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 ($N\to\infty$), but the typical strength W_i of all nodes remains finite, the total strength $W\to\infty$ 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. $w_{ij}\to w_{ij}+r\delta_{ij}$) 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. $|\mathbf{p}^i\bigcap\mathbf{p}^j|\neq 0$) 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 $(E^c_{hg})^{-1} = \sum_{i\in g, j\in h} E_{ij}^{-1}/n_g n_h$, 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-

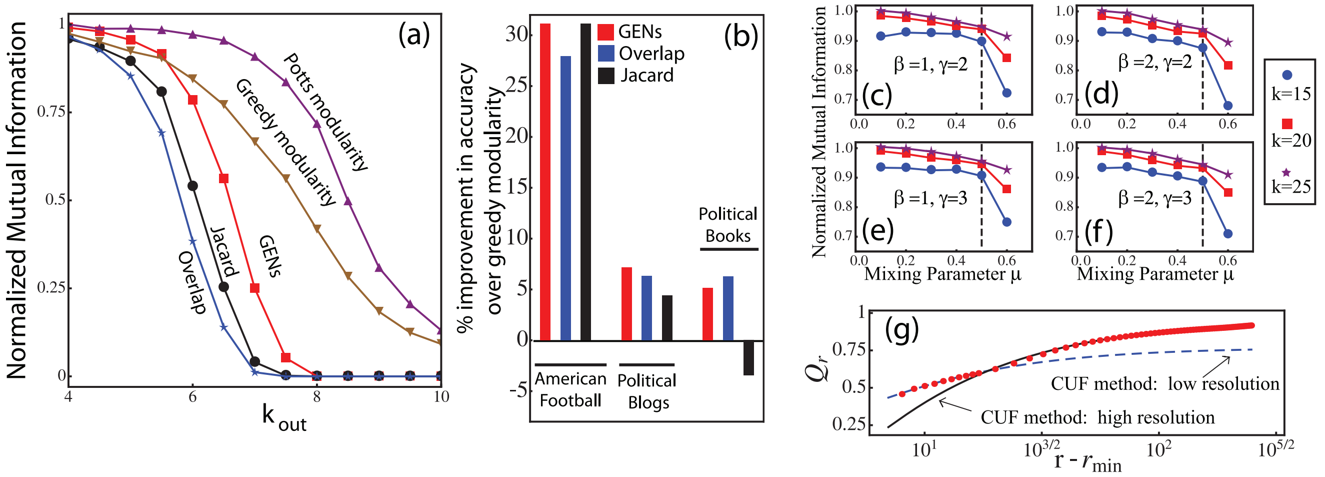

Figure 1. Benchmarks of the community detection algorithm. (a) shows the mutual information between the detected and true partitions for varying k^{out} and for different closeness measures on the Girvan-Newman benchmark [1,12]. Up and down triangles show modularity maximization using a greedy [16] (implemented 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

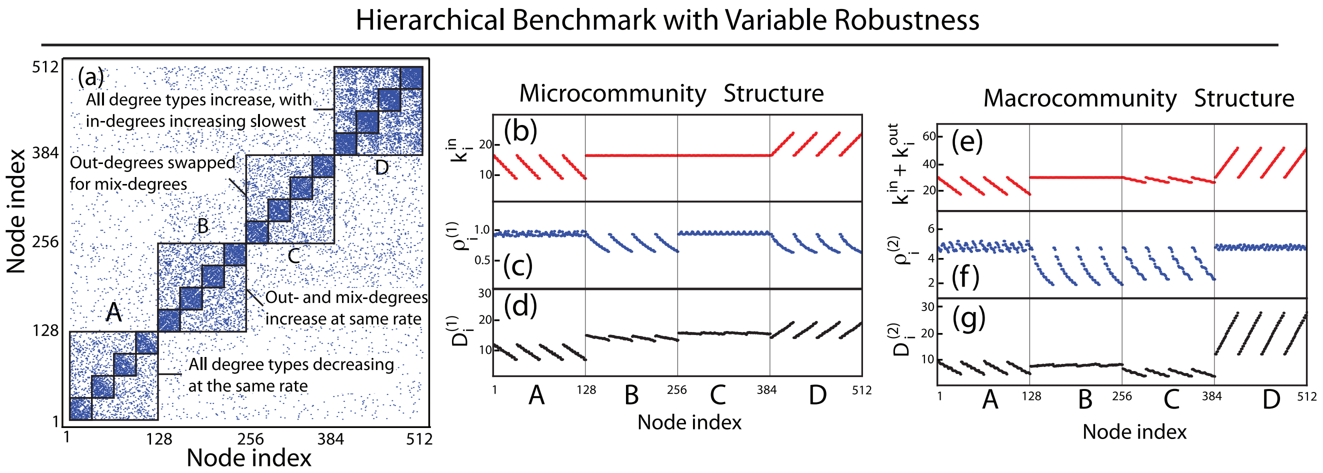

Figure 2. Benchmarks with variable node robustness. (a) A snapshot of the benchmark with hierarchical community structure and variable node robustness at $\alpha = 8$. The behavior of the nodes as a function of α and $i^* = [(i-1) \bmod 32]/31$ is described in the table, with $\rho_i^{(1)} = k_i^{in}/(k_i^{out} + k_i^{mix})$ 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	$k_0^{in} - \alpha i^*$	$k_0^{out}(1 - \alpha i^*/k_0^{in})$	$k_0^{mix}(1 - \alpha i^*/k_0^{in})$	↓	–	Less robust	↓	–	Less robust
B	k_0^{in}	$k_0^{out} + \alpha i^*/2$	$k_0^{mix} + \alpha i^*/2$	–	↓	Less robust	–	↓	Less robust
C	k_0^{in}	$k_0^{out} - \alpha i^*/2$	$k_0^{mix} + \alpha i^*/2$	–	–	Constant	↓	↓	Less robust
D	$k_0^{in} + \alpha i^*$	$k_0^{out} + 2\alpha i^*$	$k_0^{mix} + 3\alpha i^*/\rho_i^{(2)}$	↑	↓	More robust†	↑	–	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 $r_c^{(2)}$ 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

