap

Several prominent methods for finding hierarchical organization exist,[12,13] however, none are able to handle overlap since hierarchical structure always assumes disjoint community partition. In summary: CPM handles overlap, HRG handles hierarchy, but Hierarchical Link Clustering handles both. To further compare and contrast the three methods, see Fig. S17.

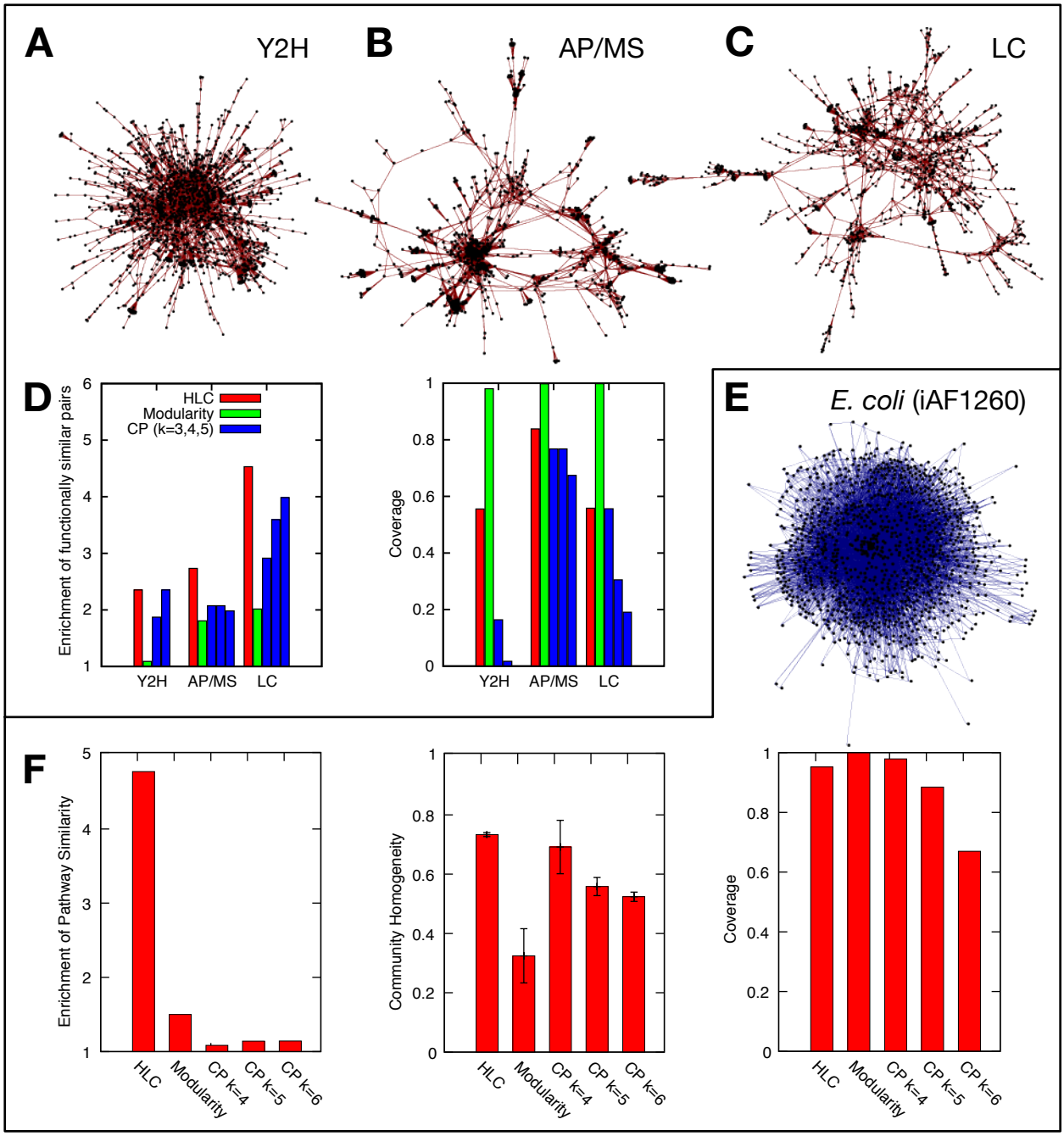

Figure S16: Evaluation of community detection methods [Hierarchical Link Clustering, Clique Percolation,[9] and modularity[45,46]] using biological networks. Top, the PPI networks of *S. cerevisiae* (**A**) Y2H, (**B**) AP/MS, and (**C**) LC, respectively. (**D**) Enrichment of functionally similar pairs[14] and coverage shows that HLC performs as well or better than other methods. (**E**) The *E. coli* metabolic network (iAF1260), which lacks observable global modular structure, in contrast to the networks used in previous studies.[11,23] (**F**) Pathway similarity and coverage in the metabolic network. Standard error is shown in the community homogeneity histogram.

