OPEN ACCESS Freely available online

PLoS one

Discovering Communities through Friendship

Greg Morrison[1]*, **L. Mahadevan**[1,2,3]

1 School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, United States of America, **2** Wyss Institute of Biological Engineering, Harvard University, Cambridge, Massachusetts, United States of America, **3** Department of Physics, Harvard University, Cambridge, Massachusetts, United States of America

Abstract

We introduce a new method for detecting communities of arbitrary size in an undirected weighted network. Our approach is based on tracing the path of closest-friendship between nodes in the network using the recently proposed Generalized Erdös Numbers. This method does not require the choice of any arbitrary parameters or null models, and does not suffer from a system-size resolution limit. Our closest-friend community detection is able to accurately reconstruct the true network structure for a large number of real world and artificial benchmarks, and can be adapted to study the multi-level structure of hierarchical communities as well. We also use the closeness between nodes to develop a degree of robustness for each node, which can assess how robustly that node is assigned to its community. To test the efficacy of these methods, we deploy them on a variety of well known benchmarks, a hierarchal structured artificial benchmark with a known community and robustness structure, as well as real-world networks of coauthorships between the faculty at a major university and the network of citations of articles published in *Physical Review*. In all cases, microcommunities, hierarchy of the communities, and variable node robustness are all observed, providing insights into the structure of the network.

Citation: Morrison G, Mahadevan L (2012) Discovering Communities through Friendship. PLoS ONE 7(7): e38704. doi:10.1371/journal.pone.0038704

Editor: Yamir Moreno, University of Zaragoza, Spain

Received February 13, 2012; **Accepted** May 13, 2012; **Published** July 20, 2012

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gmorriso@seas.harvard.edu

Introduction

The topology of networks occurring in biological or chemical [1,2], social [3,4], political [5], or technological [6] systems can give profound insights into a variety of important aspects of these systems, such as the processes that generated the network [7], the stability of the system [8] or the properties of processes occurring on it [9]. An important aspect of common real-world networks is that of community structure [10], where subsets of the network are densely connected internally and weakly connected externally. Nodes in the same community have more in common than those in distinct communities, reflected in the topology of denser intra-community edges than inter-community edges. However, the detection of communities in networks without apriori knowledge of their structure is highly nontrivial, and methods for community detection have recently attracted a great deal of interest.

Perhaps the most common approach for community detection in networks is based on modularity maximization [11,12]. Each node i in a network of N nodes and M edges is assigned to a single community, c_i, with the partition chosen to maximize

$$Q=\frac{1}{2W}\sum_{ij}(w_{ij}-\frac{W_iW_j}{2W})\delta(c_i,c_j), \qquad (1)$$

where w_{ij} is the weight of the edge between nodes i and j, $W_i=\sum_j w_{ij}$ is the strength of node i, $W=\frac{1}{2}\sum_i W_i$, and $\delta(c_i,c_j)=1$ if $c_i=c_j$ and 0 otherwise. For an unweighted network, $w_{ij}\equiv a_{ij}=$ 0 or 1, where a_{ij} is the adjacency matrix, and thus $W_i=k_i$ is the degree of the node. Modularity compares the network in question to a randomly generated network with each node constrained to have the same strength, and is maximized by a partition into communities $\{c_i\}$ that have a higher intra-community weight than would be expected randomly. This choice of a random network acts as a null model, although other choices are possible [13], and a wide variety of numerical approaches for efficiently computing the maximal partition exist, including statistical mechanical methods [14], bisection algorithms [11], and other greedy searches [15,16]. While modularity maximization is both intuitive and accurate in a variety of settings, Q has a natural system-size resolution limit [17,13]: if the number of nodes becomes large ($N\rightarrow\infty$), but the typical strength W_i of all nodes remains finite, the total strength $W\rightarrow\infty$ and the second term in the sum in Eq. 1 becomes small (since W_i and W_j do not diverge). Thus, modularity maximization may not detect small communities in large networks due to this resolution limit. Simple methods to overcome this limitation include the introduction of a resolution parameter [14,13] γ, with the redefinition of $Q=(2W)^{-1}\sum_{ij}(w_{ij}-\gamma W_iW_j/2W)\delta(c_i,c_j)$, or multiresolution methods [18] which impose a self-loop of strength r on the network (i.e. $w_{ij}\rightarrow w_{ij}+r\delta_{ij}$) in Eq. 1. Both of these approaches overcome the problem of a resolution limit by introducing an arbitrary parameter in detecting community structure that must be tuned. Alternate approaches to community detection avoid a resolution limit through other means, such as thresholding the resistance distance between nodes, with nodes having low resistance distance between each other belonging to the same community [19], maximizing the 'fitness' of each node in a greedy fashion [20], creating block models to detect communities if the number of expected communities is exactly known [21], or refining communities by finding 'statistically significant' nodes [22]. In all these approaches, at least one free parameter is required to detect the communities, which may be useful in giving the ability to tune the resolution at which

communities are detected, but with no a-priori method for determining the 'correct' value that leads to a meaningful partition.

In this paper, we develop a new parameter-free, resolution-limit-free method for community detection, most easily understood intuitively in the context of a social network: a person belongs in the same community as his or her 'closest friend' (the node to which he or she has the greatest measure of 'closeness,' discussed below). Our method requires a way to measure closeness (or friendship) between nodes in a network, and a variety of such measures are available [23]. We will focus primarily on a recently proposed non-metric measure of closeness [24], the Generalized Erdös Numbers (GENs), which have been found useful in a variety of contexts in understanding the structure of network topology. This closest-friend community detection method is shown to be able to accurately detect communities in a variety of widely used benchmarks, in some cases outperforming some modularity-maximizing detection schemes in real world networks with a known 'correct' partition. We also extend the method to detect community structure at a lower resolution (macrocommunities formed from higher resolution microcommunities) without appealing to a free parameter. Our approach has the advantages of being intuitively accessible, free of arbitrary parameters, and able to accurately find communities in complex networks. We leverage our chosen measure of closeness between nodes in determining the robustness of assignment of each node into its community (rather than a global measure of the quality of the partition using modularity). Finally, our approach is applied to a citation network and a coauthorship network, and the complex hierarchical structure of each network is examined in detail.

Methods

Communities from Closeness

In a network with community structure, nodes in a community have a higher density of edges internally (to other nodes in their community) than they do externally. While one approach to community detection maximizes global quality functions that depend on the density of edges [10], we could alternatively search for high densities of edges locally to find communities. Such a local method may use an appropriate measure of closeness between nodes, with 'close' nodes having multiple short-length paths between one another (implying a locally high density of edges; see below for examples). In the context of a social network, for example, it is natural to expect that closest friends (those who feel closest to one to another given a measure of 'closeness') should be found in the same community. Such an expectation can be enforced by determining the closest friend (CF) of each node i, denoted $f(i)$, and requiring them to be in the same community. In other words, node i is assigned to the same community as the node to which it is topologically closest. The closest friend of $f(i)$ (denoted $f(f(i))$) is also found in this community, and we generate a path of closest friendship $\mathbf{p}^i=\{i,f(i),f(f(i)),\ldots\}$ (halting when a self-intersection occurs after which the cycle would repeat). Nodes i and j that share elements of their closest friend paths (i.e. $|\mathbf{p}^i\cap\mathbf{p}^j|\neq 0$) will all trace to the same central loop, and each of the elements of $\mathbf{p}^i$ and $\mathbf{p}^j$ are placed in the same community. If the closeness measure is well chosen (such that a higher density of edges implies a stronger feeling of 'closeness'), the closest friend paths for nodes in each community will remain within the correct communities, allowing for an accurate partition of the network (discussed further in Supplementary Information S1). This approach has the advantage of generating a single partition (rather than a tree of many possible partitions from which the 'correct' partition must be chosen, commonly used in clustering algorithms) and without a system-size resolution limit [17,13], and therefore unambiguously chooses a 'natural' partition of the network.