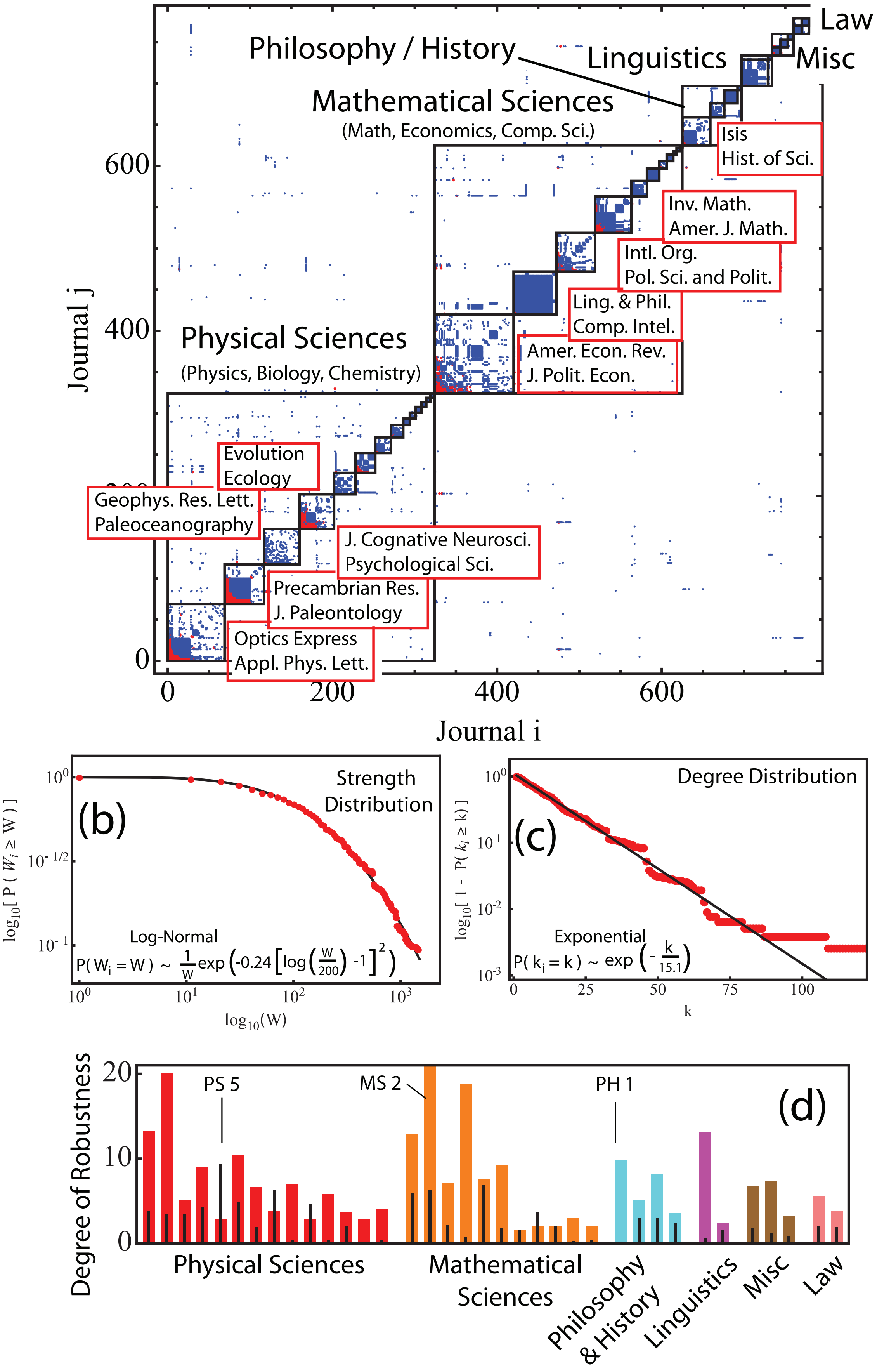
Law
Misc
Philosophy / History
Linguistics
Mathematical Sciences
(Math, Economics, Comp. Sci.)
Isis
Hist. of Sci.
Inv. Math.
Amer. J. Math.
Intl. Org.
Pol. Sci. and Polit.
Ling. & Phil.
Comp. Intel.
Amer. Econ. Rev.
J. Polit. Econ.
Physical Sciences
(Physics, Biology, Chemistry)
Evolution
Ecology
Geophys. Res. Lett.
Paleoceanography
J. Cognative Neurosci.
Psychological Sci.
Precambrian Res.
J. Paleontology
Optics Express
Appl. Phys. Lett.
600
400
0
Journal j
0
200
400
600
Journal i
Strength
Distribution
(b)
10^0
$10^{-1/2}$
10^{-1}
$\log_{10}[\,P\,(\,W_i \geq W\,)\,]$
Log-Normal
$P(W_i = W) \sim \frac{1}{W}\exp\left(-0.24\left[\log\left(\frac{W}{200}\right)-1\right]^2\right)$
10^0
10^1
10^2
10^3
$\log_{10}(W)$
Degree Distribution
(c)
10^0
10^{-1}
10^{-2}
10^{-3}
$\log_{10}[\,1 - P(\,k_i \geq k)\,]$
Exponential
$P(k_i = k) \sim \exp\left(-\frac{k}{15.1}\right)$
0
25
50
75
100
k
Degree of Robustness
20
10
0
PS 5
MS 2
PH 1
(d)
Physical Sciences
Mathematical Sciences
Philosophy & History
Linguistics
Misc
Law

Figure 3. The network of journals from the DASH data. (a) Low weight edges (with $1 \leq w_{ij} \leq 5$) are shown in blue, while higher weight nodes ($w_{ij} \geq 6$) 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