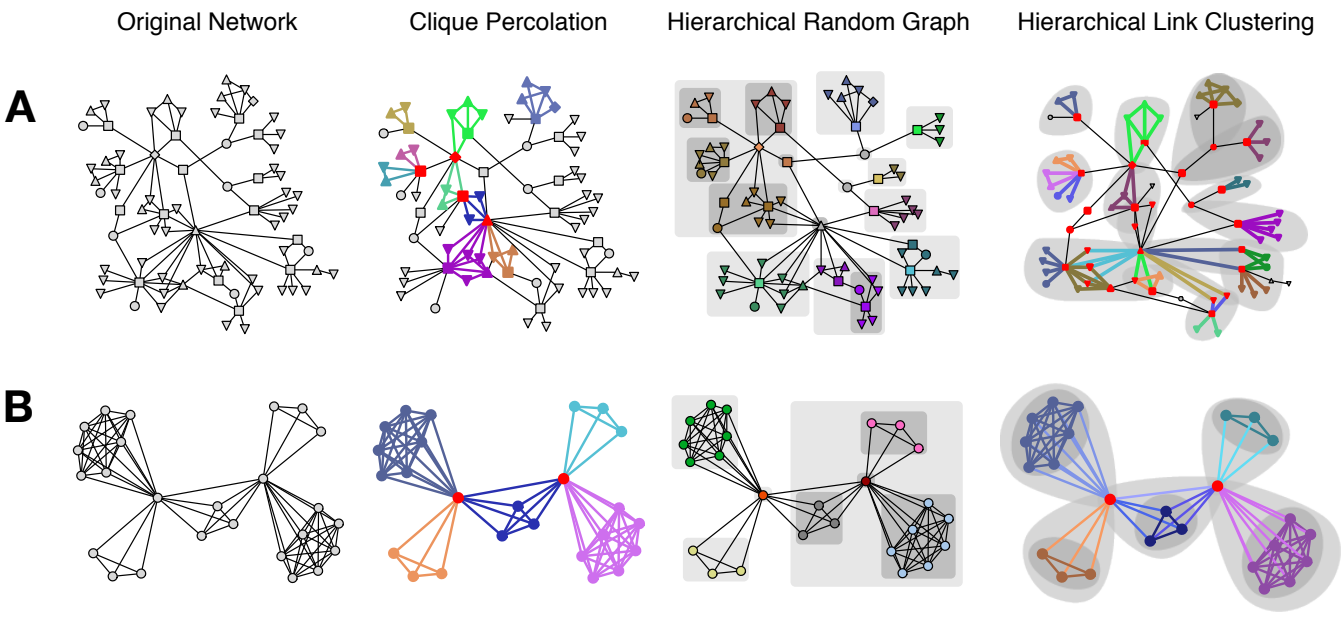

Figure S17: Comparison of methods on a network of UK grassland species interactions ,[47] which has evident hierarchical structure (**A**), and on a simple example with overlapping communities (**B**). Colors and boxes indicate community structures while nested boxes illustrate hierarchical information. Red nodes possess multiple community memberships. The performance of existing methods depends heavily on the network's structural characteristics. CPM fails to detect the structure in sparse, hierarchical networks (**A**). The HRG model captures the hierarchical structure in (**A**) but neglects overlap, and forces the middle 5-clique in (**B**) to be arbitrarily spread across branches. In the case of hierarchical link clustering, both hierarchy and overlapping structures are correctly classified. Again, real social networks possess more overlap than in (**B**).