Despite the simplicity of our method, there exist pathological network topologies may require modification of the algorithm in order to accurately detect the community structure. As a simple example, a node that is connected to every other node in the network will be *everyone's* closest friend, regardless of the topology of the rest of the network, and only one community will be detected using our approach (see Supplementary Information S1 for further discussion). Failure of the detection algorithm in this case can be avoided by searching for the closest *unpopular* friend (CUF), where the CUF is detected by sorting the closest friends of node i in descending order of node degree, and choosing the first node $f_u(i)$ who has degree less than or equal to the next-closest node. This ensures that we avoid nodes with extremely high degree (the popular close friends), who may have many out-of-community connections, and choose $f_u(i)$ to be a node that is simultaneously (a) a close friend (but not necessarily the closest) and (b) less likely to have out-of-community edges. The path of closest friendship is modified to be $\mathbf{p}_u^i=\{i,f_u(i),f_u(f_u(i)),\ldots\}$, and community detection proceeds as described above. We note that neither the CF nor CUF approaches depend on the graph being Hamiltonian: the particular path $\mathbf{p}^i$ or $\mathbf{p}_u^i$ need not span the entire graph for any starting node i (and must not, if there is to be more than one community). Additional modifications to both the CF and CUF methods are required due to community fracture: communities may be split into two or more disjoint pieces due to the random fluctuations of the edges [25] (see Supplementary Information S1 for further discussion). Fractured communities may occur for any community detection algorithms, and a greedy approach to detect and merge fractured communities is described in Supplementary Information S1.

Choosing a Closeness Measure

Before we apply the CF or CUF method for community detection, we must choose a measure of closeness between nodes in that network, with the only requirement being that nodes i and j are 'closer' if there is a higher density of edges (multiple short-ranged paths) between them. We focus on the use of a recently developed closeness measure, the Generalized Erdös numbers [24] (GENs), created with two simple principles in mind: (i) connections from node j to nodes that feel close to a specified node (nodes $\{k\}$ with low E_{ik}) are more important than connections to other nodes, and (ii) a connection of high weight from j to some node k should make node j feel more close to node k and less close to node i. This second expectation is natural if closeness is defined with a limited resource in mind, such as the time spent between people in a social or coauthorship network [24]. These expectations naturally lead to a weighted harmonic mean [24], with $E_{ii}=0$ and

$$\frac{W_j}{E_{ij}}=\sum_{k\in\mathbf{C}_j}\frac{w_{jk}}{E_{ik}+w_{jk}^{-1}}.$$

with $\mathbf{C}_j$ the set of nodes that are connected to j. E_{ij} is not a distance metric (as $E_{ij}\neq E_{ji}$), a desirable property because unpopular (low degree or low weight) individuals may feel close to popular (high weight) nodes, but not vice-versa. The GENs are computed numerically by setting $E_{ij}^{(0)}=(1-\delta_{ij})$ and iteratively computing $E_{ij}^{(t+1)}=W_j/\sum_k w_{jk}/(E_{ik}^{(t)}+w_{jk}^{-1})$, halting when $\max_{ij}|E_{ij}^{(t+1)}-E_{ij}^{(t)}|\leq\delta$ for some tolerance δ (we used 5×10^{-3}. Computing the closeness between all pairs of nodes i and j will scale as $N\times M$,

and is the slowest step in detecting communities using the CF or CUF approaches.

To see how our closeness measure works in detecting communities in a network with known community structure, we examine the Girvan-Newman benchmark [1,12] in Fig. 1(a), which consists of four equal-sized communities of 32 nodes, each with k^{out} edges leading out of the community and $16-k^{out}$ edges within the community. The connectivity between communities can also be described by the mixing parameter $\mu=k^{out}/(k^{in}+k^{out})=k^{out}/16$, with detection of the correct communities becoming difficult when $k^{out}\gtrsim 8$ or $\mu\gtrsim 0.5$. The level of agreement between the detected and correct partition is quantified using the normalized mutual information [10]:

$$I=2\frac{\sum\limits_{i\in P_t,j\in P_0} n_{ij}\log\left(\frac{Nn_{ij}}{n_i^t n_j^0}\right)}{\sum\limits_{i\in P_t} n_i^t\log(n_i^t/N)+\sum\limits_{j\in P_0} n_j^0\log(n_j^0/N)} \qquad (2)$$

with n_i^t the number of nodes in community i of the trial partition (P_t), n_j^0 is the number in community j of the true partition (P_0), and n_{ij} is the number simultaneously occurring in i and j of P_t and P_0. In Fig. 1(a), we see that the accuracy of the CUF approach does depend on the choice of closeness measure, where we compare the performance of the GEN measure with others [23] such as the overlap measure ($O_{ij}=|\mathbf{C}_i\cap\mathbf{C}_j|$ with $\mathbf{C}_j$ the set of neighbors of j) and the Jacard coefficient ($J_{ij}=|\mathbf{C}_i\cap\mathbf{C}_j|/|\mathbf{C}_i\cup\mathbf{C}_j|$). Similarly, in real-world networks with an apriori known community structure (shown in Fig. 1(b)) such as the Football network [1], the Political Blogs network [26], and the Political Books network [27] (see Supplementary Information S1), both the GENs and overlap are consistently more accurate in community detection than greedy modularity maximization. Because the GENs are the most accurate on both real world and artificial networks of all of the closeness measures attempted, we choose to focus on them as our measure of closeness in the rest of the paper.

Additional Benchmarks of Community Detection

As a systematic test of the method on a more complex benchmark, apply our detection method to the benchmark of Lancichinetti, Fortunato, and Radicchi [28]. Communities are of variable size (with the size s of each drawn from a power law distribution, $P(s)\sim s^{-\beta}$) and the degree of each node is drawn from a scale free distribution as well ($P(k)\sim k^{-\gamma}$). Each node has on average a fraction μ of its edges within its assigned community and $1-\mu$ edges outside of its community. The complex structure of this network makes community detection non-trivial, but as seen in Fig. 1(c-f) our method is accurately able to reconstruct the correct partition for various values of β, γ, and μ (for $N=1000$ and 50 realizations of the network for each data point). So long as $\mu\leq 0.5$, we typically find the normalized mutual information $I\gtrsim 0.9$, indicating a good agreement with the correct partition. Our approach produces partitions that are less accurate than the results reported in Fig. 5 of Ref. [28], in accordance with the observations in Fig. 1(a) that the method underperforms modularity maximization when the correct partition is also modularity maximizing. However, the CUF method still performs admirably, with the additional benefits of no fitting parameters or resolution limits.