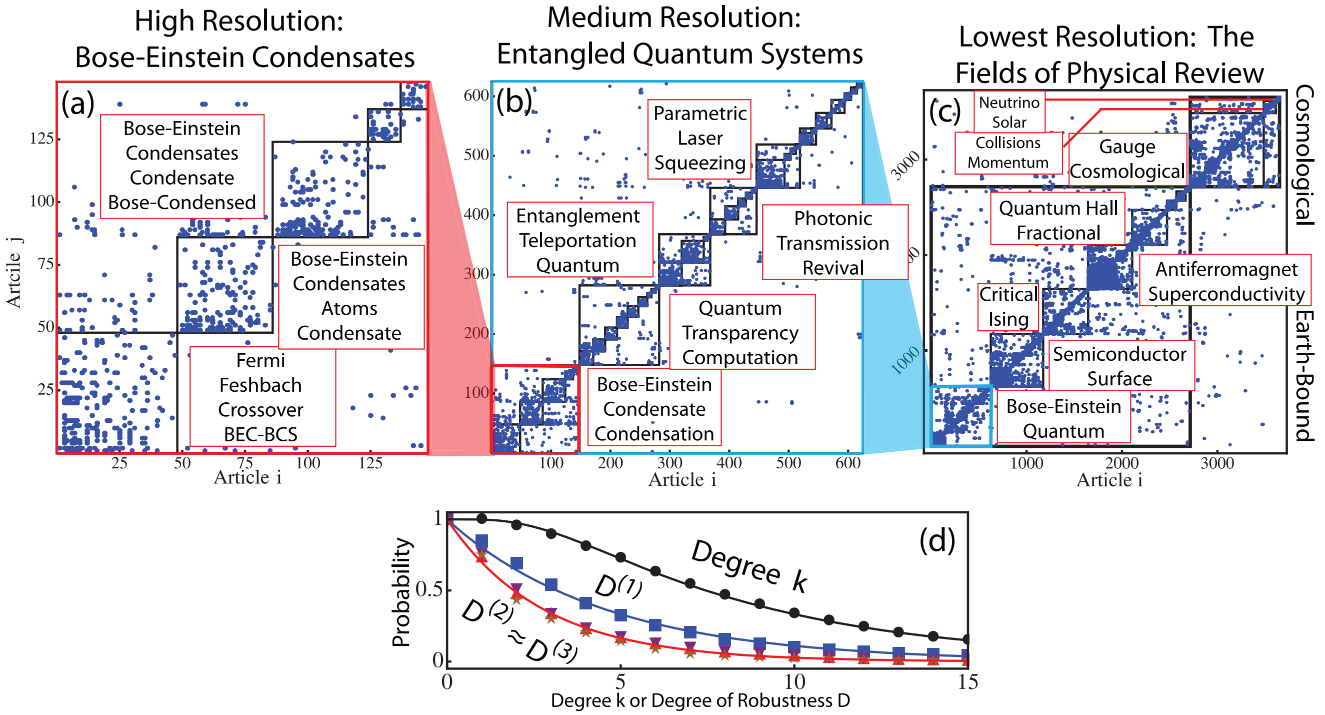

Figure 4. The hierarchical community structure of the *Physical Review* network. (a-c) shows a progressively coarsened view of the network, with the text labels of the communities composed of the most statistically significant words found in the titles of the articles in the communities. (a) shows the 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, Guirmera R, Moreira AA, Amaral LAN (2007) Extracting the hiearchical 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.

Supplementary Information

Greg Morrison and L. Mahadevan

September 30, 2012

1 Closeness Measures and Community Detection

A schematic diagram of a network with easily-detected community structure is shown in Fig. 1(a). In this network, a pair of communities with $|c| = N/2$ nodes each is connected by exactly one edge (between α and β). For any reasonable measure of closeness, a node will feel closer to other nodes within its community rather than those in a different community, with the Generalized Erdös numbers (GENs), Jacard Coefficients (JCs), and overlap explicitly demonstrated as having this property. Resistance Distance, the mean first passage time between nodes, and the Adar / Adamic coefficient[1] all behave in a similar manner (not shown). There is a clear separation between in-community and out-of-community closenesses for the network in Fig. 1(a), which can be used to determine the correct community structure. Each node i is constrained to be in the same community as their closest friend, $f(i)$. This is similar in spirit to the resistance-distance approach of Wu and Huberman[2], but does not require an arbitrary threshold for defining communities. Each measure of closeness will behave differently when fuzzy communities are detected, with some outperforming others in the ability to detect communities (as discussed in the main text). We note that no nodes in a network can be in a community by themselves using this approach, since all connected nodes necessarily have a closest friend. It may be possible to remove this restriction by introducing self-loops into the network, but we leave this to later work.

In Fig. 1(a), it is important to note that it is not possible to continuously tune an arbitrary parameter to find different partitions. Using modularity maximization with resolution parameter γ as an example, at $\gamma = 1$ we expect to detect the correct partition of two communities. For $\gamma \to 0$, we expect to find only a single community, including all nodes in the network. *No* reasonable measure of closeness will ever produce this coarsest partition of Fig. 1(a), since it would require a node in c_1 to feel closer to nodes it has *fewer* connections to than those it has many connections to. Regardless of the closeness measure chosen (and even if a tunable free parameter included in our the measure), a single community *can not* be detected using the CF approach so long as nodes feel closer to their neighbors than their non-neighbors. The coarser partition of a single community is, however, readily detected using the hierarchical approach described in the text.

In Fig. 1(b), we show a pathological network topology for which the CF method will fail: two distinct communities with each node connected to a single, central node (δ). Most measures of closeness will find all nodes feel closest to δ (and all reasonable measures will, so long as the intra-community edges are sufficiently sparse and the network sufficiently large), so the CF approach will assign all nodes to the same community as δ. In such a case, only a single community will be (incorrectly) detected. This can be avoided by searching for the closest unpopular friend: after sorting nodes into ascending order of how close node i feels to them, the closest node with degree less than or equal to the next-closest is selected as $f(i)$. Note that for the closest unpopular friend algorithm on a weighted graph, we still search for lower degree k_i rather than strength W_i, which avoids nodes that are connected to many other nodes ($k_i \gg 1$) but not nodes that are strongly connected to a few nodes ($W_i \gg 1$). The node δ is assigned to the community its closest unpopular friend is in, which will depend on the details of the network topology and the choice of closeness measure.

2 Fractured Communities

While the CF and CUF algorithms provides a intuitive method for detecting communities in an arbitrary graph, it is possible for a correct community to be fractured into two or more parts due to the local variability of the density of edges. As pictured in Fig. 2), an intended community A may be split into two groups, A_1 and A_2 due to the asymmetric connections each of these sub-communities has to the communities B and C. As pictured, there are either more edges leading from A_1 to B than there are between A_1 and C or the total weight between A_1 and B is larger than between A_1 and C. While useful information could be found in the structure of the fractured communities, is also desirable to recover the 'correct' communities despite these local variations. In order to produce a more useful method for community detection, we must supplement both of these approaches with an algorithm to merge fractured communities to better recover the 'correct' partition.

The fracture of communities can be due to two aspects of the detection: first, the decisions are purely local (even if the closeness measure incorporates the global topology). Because the decisions are not made with a global quality function, the splitting of a community into two pieces is not penalized. Second, the random nature of the networks allows for variability in the local density of edges. These fluctuations in density will affect all community detection methods[3], and in some cases may call into question the 'correctness' of the intended partition.

In Fig. 2, the detected groups A_1 and A_2 (which are in the same community in the 'correct' partition) will likely have a large number of edges between them. If we imagine a single community is mistakenly broken into two sub-communities of the same size (n nodes apiece), the number of edges between the sub-communities should scale as n^2 (with a uniform density of edges in the correct community A). This allows us to build a relatively simple greedy search for communities to merge. Once a CF or CUF partition has been determined, we perform a search for the pair of communities g and h with the largest value of $k_{g \to h}/\max(n_g, n_h)^2$, where $k_{g \to h}$ is the number of edges leading from group g to group h and n_g is the number of nodes in group g. Before we merge the communities g and h, we check to ensure that $k_{g \to h} \geq \min(k_{g \to g}, k_{g \to h})$. If the inequality is not satisfied (i.e. there are fewer edges between g and h are in either g or h alone) the greedy search is halted, otherwise g and h are merged and the search repeats.