S3 Generalizations and Extensions

S3.1 Networks with weighted, directed, or signed links

The similarity between links can be easily extended to networks with weighted, directed, or signed links (without self-loops), since the Jaccard index generalizes to the Tanimoto coefficient.[48] Consider a vector $\mathbf{a}_i = \left(\widetilde{A}_{i1}, \ldots, \widetilde{A}_{iN}\right)$ with

$$\widetilde{A}_{ij} = \frac{1}{k_i} \sum_{i' \in n(i)} w_{ii'} \delta_{ij} + w_{ij} \tag{S7}$$

where w_{ij} is the weight on edge e_{ij}, $n(i) = \{j | w_{ij} > 0\}$ is the set of all neighbors of node i, $k_i = |n(i)|$, and $\delta_{ij} = 1$ if $i = j$ and zero otherwise. The similarity between edges e_{ik} and e_{jk}, analogous to Eq. (S5), is now:

$$S(e_{ik}, e_{jk}) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{|\mathbf{a}_i|^2 + |\mathbf{a}_j|^2 - \mathbf{a}_i \cdot \mathbf{a}_j} \tag{S8}$$

S3.2 Multi-partite networks

A multi-partite network is a network in which the nodes can be divided into K disjoint sets and all links must terminate in two distinct sets. This creates additional constraints on the existence of certain edges which must be accounted for in both the link similarity and the partition density.

Link similarity: The similarity measures, Eqs. (S5) and (S8), depend only upon connectivity, and therefore automatically account for multi-partite structure. The one change necessary is incorporating the forbidden connections between the same kind of nodes, which can be achieved by using the set of neighbors instead of the inclusive neighbor set when calculating the similarity.

Partition density: We must modify the definition of the partition density since a fully connected K-partite clique is much sparser

than a clique in a unipartite network. In general, the K-partite partition density of a subset c can be written as

$$D_c^{(K)} = \frac{m_c + 1 - \sum_k n_c^{(k)}}{\sum_k \left(n_c^{(k)} \sum_{k' \neq k} n_c^{(k')} \right) - 2 \left[\left(\sum_k n_c^{(k)} \right) - 1 \right]}, \tag{S9}$$

where the index k runs over the K node types and the notation $n_c^{(k)}$ refers to nodes of type k. The full partition density is achieved by summing over individual communities, $D^{(K)} = 2M^{-1} \sum_c m_c D_c^{(K)}$.

S3.3 Local methods

Since our definition of similarity between links only uses local information, a local version [49–51] of HLC can be trivially obtained. One can simply choose a starting *link*, compute its similarity S with all adjacent links, agglomerate the one with the largest S into the community, compute any new similarities between edges inside the community and bordering it, and repeat. A stopping criteria to determine when the community has been fully agglomerated is still necessary .[50] For instance, one can monitor the partition density as links are agglomerated, in order to establish a reasonable community boundary. Another, simpler, approach is to fix the similarity threshold and agglomerate only links with similarity larger than that threshold. To find all the overlapping communities of a node one can simply begin the above methods with each of that starting node's links or start from one link, find its community (which may end up including another starting node link), then pick another unassigned link from the starting node, find that community, and repeat until all the starting node's links are contained within communities.

S3.4 Partition density optimization

Since the partition density is a quality function of link community structures in networks, it is possible to find link communities by direct optimization. Begin by assigning links to communities at random, then use, e.g. simulated annealing. The disjoint nature of link communities enables us to apply many traditional optimization techniques to find overlapping communities.

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

[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).

[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] 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).

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

[17] 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).

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

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

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

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

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

[23] Guimerà, R. & Amaral, L. A. N. *Nature* **433**, 895 (2005).

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

[25] Evans, T. S. & Lambiotte, R. Line graphs, link partitions and overlapping communities. *arXiv:0903.2181* (2009).

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

[27] Yu, H., Jansen, R., Stolovitzky, G. & Gerstein, M. Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics* **23**, 2163–2173 (2007).