Hierarchical Communities

In many cases [29,20] networks have community structure at multiple resolutions, begging the question of how to detect such a hierarchical community structure. Instead of using a tunable resolution parameter whose 'correct' value(s) are unknown a-priori, the CF/CUF method naturally suggests a simpler approach: to iteratively coarse grain the network using a high-resolution partition (detected as described above) and then reapply our detection method on the lower resolution network. Communities in the high-resolution partition act as coarse grained nodes, and the average closeness felt between communities serves to determine closest friends. If the GENs are chosen as the measure of closeness, the averages are taken as $(E_{hg}^c)^{-1}=\sum_{i\in g,j\in h}E_{ij}^{-1}/n_g n_h$, where n_g is the number of nodes in g. While the choice of a method of coarse graining the network implies an additional degree of freedom in our algorithm, it is important to note the differences between the CUF method and modularity maximization with a variable resolution parameter. In the CF/CUF method, the resolution can not be tuned continuously by choosing different closeness measures or methods of coarse graining. Rather, the choice of measure and method set an optimal apriori resolution for hierarchical community detection, which is likely to be robust to changes in the method if the closeness measure and the coarse graining method are well chosen.

The accuracy of our hierarchical detection method on a commonly used artificial benchmark, implemented in Ref. [18], is shown in Fig. 1(g), with additional benchmarks discussed further in Supplementary Information S1. A network of 256 nodes is formed from 16 communities of 16 nodes each, in turn composed of 4 macrocommunities containing 4 communities each. Each node has on average 13 edges within its community and 4 edges outside of its community but within its macrocommunity, and 1 edge outside of its macrocommunity. This is similar to the Reichardt and Bornholdt [14,20] benchmark discussed in Supplementary Information S1 and adapted in the next section. We compare the partitions detected using the CUF algorithm with a simulated annealing maximization of the multiresolution modularity (that is, Eq. 1 with $w_{ij}\rightarrow w_{ij}+r\delta_{ij}$, where r is a resolution parameter ranging from $r_{min}=-W/N$ to ∞). The average modularity Q_r for the modularity maximizing partition is shown by the red points in Fig. 1(g), and this modularity maximizing partition transitions smoothly between the high-resolution communities detected using our CUF algorithm for large r and the low-resolution coarse grained using our hierarchical algorithm for small r. Additional analysis of a similar benchmark for our hierarchical detection algorithm can be found in Supplementary Information S1.

Robustness of Individual Nodes

It is desirable that any method for community detection be relatively robust to small changes in network connectivity. Modularity may be used to assess the quality of a partition on a global level at a particular resolution, but not the robustness of a individual node. The assignment of node i to a particular community may be fragile (non-robust) if it (a) has few edges within its assigned community (i.e. small $k_i^{in}=\sum_{j\in c_i}a_{ij}$) or (b) has a small ratio of in-community and out-of-community edges (i.e. small $k_i^{in}/(k_i-k_i^{in})=k_i^{in}/k_i^{out}$). It is useful to incorporate both of these elements into a single measure, which we call the degree of robustness: $d_i^{(1)}$ is the number of the k_i^{in} nodes to which i feels closest that are in i's microcommunity. Nodes with high robustness can be considered the 'core' of their community, since of all of the nodes in the community they have the largest number of close friends amongst the other community members. In networks with a hierarchical community structure, nodes may have varying robustness at each resolution. Nodes that are robustly assigned to a microcommunity may have a fragile assignment to its macro-

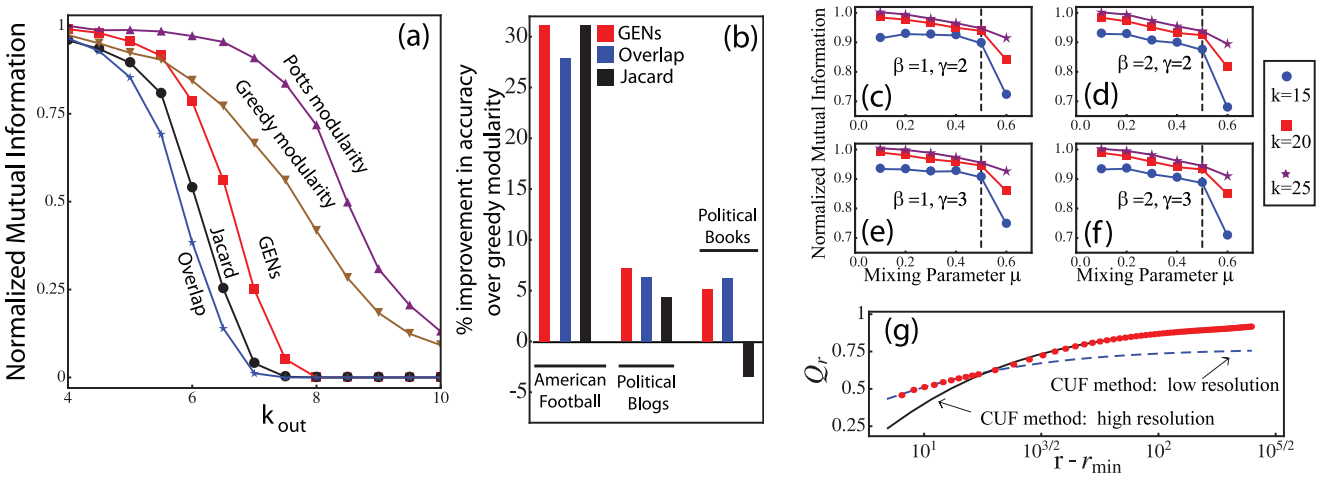


Figure 1. Benchmarks of the community detection algorithm. (a) shows the mutual information between the detected and true partitions for varying k^{out} and for different closeness measures on the Girvan-Newman benchmark [1,12]. Up and down triangles show modularity maximization using a greedy [16] (implemented in Mathematica) and Potts model [14,32] for comparison with the CUF method implemented using the Jacard Coefficients (black circles), GENs (red squares) and overlap (blue stars) as closeness measures. (b) Percent improvement of the CUF approach over a greedy modularity maximization [16] using the GENs (red), overlap (blue), and Jacard Coefficients (black) as a closeness measure for real world networks with a 'correct' partition known apriori. Taken together, (a) and (b) suggest the GENs are typically more accurate measure of closeness. (c-f) show the CUF method implemented on the benchmark of Lancichinetti, Fortunato and Radicchi for varying k, β and γ (compare to Fig. 5 and 7 of Ref. [28]). The CUF method performs well for $\mu \leq 0.5$, although modularity maximization is more accurate (as is the case in (a)), and beings to fail significantly for $\mu > 0.5$ as expected. (g) shows the multiresolution modularity [18] Q_r of the high (solid black line) and low (dashed blue line) resolution partitions using our CUF algorithm, alongside the maximum modularity determined via simulated annealing. The modularity maximizing solutions transition smoothly between the coarser partition for small r and the finer partition for larger r as expected, indicating that our CUF method does indeed detect the two levels of hierarchy accurately without appealing to arbitrary parameters.
doi:10.1371/journal.pone.0038704.g001

community, and vice versa. To assess the robustness at each level of the hierarchy, we can compute $D_i^{(j)} = d_i^{(j)} - d_i^{(j-1)}$, where $d_i^{(j)}$ is the robustness of a node i at the j^{th} resolution in the hierarchy, setting $d_i^{(0)} = 0$ for notational convenience so that $D_i^{(1)} = d_i^{(1)}$. Nodes with small $D_i^{(j)}$ are weakly connected to the other nodes in their community (i.e. their assignment to the micro- or macro-community is fragile, regardless of the robustness in communities of other resolutions). Note that the normalized degree of robustness $D_i^{(j)}/k_i$ is useful in detecting nodes on the boundary between communities (having many edges, but few close friends in their assigned community), but that $D_i^{(j)}$ more directly indicates robustness as the number of strong in-community edges. At each level of resolution, the average robustness of any community can be estimated as $r_c^{(j)} = \langle D_i^{(j)} \rangle_{i \in c} = n_c^{-1} \sum_{i \in c} D_i^{(j)}$.