In Fig. 3, we use the GENs to compare the CF and CUF methods for with or without community merging (averaged over 100 realizations of the network). We see that the CUF approach with merging gives the best overall results, with the largest normalized mutual information[4, 5] (as defined in eq. 2 of the main text) for all values of k^{out}, with only a moderate improvement over the CF approach. However, the CUF approach is more prone to community fracture (where the black circles do not converge to $I = 1$ as $k^{out} \to 0$), and greedy merging is therefore essential for reliable reconstruction of the network. We note that the merging of fractured communities as implemented here could also be used with modularity maximizing methods and may improve the spurious splitting of communities in some cases.

As noted in Fig. 2, variability in the local density of edges can lead to fractured partitions. The propensity of modularity maximization for finding spurious sub-communities can perhaps be most clearly seen by considering a random network *without* any community structure. We generate networks with the probability of an edge between any nodes is p_{edge}, with $0.04 \leq p_{edge} \leq 1$. In Fig. 4, we show the number of communities detected in these networks using both greedy modularity maximization (squares) and the CUF approach (circles). The CUF performs far better than the greedy modularity maximization for $p_{edge} \gtrsim 0.1$, while modularity maximization consistently finds more than one community for all $p_{edge} < 1$. For small p_{edge}, we expect fluctuations in the edges will produce a locally higher density of edges randomly, which may be detected using any community detection method[6]. However, as p_{edge} increases, these fluctuations should be less significant, and the CUF that detects only a single community may be preferable.

3 A Common Hierarchical Benchmark

It is natural to define a coarse grained network formed with the communities in the higher-resolution partition acting as nodes in the new, lower-resolution network in order to detect a hierarchy of community structure in the network. However, it is not immediately obvious how to choose the new edges in the new network, and rather than attempt to define coarse-grained edges at this resolution, we take the closeness between the coarse-grained nodes to be the average closeness between communities in the high-resolution partition. In the particular case of the GENs, the harmonic mean is the appropriate way to average the closeness, as the closeness between communities should be dominated by the nodes in each that feel close to one another (small E_{ij}, thus more significant in the harmonic mean), rather than the nodes that do not feel close to one another (large E_{ij}, less significant in the harmonic mean). For other closeness measures, it a linear mean may be the more reasonable choice for averaging the closeness between communities.

It is worth re-emphasizing that we do not expect to be able to continuously tune the resolution of the coarse-grained network with a free parameter. Our ability to detect hierarchical structure of course depends on the accuracy of the higher resolution partition (with an inaccurate partition unlikely to accurately detect the correct macrocommunities), and the existence of a 'correct' hierarchical partition. If nodes in communities g and h are very close to one another in the original network, a reasonable method of averaging should ensure they are close in the coarse grained network. While the detected partition will weakly depend on the method of coarse graining, it is not possible to tune the averaging as it is in the case of modularity maximization (or other approaches), where choosing a resolution $\gamma \ll 1$ will assign all coarse-grained nodes to the same community.

We apply our coarse graining approach to detect the community structure of the benchmark presented by Rechardt and Bornholdt [7] depicted in Fig. 5. The network of $N = 512$ nodes is composed of 16 microcommunities with on average $k^{in} = 16$ edges internally per node. Four of these microcommunities form a macrocommunity, with on average k^{out} edges per node within a macrocommunity and k^{mix} edges per node between macrocommunities. Note that this is the benchmark that is modified in order to produce the benchmark of variable robustness as described in the main text. The mutual information between the correct and detected partitions of the micro-communities (using the CUF approach with the GENs as the closeness measure) is shown in Fig. 6(a) for varying k^{out} and k^{mix}. The microcommunities are detected accurately for small k^{out}, with the transition from 'good' to 'bad' detection occurring for $k^{out} + k^{mix} \approx 34$ (the point at which $I = 0.5$, averaged over the four curves shown), more than twice the value of $k^{in} = 16$. It is worth noting that for the larger values of k^{out} (12 or 14), often the failure to saturate to $I = 1$ at $k^{mix} = 0$ is due to the fact that the method will fail to detect the microcommunity structure of 16 communities, but rather the macrocommunity structure of 4 macrocommunities. For sufficiently dense connections within the macrocommunities, the CUF method does fail to detect the finest resolution of the network. However, so long as the microcommunities are accurately detected, the macrocommunity structure is also correctly determined (as shown in Fig. 6(b)). For $k^{out} = 16$, we generally fail to find the macrocommunity structure because of the poor detection of the fine-resolution structure, while for $k^{out} = 12$ or 14, the macrocommunity structure is not reliably found as $k^{mix} \to 0$. Modularity-based methods or other approaches may outperform these results[7, 8] if the correct (but a priori unknown) resolution parameter is chosen. However, our approach gives a single partition for each scale (both micro- and macro-), and performs very well so long as the micro-communities are not too fuzzy (k^{out} is sufficiently small), without using an unknown parameter.

4 Common Real-World Community Benchmarks

Modularity maximization performs quite well on the artificial GN benchmark precisely because of the modular structure inherent in the test: the correct solution was also the modularity maximizing one. This may not be the case in real world networks, where the 'correct' partition is determined from external information and is independent of the partition's modularity. To see the utility of the CF or CUF methods, we examine three simple real-world benchmarks with an a priori known partition in the main text. The football network[9] is comprised of nodes representing american football teams, with edges denoting games played between them in 2000. The 'correct' partition groups each team within their externally-defined division. The political blogs network[10] is a set of blogs in the leadup to the 2006 US midterm election, with an edge representing a link from one blog to another (we use an undirected version of this network). The political books network[11] is a set of books purchased on amazon.com around the 2004 US presidential elections, with an edge representing a co-purchase of a pair of books. In the political blogs and books networks, the 'correct' partition is the node's apparent political leaning: liberal vs. conservative in the former and liberal, independent, or conservative in the latter. All of these benchmarks are unweighted networks (with $w_{ij} = 0$ or 1).

One common benchmark with a known community structure not mentioned in the main text is Zachary's Karate club[12]. This is a very small network of 34 nodes representing members of a karate club at an unnamed university, with edges denoting the out-of-club interactions between individuals. The club split into two parts due to a disagreement over the club's leadership, and the 'correct' partition denotes which individuals fell on a particular side of the disagreement. The karate club is partitioned using a number of approaches in Fig. 7, with from left to right modularity maximization, CF/CUF using the GENs, using the JCs, and using overlap. For the Karate club benchmark, we find surprisingly that both overlap and the GENs perform extremely poorly while the Jacard coefficients (JCs) perfectly reconstruct the correct partition. However, if the closest friend (CF) approach is used (rather than the closest unpopular friend approach, which avoids high degree nodes when assigning communities and is implemented throughout the main text), the GENs perfectly reconstruct the network, followed by overlap and then by the JCs. This illustrates that pathological networks do indeed exist that have not been fully accounted for in the CUF methodology, and it is difficult to predict exactly which method will be optimal a priori. We also note that a CF partition can be generated rapidly when generating a CUF partition, and by examining a global quality function (such as modularity), one can easily distinguish which partition better represents the structure of the network. Thus, despite the unexpected behavior of our approach when considering the Karate Club network, we determine that (a) the GENs remain a reasonable choice for the closeness measure and (b) that it may be necessary to compare the results of the CUF approach to a global quality function (such as modularity) to determine if the partition is reasonable.