[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). URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/20/18/3710. http://bioinformatics.oxfordjournals.org/cgi/reprint/20/18/3710.pdf.

[29] Doyon, Y., Selleck, W., Lane, W. S., Tan, S. & Côté, J. Structural and functional conservation of the nua4 histone acetyltransferase complex from yeast to humans. *Mol. Cell. Biol.* **24**, 1884 (2004).

[30] Dotson, M. R. *et al.* Structural organization of yeast and mammalian mediator complexes. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 14307–14310 (2000). URL http://www.pnas.org/content/97/26/14307.abstract. http://www.pnas.org/content/97/26/14307.full.pdf+html.

[31] Wu, P.-Y. J., Ruhlmann, C., Winston, F. & Schultz, P. Molecular architecture of the s. cerevisiae saga complex. *Molecular Cell* **15**, 199 – 208 (2004). URL http://www.sciencedirect.com/science/article/B6WSR-4CXCPJ1-8/2/d893d6c2ca8fb5606c8c186f8885

[32] Saleh, A. *et al.* Tra1p Is a Component of the Yeast AdaSpt Transcriptional Regulatory Complexes. *J. Biol. Chem.* **273**, 26559–26565 (1998). URL http://www.jbc.org/cgi/content/abstract/273/41/26559. http://www.jbc.org/cgi/reprint/273/41/26559.pdf.

[33] Brown, C. E. *et al.* Recruitment of HAT Complexes by Direct Activator Interactions with the ATM-Related Tra1 Subunit. *Science* **292**, 2333–2337 (2001). URL http://www.sciencemag.org/cgi/content/abstract/292/5525/2333. http://www.sciencemag.org/cgi/reprint/292/5525/2333.pdf.

[34] Bhaumik, S. R., Raha, T., Aiello, D. P. & Green, M. R. In vivo target of a transcriptional activator revealed by fluorescence resonance energy transfer. *Genes & Development* **18**, 333–343 (2004). URL http://genesdev.cshlp.org/content/18/3/333.abstract. http://genesdev.cshlp.org/content/18/3/333.full.pdf+html.

[35] Baumeister, W., Walz, J., Zühl, F. & Seemüller, E. The proteasome: Paradigm of a self-compartmentalizing protease. *Cell* **92**, 367 – 380 (1998). URL http://www.sciencedirect.com/science/article/B6WSN-419K592-8/2/de10e433366c37b67fcd75cfbe88a

[36] Palla, G., Barabási, A.-L. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664 (2007).

[37] Kanehisa, M. & Goto, S. *Nucleic Acids Res.* **28**, 27–30 (2000).

[38] Huss, M. & Holme, P. Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *Systems Biology, IET* **1**, 280–285 (2007).

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

[40] Jaccard, P. étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* **37**, 547–579 (1901).

[41] Danon, L., Diaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* **2005**, P09008 (2005).

[42] Shen, H., Cheng, X., Cai, K. & Hu, M.-B. Detect overlapping and hierarchical community structure in networks. *Physica A* **388**, 1706–1712 (2009).

[43] Nicosia, V., Mangioni, G., Carchiolo, V. & Malgeri, M. Extending the definition of modularity to directed graphs with overlapping communities. *J Stat Mech-Theory E* P03024 (2009).

[44] Knuth, D. E. *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, Reading, MA, 1993).

[45] Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Physical Review E* **69**, 066133 (2004).

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

[47] Martinez, N. D., Hawkins, B. A. & adn B. P. Feifarek, H. A. D. *Ecology* **80**, 1044–1055 (1999).

[48] Tanimoto, T. T. An elementary mathematical theory of classification and prediction. Tech. Rep., IBM Internal Report (1958).

[49] Bagrow, J. P. & Bollt, E. M. A local method for detecting communities. *Phys. Rev. E* **72**, 046108 (2005).

[50] Bagrow, J. P. Evaluating local community methods in networks. *J. Stat. Mech.* **2008**, P05001 (2008).

[51] Clauset, A. Finding local community structure in networks. *Physical Review E* **72**, 026132 (2005).