An Artificial Benchmark with Variable Robustness

In order to introduce variable node robustness into an artificial benchmark, we modify the benchmark of Reichardt and Bornholdt [14,20] (similar to that in Fig. 1(g)) which includes 512 nodes, 16 microcommunities of 32 nodes, and 4 macro-communities of 128 nodes (see Supplementary Information S1 for more details). Each node i has on average k_i^{in} edges connecting it to its microcommunity, $k_i^{out} + k_i^{in}$ edges in its macrocommunity, and k_i^{mix} edges outside of its macrocommunity. In order to modify the benchmark to allow for variable node robustness, we choose k_i^{in}, k_i^{out}, and k_i^{mix} to depend on i in a simple fashion, depending on the macrocommunity it is assigned to (labelled A–D in Fig. 2(a)) and an asymmetry parameter $\alpha \geq 0$, with $\alpha = 0$ corresponding to the standard Reichardt-Bornholdt benchmark [14] (see the table in the caption of Fig. 2 and discussion in Supplementary Information S1). This modified benchmark allows us to examine the effectiveness of the multi-level hierarchical community detection as well as the utility of the degree of robustness $D_i^{(j)}$.

An example of the benchmark is shown explicitly in Fig. 2(a) for $\alpha = 8$, for which the in-, out-, and mix-degrees of nodes vary significantly with i (see the caption of Fig. 2). Fig. 2(b-c) show the in-degrees and in-out ratios for the highest resolution of the hierarchy and (e-f) for the coarsest resolution, with a decrease in k_i^{in} implying a node is less connected to its community and a decrease in $\rho_i^{(1)} = k_i^{in}/(k_i^{out} + k_i^{mix})$ indicating a node is highly connected to nodes outside of its community. When we apply our community detection algorithm, the CUF approach recovers the correct partition with a mutual information of $\langle I_{micro} \rangle = 0.95$ on the micro-scale and $\langle I_{macro} \rangle = 0.85$ on the macro-scale (see eq. 2) at $\alpha = 8$. The mutual information at each scale increases for for decreasing α, but begins to drop rapidly near $\alpha \gtrsim 10$. The high value of the mutual information shows that the CUF algorithm accurately detects the intended communities for reasonably large asymmetry in the community structure (see Supplementary Information S1 for further hierarchical benchmarking).

The benchmark shows that the degree of robustness $D_i^{(j)}$ accurately determines nodes that are less robustly assigned to their intended community at both levels of resolution (shown in Fig. 2(d) and (g)). Nodes in macrocommunity A are less connected to the network overall (and are less robustly assigned at all scales), with and unsurprisingly both $D_i^{(1)}$ and $D_i^{(2)}$ are decreasing with $i^* = [(i-1) \bmod 32]/31$ as expected. In macrocommunity B, nodes have a constant in-community degree and a decreasing ratio of in- to out-of-community degree at each scale, so nodes should

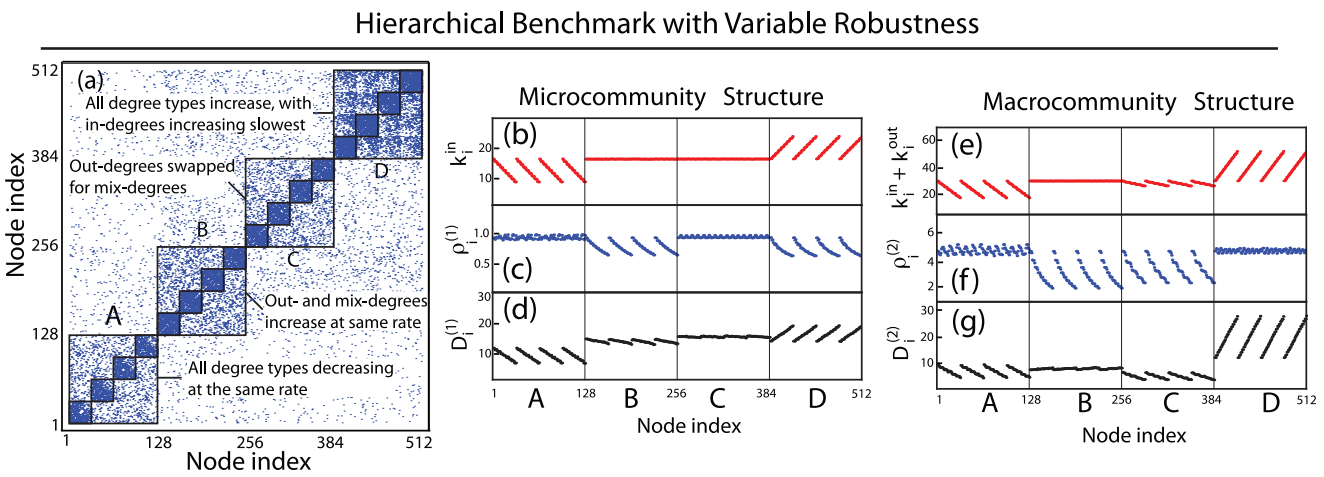


Figure 2. Benchmarks with variable node robustness. (a) A snapshot of the benchmark with hierarchical community structure and variable node robustness at $\alpha=8$. The behavior of the nodes as a function of α and $i^*=[(i-1) \bmod 32]/31$ is described in the table, with $\rho_i^{(1)}=k_i^{in}/(k_i^{out}+k_i^{mix})$ the average in-out ratio at the microcommunity resolution, and $\rho_i^{(2)}=(k_i^{in}+k_i^{out})/k_i^{mix}$ is the in-out ratio at the macrocommunity resolution. In the table, down arrows, up arrows, and dashes denote increasing, decreasing, and constant values (respectively) of the quantities on average. (b) and (e) show the in-degrees at each resolution, k_i^{in} for microcommunities and $k_i^{in}+k_i^{out}$ for macrocommunities. Likewise, (c) and (f) show the ratio of in- and out-degrees at each resolution, $\rho_i^{(1)}$ and $\rho_i^{(2)}$. (d) shows the degrees of robustness $D_i^{(1)}$ at the micro-scale and (g) shows the robustness $D_i^{(2)}$ on the macro-scale. The behavior of the degrees of robustness at both resolutions agrees with the expectations in most cases: if the in-degrees or in- to out-degrees decrease, the nodes become less robust.
doi:10.1371/journal.pone.0038704.g002

Table 1.