5 Simulated Annealing of the Benchmark

In order to generate the network used in benchmarking the community detection and robustness measure in the main text, we used simulated annealing to produce a network with the desired properties. The desired in-, out-, and mixing-degree of each node were computed: $k_i^{in,0}$, the desired number of edges from node i to nodes in its microcommunity, $k_i^{out,0}$, the desired number of edges leading from i to any node in its macrocommunity (but not in the microcommunity) and $k_i^{mix,0}$, the number of edges leading from i out of its macrocommunity. From these the total number of edges $M = \frac{1}{2}\sum_i (k_i^{in,0} + k_i^{out,0} + k_i^{mix,0})$ was determined, and a network of $N = 512$ nodes was generated having precisely M randomly distributed edges. The network was then randomly rewired, with a new trial configuration generated by removing one edge connecting the randomly chosen i and j, and a new edge being drawn between i and k. This trial configuration was accepted using a metropolis criterion: $p_{acc} = \min(1, e^{-\beta(E_{old}-E_{trial})})$, with the energy of a configuration

$$E = \frac{1}{2}\sum_i \left[(k_i^{in} - k_i^{in,0})^2 + (k_i^{out} - k_i^{out,0})^2 + (k_i^{mix} - k_i^{mix,0})^2 \right] \qquad (1)$$

where the first term of E is minimized if the in-, out-, and mix-degrees of each node satisfy our desired conditions.The temperature parameter β is set to $\beta = 1$ initially, and incrementally increased by 2×10^{-5} at each attempted rewiring. A total of 500,000 rewiring attempts were made, with each edge on average experiencing ≈ 975 attempted rewirings.

6 How Ties are Handled

Unlike many real world networks, the network in Fig. 1(a) is highly symmetric and the closeness between nodes in groups A or B is likewise symmetric so that there is not a unique closest friend. In this case, we must develop a rule for handling ties in the closeness. In the case of a tie, we randomly but consistently select the 'closest' neighbor of i, $f(i)$. This is accomplished by initially randomizing the node index, and choosing the node with the lowest (random) index as closest. In practice, the importance of ties in the artificial or real world networks networks depends on the choice of closeness measure. The Jacard coefficient $\mathrm{JC}_{ij} = |C_i \cap C_j|/|C_i \cup C_j|$ can easily produce ties[13] for complex networks, whereas the GENs require highly symmetric networks to see a tie. The lack of ties is an additional advantage of measures that incorporate the global topology of the network, rather than purely local information.

7 Details of the DASH robustness

The DASH database, downloaded in June 2010 contained $N_0 = 918$ journals and 2404 articles published by 3385 unique author names, not all of which work at Harvard. Because of the interdisciplinary and highly connected nature of the journals *Science*, *Nature*, and *Proc. Natl. Acad. Sci*, these three journals are removed from the network. This alteration does not alter the shape of either the degree or weight distributions (although the removal of edges does affect their particular fitting parameters).

While briefly discussed in the text, it is worthwhile to examine the structure of the DASH network in detail, to determine the power of the degree of robustness in finding complex topologies or incorrectly assigned nodes. When we examine the degrees of robustness observed in the network, nodes with few edges connecting them to their community have a correspondingly low degree of robustness, reflecting the fact that they are only weakly connected to their assigned community. Low values for the degree of robustness $D_i^{(1)}$ for these weakly connected nodes is unsurprising. We can use the degree of robustness to find nodes that are on the boundary between communities (i.e. that are strongly connected both to their assigned community as well as to a different community to which they are not assigned). We find 142 nodes with $D_i^{(1)} \leq 2$, 53% of which have $k_i^{in} \leq 2$ (indicating that they are simply of low degree, rather than on the boundary of a community). However, there are a few nodes that have $D^{(1)} \leq 2$ but are strongly connected to their respective communities (having high degree and weight directed into c_i). Due to their large values of k_i^{in}, these nodes are most likely on the boundary of their respective communities. The five journals with smallest $D_1^{(i)}/k_i$ with $D_i^{(1)} = 1$ or 2 are shown in Table 1. Some of these journals have a $k_i^{in} \ll k_i$ (so many edges lead from i to different communities), while others have $k_i^{in} \approx k_i$ (so most of the edges from i are within its assigned community).

Examining the topology of the DASH network connected to these nodes that are boundary-like shows two distinct causes of high in-degree and low degree of robustness. *Cognition*, the second journal in Table 1 has more than twice as many out-edges as in-edges, but these out-edges are distributed amongst a wide range of communities. In Table 1, *Cognition* has the most weight ($W_i^{in} = 14$) directed towards its community (Phys. Sci. 4, primarily focused on Oceanography and Atmospheric Science), but has a large weight of 12 directed towards the Phys. Sci. 3 community (focused primarily on Psychology and Neuroscience, a more natural choice of community assignment for *Cognition*). It is likely that this node was incorrectly assigned, but the fact that the highest weight points towards Phys. Sci. 4 makes the misassignment understandable. The degree of robustness has allowed us to locate this possible error with ease, while the in-degree ($k_i^{in} = 8$), total degree, the ratio of in- to total degrees ($k_i^{in}/k_i = 0.32$, and is the 17^{th} worst of all journals), or the ratio of in- to total strengths ($W_i^{in}/W_i = 0.34$, the 10^{th} worst of all journals) would not highlight *Cognition* as a particularly troublesome node.

The other journals in Table 1 all have a low degree of robustness for a different reason. For these, the largest number of edges point towards their assigned communities, and in all but one case (the *Journal of Economic History*) the largest weights are also pointed towards their respective communities. However, in each case the journal is connected to the 'core' of a different community: nodes in a different community with both high in-weights or in-degrees and high robustness. While the assignment of each node in Table 1 to its respective community is often reasonable (since the majority of edges are within its assigned community), each of these nodes is also connected to one or more nodes that effectively define a neighboring community. These journals act as a bridge between the (generally less robust) communities to which they are assigned and the core of a robust, strongly connected community.

It is also of interest to determine the quality of the assignment of each microcommunity to its macrocommunity. The thin black lines in Fig. 2 of the main text denote the macrocommunity robustness $r_c^{(2)} = \langle D_i^{(2)} \rangle_{i \in c}$ of each assignment. We note that a robust microcommunity (with high $r_c = \langle D_i^{(1)} \rangle_{i \in c}$) does not necessarily imply a robust assignment to its macrocommunity, and that many well formed microcommunities have a very low value of $r_c^{(2)}$. Table 2 shows that the lowest values of $r_c^{(2)}$ typically occur for communities that have relatively few out-edges (and thus their assignment to their macrocommunity is expected to be fragile). However, the assignment of the Philosophy and History 1 (PH1) microcommunity to its macrocommunity is surprising, as it has a very low ratio of in- to out-degree and in- to out-strength. While the placement of PH1 to the Philosophy and History macrocommunity may appear to be an error, the surprising assignment is due to the fact that 75% of the out-of-macrocommunity edges and 84% of the out-of-macrocommunity weight are due to only two journals: the strong connections that *Social Studies of Science* and *Annual Review of Sociology* have towards Mathematical Sciences 3 (also focused on the Social Sciences). There are three journals in PH1 that are connected to the Philosophy and History macrocommunity, *Isis*, *Persepectives on Science*, and *Journal of the History of Ideas.* Two of these journals are in the 'core' of PH 1 (with $D^{(1)} = 17$), while only one of the journals strongly connected to Math. Sci. is in the core (with $D^{(1)} = 16$). Thus, the assignment of PH1 to the Philosophy and History macrocommunity is due to the fact that, while more weight is directed out of the assigned macrocommunity, the core journals of PH1 are more strongly connected to Philosophy and History journals. PH 1 is clearly boundary-like, and our robustness measure of $r_c^{(2)}$ accurately detects this fragile assignment.

8 Additional details of the Phys. Rev. Network

The Physical Review network included over 462,000 articles published in any Physical Review journal up to July 2010. Due to the size of the network , we consider only the subset of articles that have garnered at least 100 citations, with the largest connected component including 3651 articles and over 16,000 edges. While the network is unweighted (one citation is neither stronger nor weaker than another, thus $w_{ij} = 0$ or 1) and directed (article i cites article j, but not vice-versa), we consider the non-directed version (with $w_{ij} = w_{ji}$=0 or 1). The community structure at one resolution of the Phys. Rev. network up to 2007 has previously been determined[14]. The detected communities are similar in many respects to the community structure we have detected, although these other papers did not report an examine of any additional hierarchical structure, as we discuss in the main text.

References

[1] Adamic L, Adar E (2001) Friends and neighbors on the web. Social Net 25: 211-230.

[2] 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.

[3] Sun Y, Danila B, Josic K, Bassler KE (2009) Improved community structure detection using a modified fine-tuning strategy. Europhys Lett 86: 28004.

[4] Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78: 46110.

[5] Guimerà R, Sales-Pardo M, Amaral L (2007) Module identification in bipartite and directed networks. Physical Review E 76: 36102.

[6] Guimera R, Sales-Pardo M, Amaral LAN (2008) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70: 025101.

[7] Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Physical Review E 74: 16110.

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

[9] 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.

[10] Adamic LA, Glance N (2005) The political blogosphere and the 2004 us election. Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem .

[11] Krebs V. http://www-personal.umich.edu/ mejn/netdata/, maintained by M. E. J. Newman.

[12] Zachary W (1977) An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33: 452–473.

[13] Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466: 761–764.

[14] Chen P, Redner S (2010) Community structure of the physical review citation network. J Infometrics 4: 278.

Tables

Name	Community	$D_i^{(1)}$	k_i^{in}	k_i	W_i^{in}	W_i
J. Econ. Hist.	Math. Sci. 3	2	16	24	71	157
Cognition	Phys. Sci. 4	1	8	25	14	41
Ecol. Appl.	Phys. Sci. 5	1	8	10	18	22
Oikos*	Phys. Sci. 5	1	8	10	18	22
Brit. Med. J.	Math. Sci. 6	2	14	15	29	32

*Oikos is an ecology journal published by the Nordic Ecol. Soc., and has exactly the same connections as Ecol. Appl.

Table 1: The five most boundary-like nodes (with the lowest non-zero values of $D_i^{(1)}/k_i^{in}$). The first, *J. Econ. Hist.*, has a high degree and strength and large k^{in} and k^{out}. Similarly, *Cognition* has the smallest ratio of in-edges to total node degree, and is connected to a large number of other communities. The last three elements in the table are surprising in that they have a few connections outside of their communities ($k_i^{out} = 1$ or 2 compared to $k^{in} = 8$ or 10) but still have low degrees of robustness. This is because while they have many in-community connections, their few out-of-community connections lead to strong, central nodes in other communities. These boundary-like nodes would not be easily detected by simply looking at the in-degrees or in-out ratio.

Community	Focus	$r_c^{(2)}$	k_c^{in}	W_c^{in}	k_c^{out}	W_c^{out}
Phil. & Hist. 1	History	0.12	3	7	28	45
Phys. Sci. 13	Info. Theory	0.2	1	1	0	0
Math. Sci. 10	Drug Addiction	0.25	1	1	0	0
Math. Sci. 11	Crystallography	0.33	1	1	0	0
Phys. Sci. 9	High En. Physics	0.38	3	6	0	0

Table 2: The five least robust macrocommunity assignments. k_c^{in} and W_c^{in} denote the total number of edges and total weight from the microcommunity to other microcommunities in its macrocommunity respectively, while k_c^{out} and W_c^{out} denote the total number and weight of edges into any other macrocommunity. Philosophy and History 1 (PH 1) is the worst, and lies on the boundary of the Philosophy and History macrocommunity and the Mathematical Sciences macrocommunity. The other macrocommunity assignments are very fragile do to the very small number of connections, and are peripheral microcommunities.

Figure Captions

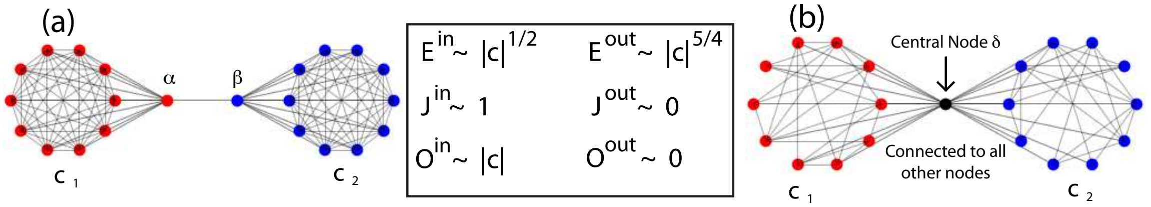

Figure 1: Detecting communities with the CF and CUF methods. (a) An example of two clearly defined communities (c_1 and c_2), each of size $N/2$ with exactly one edge connecting them. Any plausible measure of closeness based on the network topology will clearly distinguish between intra- and inter-community connections. The closeness between nodes within the community as measured by the GENs (E^{in}), JCs (J^{in}), and overlap (O^{in}), with $|c|$ the number of nodes in each community. Likewise, the closeness between nodes in different communities is shown with the superscript 'out'. (b) A schematic network of a single highly connected node (δ) to which all nodes in the network will feel closest. Assigning each node to the same community as their closest friend (the CF approach) will assign all nodes to the same community as δ, thus detecting only one community. By avoiding high-degree nodes (the CUF approach), the two communities are correctly detected, with δ assigned to one or the other.