Macrocom.	k_i^{in}	k_i^{out}	k_i^{mix}	k_i^{in}	$\rho_i^{(1)}$	Behavior	$k_i^{in}+k_i^{out}$	$\rho_i^{(2)}$	Behavior
A	$k_0^{in}-\alpha i^*$	$k_0^{out}(1-\alpha i^*/k_0^{in})$	$k_0^{mix}(1-\alpha i^*/k_0^{in})$	$\downarrow$	–	Less robust	$\downarrow$	–	Less robust
B	k_0^{in}	$k_0^{out}+\alpha i^*/2$	$k_0^{mix}+\alpha i^*/2$	–	$\downarrow$	Less robust	–	$\downarrow$	Less robust
C	k_0^{in}	$k_0^{out}-\alpha i^*/2$	$k_0^{mix}+\alpha i^*/2$	–	–	Constant	$\downarrow$	$\downarrow$	Less robust
D	$k_0^{in}+\alpha i^*$	$k_0^{out}+2\alpha i^*$	$k_0^{mix}+3\alpha i^*/\rho_i^{(2)}$	$\uparrow$	$\downarrow$	More robust†	$\uparrow$	–	More robust

†The robustness with increasing i^* depends on how slowly k_i^{in} increases.
doi:10.1371/journal.pone.0038704.t001

be less robust with increasing i^*. While the expected decrease in robustness is clearly observed for $D_i^{(1)}$, at the macro-scale there is a slight (but unexpected) increase in the robustness of each node as i^* increases. This is due to errors in the macro-scale community detection, with macrocommunity *B* being the most difficult to detect of all of them. Nodes in macrocommunity *C* have constant in-degree and in-out ratio at the micro-scale (with the corresponding robustness $D_i^{(1)}$ nearly constant), but at the macro-scale are less robust with both the in-degree and in-out ratio decreasing (leading to an expected decrease in $D_i^{(2)}$ with increasing i^*). Finally, the nodes on the micro-scale in macrocommunity D simultaneously have increasing in-degree but decreasing in-out ratio with increasing i^*. While we find the degree of robustness $D_i^{(1)}$ increasing, the rate of increase of $D_i^{(1)}$ depends on the interplay between the increased robustness due to more in-community edges and the decreased robustness due to more out-of-community edges. $D_i^{(1)}$ in macrocommunity B and D and $D_s^{(2)}$ in macrocommunity D are both clear examples of the dependence of the rate of increase in $D_s^{(j)}$ on both k^{in} and $\rho^{(j)}$. The successes in correctly determining not only the hierarchical community structure but also node robustness of this simple benchmark suggest that our approach may be fruitfully applied to complex real world networks with hierarchical structure.

Results and Discussion

The Harvard Coauthorship Network

Turning now to real examples, we look at the network of scientific journals which we expect can be divided into sub-fields at varying resolutions. We construct a network from publications found in the Digital Access to Scholarship at Harvard (DASH) repository, a database of journals, book chapters, and conference proceedings uploaded by Harvard faculty. The available metadata includes the authors and the journal of publication, which we use to generate a weighted network with each journal as a node. The weight of the edge between nodes i and j, w_{ij}, is the number of article pairs that have at least one author in common, with one article published in journal i and the other in journal j. The largest connected component of this network (comprising $N=779$ journals as nodes, shown in Fig. 3(a)) has a complex structure: while the degree of each node (the number of edges with non-zero weight) is exponentially distributed, $P(k_i=k)\sim e^{-k/15.1}$, the strength of each node is log-normally distributed, with a good fit given by $P(W_i=W)\sim W^{-1}e^{-0.24[\log(W)-5.3]^2}$ (see Fig. 3(b-c)). It is

interesting to note that an exponentially distributed degree sequence is indicative of network growth *without* preferential attachment [30], while log-normally distributed strengths may indicate growth with a localized preferential attachment in the weight (see ref. [31] and below for further discussion). This may illuminate some of the details of how a publication network grows: while authors preferentially publish in high-profile journals or proceedings (leading to the fat tail on the strength distribution), they may choose to publish in new or lower profile journals if necessary (leading to the exponential, non-preferential attachment distribution of the degree sequence).

In Fig. 3(a), 36 microcommunities in the DASH network are found, and in most cases an inspection of the group memberships showed the members of each community were related (a full list is found in Supplementary Information S1). It is worth noting that using a Potts model approach to modularity maximization [14,32] (with resolution $\gamma = 1$) yields 32 distinct microcommunities, and the partitions generated by the two methods share much in common, suggesting the CUF results are reasonable. The hierarchical detection scheme shows that each of the microcommunities falls into 6 natural macrocommunities (see Fig. 3(a)). The two largest macrocommunities show a division between the Physical Sciences (physics, biology, chemistry, and geology) and the Mathematical Sciences (pure mathematics, economics, and computer science). Three additional macrocommunities consist of a combination of Philosophy and the History of Science, Linguistics, and Law, and a final macrocommunity having no obvious meaning on inspection (see Supplementary Information S1 for the member journals of each community). We note that this hierarchical partition is not easily detected using the Potts modularity maximization approach: even for $\gamma = 0.02$, there are still 23 microcommunities detected via modularity maximization. Thus, the partition into distinct scientific fields naturally arises from the coarse graining in our approach, but is difficult to detect using modularity methods alone. Further coarse graining shows that there is no additional hierarchical structure to be found in the DASH network.

The average robustness of the nodes in each community of the DASH data is very heterogeneous (the multi-colored bars in Fig. 3(d)), which can be of use in determining which microcommunities are held together weakly, either because of the complex network topology involving the nodes in the community or due to an incorrect partitioning of the network. Many of the detected communities have few nodes, and are correspondingly less robust on average. Even some large communities have low average robustness, which could indicate an incorrect assignment or an unexpected network topology around a community. For example, Phys. Sci. 5 (PS5 in Fig. 3(d)) consists of 26 journals, with a very small average degree of robustness of $r^{(1)}_{PS5} = 2.8$. The surprisingly low robustness of PS5 is not due to sparse connections between nodes within the community (the average degree of nodes in PS5, $\langle k_i^{in} \rangle = 7.6$), but is because of the fact that these journals are highly connected externally ($\langle k^{out} \rangle = 5.5$).