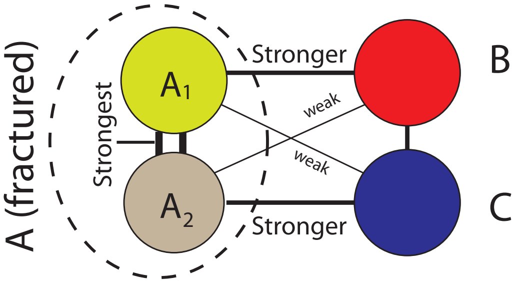

Figure 2: Merging of fractured communities. Community A is fractured into two communities, A_1 and A_2 due to the fact that A_1 is more strongly connected to B (connections labelled 'stronger') than to C (connections labelled 'weak'), while community A_2 is more strongly connected to C than B. In this coarse-grained schematic, 'stronger' may represent either high weight or many edges between them. Because A_1 and A_2 are truly subsets of the same community in the 'correct' partition, we expect a large number of edges between them.

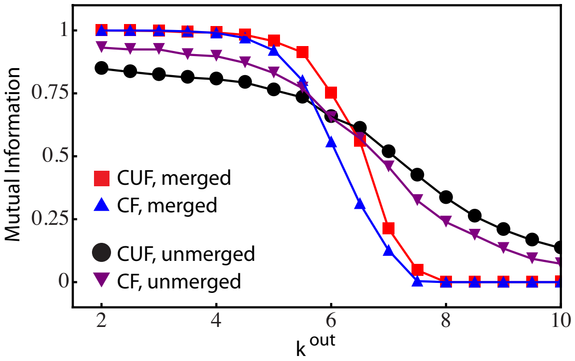

Figure 3: Improvements in the methods using fracture merging. A comparison of the approaches for community detection using the GENs, using the Newman-Girvan benchmark. Red squares denote the CUF after community merging, which gives the best overall results. Black circles denotes the result of the CUF without merging, and has a low mutual information to the expected partition due to fracture (even for clear communities, with low k^{mix}). The blue up and purple down triangles are the results for the CF algorithm with and without fracture correction, respectively.

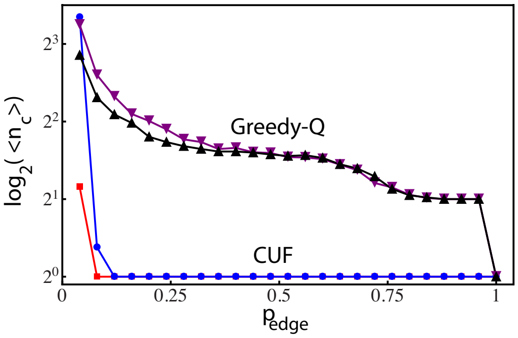

Figure 4: Community detection in unstructured networks. The number of communities n_c detected using greedy modularity maximization (up and down triangles) or the CUF method (squares and circles) for a randomly linked network (with no intended community structure) as a function of the probability of an edge between two nodes, p_{edge}. Greedy modularity maximization is shown in the purple down triangles for $N = 100$ nodes and black up triangles for $N = 200$ nodes, while the blue circles shows CUF detection for $N = 100$ and red squares for CUF with $N = 200$. When there is no intended structure in the network, modularity maximization tends to find a relatively large number of communities, while the CUF method typically finds only one community (for sufficiently large $p_{edge} \gtrsim 0.1$.

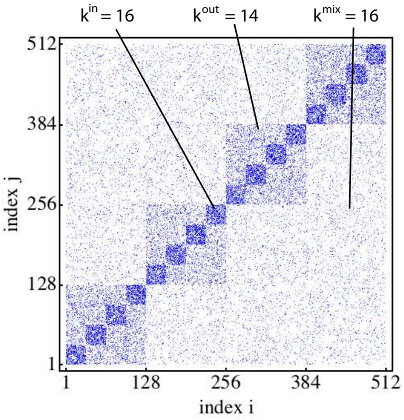

Figure 5: The adjacency matrix of the Reichardt-Bornholdt hierarchical benchmark. Each node is a member of a micro-community of 32 nodes, with $k^{in} = 16$ connections to the other nodes in its micro-community on average. Each micro-community is a member of one of four macro-communities, and each node in a macro-community has k^{out} edges internally on average. Each node has on average k^{mix} edges to nodes outside of their macro-community.

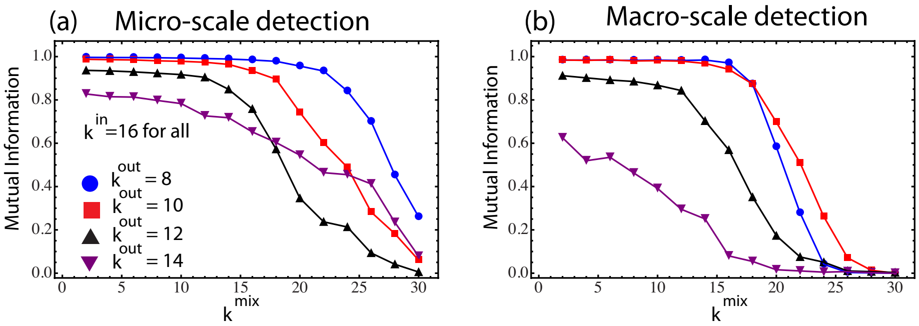

Figure 6: Accuracy of the hierarchical benchmark. The detection of (a) micro- and (b) macro-communities averaged over 100 realizations of the network. For all samples, $k^{in} = 16$ is held fixed. k^{out} is varied as $k^{out} = 8$ (blue circles), 10 (red squares), 12 (black up triangles) and 14 (purple down triangles). The mixing between macrocommunities is varied with $2 \leq k^{mix} \leq 30$. The CUF approach accurately detects the microcommunities over a wide range of values of k^{out} and k^{mix}, and is clearly able to accurately detect the microcommunity structure for sufficiently clear communities. So long as the microcommunity structure is accurately detected, the macrocommunity structure seems reliably determined as well.