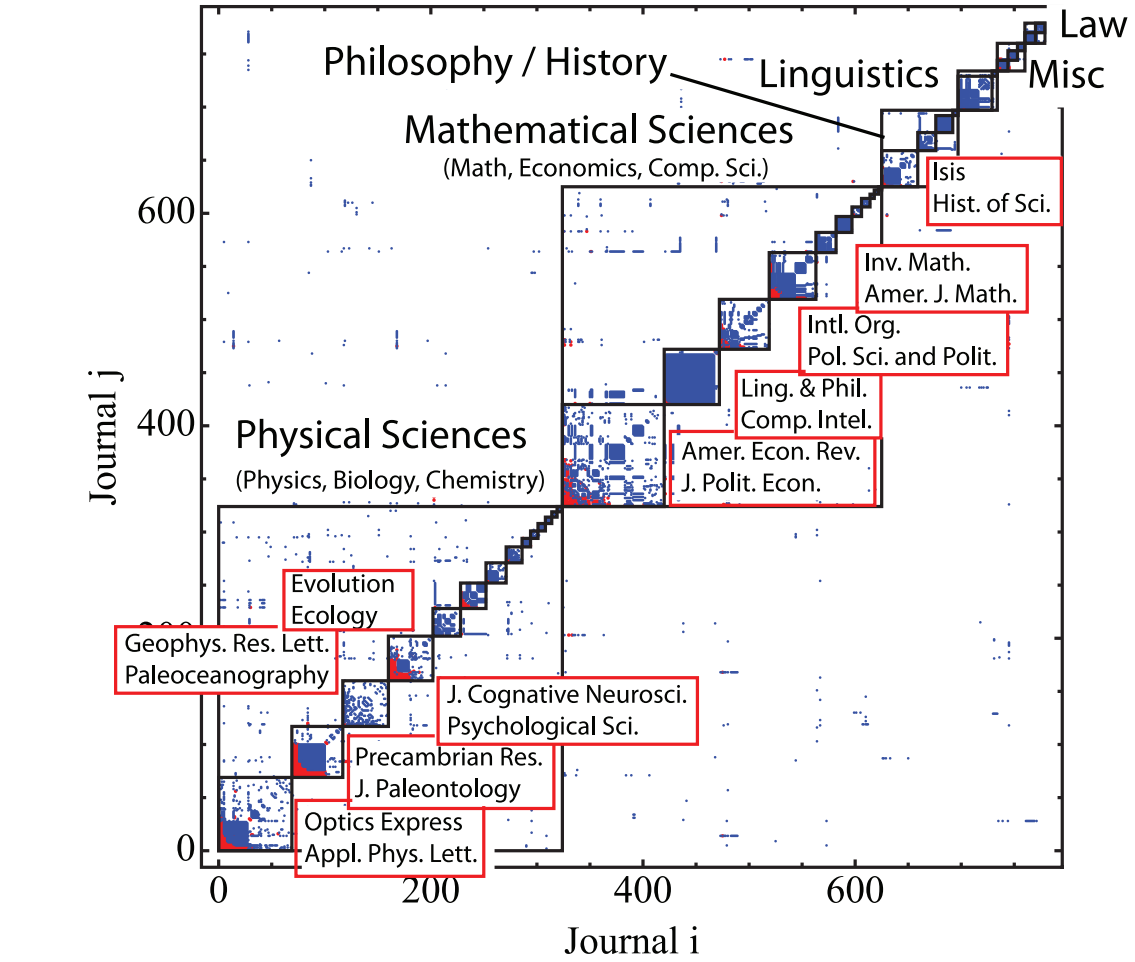
The robustness of a node's assignment to its macrocommunity (the thin black bars in Fig. 3(d)) is not determined by how robustly assigned it is to its microcommunity. The average robustness $r^{(2)}_c$ gives an indication of how strongly a microcommunity is attached to its macrocommunity, and we find that Philosophy/History 1 (PH1) is the most weakly assigned, with $r^{(2)}_{PH1} \approx 0.12$, despite the very robust assignment of the nodes in the microcommunity ($r^{(1)}_{PH1} = 9.8$). Two journals in PH1 are very strongly connected to the Mathematical Sciences macrocommunity (so much weight is directed to Math. Sci. from PS1), while many journals in PH1 are more weakly connected to the journals in its own macrocommunity (so more edges are directed towards Philosophy and History). The degree of robustness is thus able to home in on microcommunities that may be on the boundary between macrocommunities and identifying particularly complex topologies.

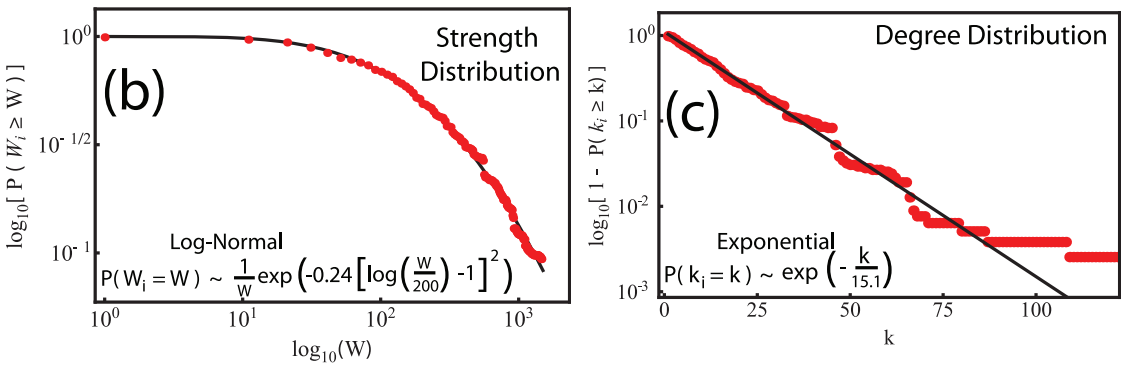
The *Physical Review* Citation Network

Another real-world network where one may expect a hierarchical structure is that of a citation network (independent of their journal of publication), with an expectation of divisions between fields and sub-fields as was observed in the DASH network. We examine the citation network of articles published in the *Physical Review* journals [33,31], with articles as nodes and citations between articles as edges. Citations naturally form directed edges (a citation between i and j does not imply a citation between j and i), but to apply our methods we study the undirected ($w_{ij} = w_{ji}$) version. The degree distribution of this network has been previously shown to be log-normally distributed [31], which may indicate the underlying dynamics of the growth of the network. Network growth coupled with with preferential attachment produces a scale free degree distribution [30,7], but Redner [33] has noted that a modified, locally defined preferential attachment process explains the emergence of a log-normally distributed data. Rather than citing the most important papers, an author chooses to cite either a randomly chosen paper or one of the citations of that paper (with the latter likely to be highly cited [34]). The log-normal distribution is also observed in the highly-cited subset of the network considered (see below for further discussion), suggesting that this smaller sample is reasonably representative of the structure of the full network.

Applying the CUF method to the *Physical Review* network detects four distinct hierarchies of community structure, ranging from the finest resolution of numerous small microcommunities to the coarsest resolution with two large macrocommunities (see Fig. 4(a-c) for a schematic ranging from coarsest to finest). At the highest resolution, 266 communities are detected, and the partition has the modularity $Q_1 = 0.63$ (at $\gamma = 1$). This is in reasonable agreement with a similar previously studied *Phys. Rev.* network [33] with 274 detected communities and a modularity of $Q = 0.54$, suggesting that this fine resolution partition of the more current data is reasonable. High-modularity partitions are also detected using our coarse graining method, with the modularities $Q_2 = 0.75$ for the 62 communities on the second level of the hierarchy and $Q_3 = 0.74$ for the 11 communities at the third level (see Fig. 4(a-b)). The final level of coarse graining does not produce a very high modularity (with $Q_4 = 0.33$) for two macrocommunities, but the meaning of the partition recognizable on inspection of the component communities for its distinction between earth-bound and cosmological research. At each level of hierarchy, the partitioning is both reasonable from a scientific perspective as well as generally producing a large modularity, suggesting that CUF approach is able to discern the natural partitions of the network without need for a resolution parameter.

The distribution of the degrees of robustness found in the *Physical Review* network is shown in Fig. 4(d), along side the degree distribution of the nodes. As mentioned earlier, the degree distribution is well fit by a log-normal distribution [31] $P(k_i = k) \sim k^{-1} e^{-1.1[\log(k)-2]^2}$, with a fatter tail than exponential but vanishing faster than a power law. The distribution of node robustness $D_i^{(j)}$, which indicates how robustly the node i is assigned at the j^{th} level of the hierarchy, decays much more rapidly for large $D_i^{(j)}$ for all four of the hierarchical levels. At the finest resolution (blue squares in Fig. 4(d)), the degrees of robustness are well fit by an exponential decay $P(D_i^{(1)} = D) \sim e^{-D/4.5}$, and although the tail


Law
Philosophy / History
Linguistics
Misc
Mathematical Sciences
(Math, Economics, Comp. Sci.)
Isis
Hist. of Sci.
Inv. Math.
Amer. J. Math.
Intl. Org.
Pol. Sci. and Polit.
Ling. & Phil.
Comp. Intel.
Amer. Econ. Rev.
J. Polit. Econ.
Physical Sciences
(Physics, Biology, Chemistry)
Evolution
Ecology
Geophys. Res. Lett.
Paleoceanography
J. Cognative Neurosci.
Psychological Sci.
Precambrian Res.
J. Paleontology
Optics Express
Appl. Phys. Lett.
Journal j
Journal i
0
200
400
600


(b)
Strength
Distribution
Log-Normal
$P(W_i = W) \sim \frac{1}{W} \exp\left(-0.24\left[\log\left(\frac{W}{200}\right) - 1\right]^2\right)$
$\log_{10}[P(W_i \geq W)]$
$\log_{10}(W)$
(c)
Degree Distribution
Exponential
$P(k_i = k) \sim \exp\left(-\frac{k}{15.1}\right)$
$\log_{10}[1 - P(k_i \geq k)]$
k

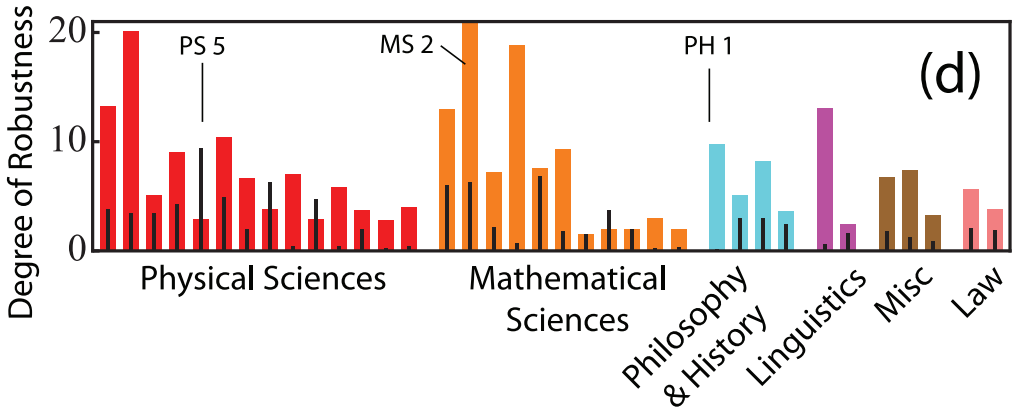

(d)
PS 5
MS 2
PH 1
Degree of Robustness
Physical Sciences
Mathematical Sciences
Philosophy & History
Linguistics
Misc
Law

Figure 3. The network of journals from the DASH data. (a) Low weight edges (with $1 \le w_{ij} \le 5$) are shown in blue, while higher weight nodes ($w_{ij} \ge 6$) are shown in red. Nodes are ordered in order of descending macrocommunity size, then descending microcommunity size, and finally in descending strength. The 36 microcommunities are denoted by the smaller black squares, while the 6 macrocommunities are shown in the larger thick black squares. Some microcommunities are labelled with their two most robust nodes (having largest $D_i^{(1)}$). The degree distribution of the DASH data in (b) is exponential, while the distribution of node strengths in (c) appears to be log-normal. In (d), the average robustness of nodes in the microcommunities ($r_c^{(1)}$, thick bars of varying color) and macrocommunities ($r_c^{(2)}$, thin black bars) for the DASH data. In (d), the bar for Mathematical Sciences 2 (MS2) is cut off, having a very high average degree of robustness of $r_{MS2}^{(1)} = 39.8$.
doi:10.1371/journal.pone.0038704.g003

beyond $D = 20$ (incorporating below 2.5% of the nodes) is slower than exponential, it remains faster than log-normal. The far more rapid decay of the degrees of robustness suggest that highly-cited papers have applications in a wide variety of fields (i.e. are have many out-of-community edges). The robustness of the nodes at the lower-resolution partitions are all similar to one another (triangles and stars in Fig. 4(d)), all satisfying an exponential initial decay of $P(D_i^{(j)} = D) \sim e^{-D/2.8}$ over a somewhat shorter range. Each node has roughly the same robustness on each level of the hierarchy, suggesting that an equal fraction of nodes are involved in forming the edges of the different levels of the hierarchies.

Conclusions

In this paper, we have described a new and intuitive method for detecting hierarchical community structure in complex networks that does not rely on free parameters or require advanced knowledge of the number or size of the communities. Given a method for measuring the 'closeness' between two nodes in a network, one can trace a path of closest friendship that defines a high-resolution partition of the community, resulting in a method with (1) reasonable computational complexity in comparison to other methods [10], (2) easy detection of multiple levels of community structure without the need for an (unknown apriori) resolution parameter [17,13], and (3) a simple yet powerful method of measuring the robustness of the assignment of an individual node to its community. We must note that there are also limitations to our approach, including the free choice of a closeness measure, pathological network topologies (which, for example, necessitates the use of the CUF over the CF; see Supplementary Information S1), and the requirement that no community can be formed from only one node. Despite these possible limitations, the advantages of our approach in automatically detecting and evaluating hierarchical community structure are significant. Using the recently proposed Generalized Erdös Numbers [24] as a closeness measure (which performs better than other measures in benchmarks) we examined two real world systems where a hierarchical community structure is naturally expected: a

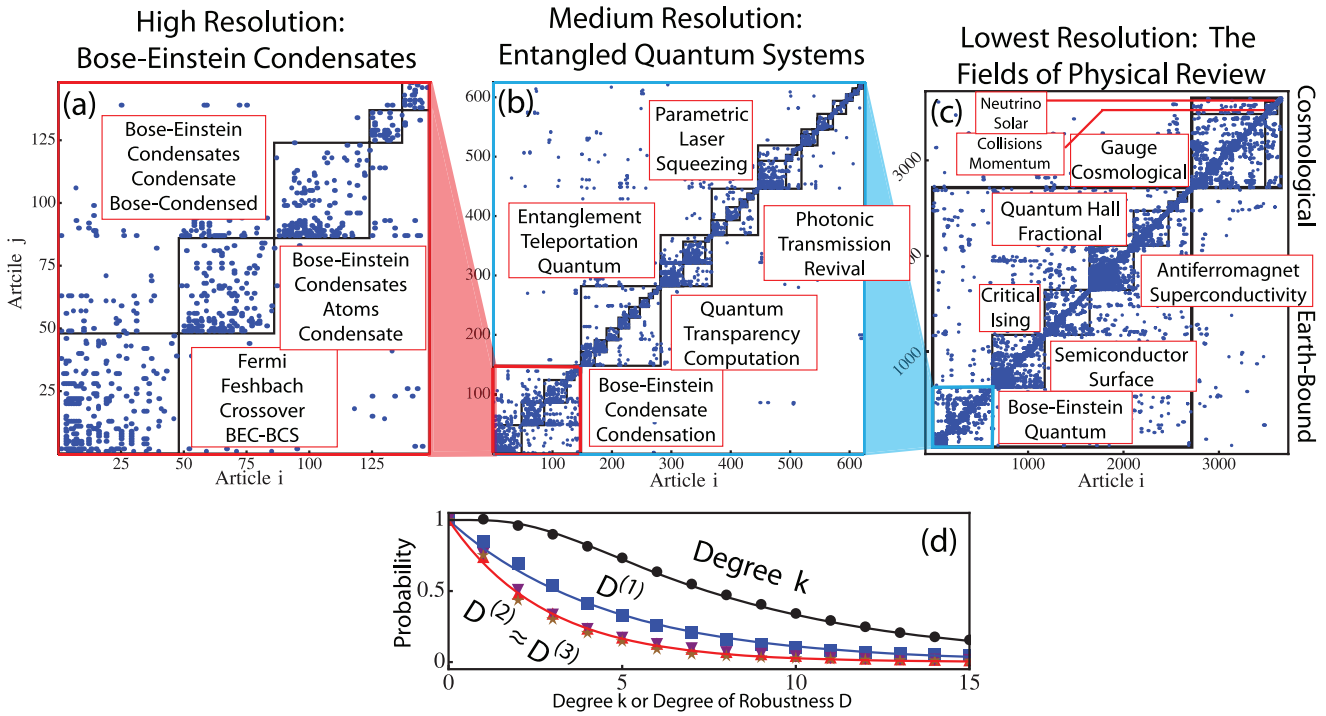


Figure 4. The hierarchical community structure of the *Physical Review* network. (a-c) shows a progressively coarsened view of the network, with the text labels of the communities composed of the most statistically significant words found in the titles of the articles in the communities. (a) shows the microcommunity structure of 148 nodes, with (b) a zoomed-out picture of the 625 nodes in one macrocommunity of the second level of the hierarchy, and (c) the full network (showing the final two levels of hierarchy). (d) shows the degree distribution as well as the distribution of node robustness at each level of the hierarchy (shown log-linear in the inset). Black circles show the degree distribution, which is log-normally distributed [31] (the best fit is the black line). The distribution of robustness on the micro-scale, $D_i^{(1)}$, is shown with the blue squares, while the distribution for the other hierarchical degrees of robustness $D_i^{(j)}$ are all quite similar (shown with the up triangles, down triangles, and stars). The initial decay of the robustness is well-fit by an exponential in all cases (with the best fit for each shown as lines).
doi:10.1371/journal.pone.0038704.g004

coauthorship network defined by the DASH data and a citation network generated from the *Physical Review* data. Our approach is able to detect a high-resolution partition of each dataset that is composed of well defined communities of variable size, and an inspection of the member nodes suggests that the partition is meaningful in both the DASH- and *Phys. Rev.* networks. Our coarse graining method of detecting hierarchy finds a reasonable macrocommunity partition for the DASH data (with each of the macrocommunities clearly linked upon inspection), with this coarse-grained partition not obviously detected using modularity maximization. By examining the degree of robustness of these communities on the micro- and macro-scale, we are able to rapidly home in on the most interdisciplinary communities (those with many significant connections to other communities). The *Phys. Rev.* citation network naturally partitions into four distinct hierarchies of communities (without any apriori assumption of the correct number of hierarchies), with the nodes in the communities generally related to each other upon inspection. The ability to find communities of arbitrary size, detect the structure of a natural (and system-defined) number of hierarchies, and locate particularly insular or interdisciplinary communities are all significant advantages of our method, and clearly displayed in the analysis of both the DASH and *Phys. Rev.* networks.

Supporting Information

Supplementary Information S1
(PDF)

Acknowledgments

We would like to thank Reinhard Engels for providing us with a easily processed copy of the DASH data, Levi Dudte for many useful conversations on the methods and paper, and the Wyss Institute for Biologically Inspired Engineering at Harvard.

Author Contributions

Conceived and designed the experiments: GM LM. Performed the experiments: GM. Analyzed the data: GM. Contributed reagents/materials/analysis tools: GM LM. Wrote the paper: GM LM.

References

1. Girvan M, Newman M (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99: 7821.
2. Bilke S, Peterson C (2001) Topological properties of citation and metabolic networks. Physical Review E 64: 36106.
3. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Reviews of modern physics 81: 591–646.
4. Barabási A, Jeong H, Néda Z, Ravasz E, Schubert A, et al. (2002) Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications 311: 590–614.
5. Porter M, Mucha P, Newman M, Friend A (2007) Community structure in the united states house of representatives. Physica A: Statistical Mechanics and its Applications 386: 414–438.
6. Yan K, Fang G, Bhardwaj N, Alexander R, Gerstein M (2010) Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. Science's STKE 107: 9186.
7. Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286: 509.
8. Albert R, Jeong H, Barabasi A (2000) Error and attack tolerance of complex networks. Nature 406: 378–382.
9. Moore C, Newman M (2000) Epidemics and percolation in small-world networks. Physical Review E 61: 5678–5682.
10. Fortunato S (2010) Community detection in graphs. Physics Reports 486: 75–174.
11. Newman M (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103: 8577.
12. Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Physical Review E 69: 26113.
13. Kumpula J, Saramäki J, Kaski K, Kertesz J (2007) Limited resolution in complex network community detection with potts model approach. The European Physical Journal B 56: 41–45.
14. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Physical Review E 74: 16110.
15. Newman M (2004) Fast algorithm for detecting community structure in networks. Physical Review E 69: 066133.
16. Clauset A (2005) Finding local community structure in networks. Phys Rev E 72: 026132.
17. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proceedings of the National Academy of Sciences 104: 36.
18. Arenas A, Fernandez A, Gomez S (2008) Analysis of the structure of complex networks at different resolution levels. New Journal of Physics 10: 053039.
19. Wu F, Huberman B (2004) Finding communities in linear time: a physics approach. The European Physical Journal B-Condensed Matter and Complex Systems 38: 331–338.
20. Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics 11: 033015.
21. Karrer B, Newman M (2011) Stochastic blockmodels and community structure in networks. Phys Rev E 83: 016107.
22. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS One 6: e18961.
23. Liben-Nowell D, Kleinberg J (2007) The link prediction problem for social networks. Journal of the American Society for Information Science and Technology 58: 1019–1031.
24. Morrison G, Mahadevan L (2011) Asymmetric network connectivity using weighted harmonic averages. Europhys Lett 93: 40002.
25. Guimera R, Sales-Pardo M, Amaral LAN (2008) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70: 025101.
26. Adamic LA, Glance N (2005) The political blogosphere and the 2004 us election. Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem.
27. Krebs V. Network data website. Available: http://www-personal.umich.edu/~mejn/netdata/, maintained by M. E. J. Newman. Accessed 2012 Jun 1.
28. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78: 46110.
29. Sales-Pardo M, Guimera R, Moreira AA, Amaral LAN (2007) Extracting the hierarchical organization of complex systems. Proc Natl Acad Sci 104: 15224.
30. Barabasi AL, Albert R, Jeong H (1999) Mean-field theory for scale-free networks. Physica A 272: 173.
31. Redner S (2005) Citation statistics from 110 years of physical review. Physics Today 58: 49.
32. Csrdi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems : 1695.
33. Chen P, Redner S (2010) Community structure of the physical review citation network. J Infometrics 4: 278.
34. Feld SL (1991) Why your friends have more friends than you do. Amer J Sociol 96: 1464.