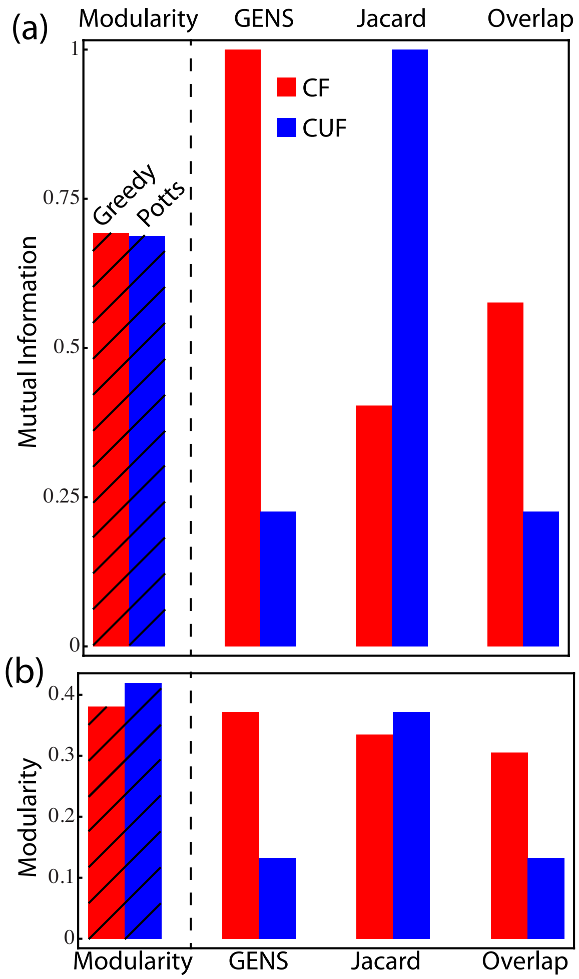

Figure 7: The karate club network. Mutual information (a) and modularity (b) of the partitions of the Karate club[12] detected by a variety of approaches, with the a priori correct partition known. The leftmost results show the results of partitions using the greedy (striped red) and Potts model (striped blue) modularity maximizing partitions. For the remainder, red denotes the closest friend (CF) while blue denotes the closest unpopular friend (CUF) approach, with the GENs, JCs, and overlap shown. Surprisingly, the GENs implemented using the CUF method performs the worst in all respect (in contrast to most other benchmarks where it performs the best). For the Karate club network, the GENs do reconstruct the exact 'correct' partition if the CF method is used.

Community Membership of the DASH Network

Greg Morrison and L. Mahadevan

May 29, 2012

Natural Sciences, Community 1		
Accounts of Chemical Research	ACS Nano	Acta Materialia
Advanced Materials	Animal Behavior	Annals of Applied Statistics
Annals of Statistics	Applied and Environmental Microbiol...	Applied Physics A
Applied Physics Letters	Behavioral Ecology and Sociobiology	BMC Genomics
Chemistry and Biology	Defect and Diffusion Forum	Environmental Science and Technolog...
IEEE Journal of Quantum Electronics	IEEE Journal of Selected Topics in ...	IEEE Photonics Technology Letters
International Journal of Primatolog...	Journal of Applied Physics	Journal of Archeological Science
Journal of Bacteriology	Journal of Chemical Physics	Journal of Computer Aided Materials...
Journal of Crystal Growth	Journal of Dramatic Theory and Crit...	Journal of Materials Science
Journal of Microelectromechanical S...	Journal of Physical Chemistry C	Journal of Physics: Condensed Matte...
Journal of the American Chemical So...	Journal of The Electrochemical Soci...	Journal of Vacuum Science and Techn...
Journal of Vacuum Science & Techno...	Lab on a Chip	Materials Research Society Symposia...
Materials Science and Engineering	Materials Science and Engineering A	Materials Science in Semiconductor ...
Metallurgical and Materials Transac...	Microbiology	Modern Drama
Molecular and Cellular Biology	Molecular Biology and Evolution	MRS Bulletin
Nano Letters	Nanoscale Research Letters	Nanotechnology
Nature Biotechnology	Nature Methods	Nature Physics
New Journal of Physics	Nuclear Instruments and Methods in ...	Nuclear Instruments and Methods in ...
Nucleic Acids Research	Optics and Photonics News	Optics Express
Optics Letters	Physica C	Physical Review A
Physical Review B	Physical Review D	Physical Review Letters
Plant Cell and Environment	Proceedings of SPIE	Progress in Biophysics and Molecula...
Social Text	The Open Inorganic Chemistry Journa...	Trends in Ecology and Evolution

Natural Sciences, Community 2		
ACM SIGCSE Bulletin	American Journal of Botany	American Journal of Science
Annual Review of Earth and Planetar...	Annual Review of Ecology, Evolution...	Annual Review of Microbiology
Astrobiology	Canadian Journal of Earth Sciences	Canadian Journal of Forest Research...
Chemical Geology	Earth and Planetary Science Letters	Elements
Geobiology	Geochemistry Geophysics and Geosyst...	Geochimica et Cosmochimica Acta
Geological Magazine	Geological Society of America Bulle...	Geology
Harvard Papers in Botany	Icarus	International Journal of Plant Scie...
International Workshop on Wearable ...	Journal of Geophysical Research - P...	Journal of Mathematical Biology
Journal of Paleontology	Journal of Petrology	Journal of Plant Growth Regulation
Lethaia	Nature Reviews Neuroscience	New Phytologist
Oceanography	Organic Geochemistry	Origins of Life
Origins of Life and Evolution of th...	Palaios	Paleobiology
Philosophical Transactions of the R...	Physical Review	Physics Today
Phytochemistry	Plant Physiology	Plos One
Precambrian Research	Proceedings of the American Philoso...	Review of Scientific Instruments
Sedimentary Geology	Taxon	The Sciences

Natural Sciences, Community 3		
American Journal of Psychiatry	Archives of Neurology	Biological Psychology
California Law Review	Cognitive Brain Research	Cognitive Neuropsychology
Current Directions in Psychological...	Developmental Biology	European Review of Social Psycholog...
IEEE Transactions on Information Te...	Journal of Adult Development	Journal of Cognitive Neuroscience
Journal of Consulting and Clinical ...	Journal of Experimental Psychology:...	Journal of Experimental Psychology ...
Journal of General Internal Medicin...	Journal of Mathematical Psychology	Journal of Neurophysiology
Journal of Neuroscience	Journal of Personality and Social P...	Journal of Physiology - Paris
Journal of the American Academy of ...	Journals of Gerontology Series B	Language and Cognitive Processes
Memory & Cognition	Mind, Brain, and Education	Molecular Psychiatry
New Ideas in Psychology	Personality and Individual Differen...	Personality and Social Psychology B...
Perspectives on Psychological Scien...	Psychiatry Research	Psychological Science
Psychology and Aging	Psychoneuroendocrinology	Psychonomic Bulletin & Review
Public Opinion Quarterly	Review of General Psychology	Self and Identity
Small Group Research	Social Cognitive and Affective Neur...	Trends in Cognitive Sciences
	Visual Cognition	

Natural Sciences, Community 4

Aerosol Science and Technology	Agricultural and Forest Meteorology	Atmospheric Chemistry and Physics
Atmospheric Environment	Child Development	Cladistics
Climate of the Past	Cognition	Cognitive Psychology
Computers and Geosciences	Cortex	Current Biology
Deep Sea Research Part A. Oceanogra...	Dynamics of Atmospheres and Oceans	Earth Science Reviews
ECS Transactions	Europhysics Letters	General Psychologist
Geophysical Research Letters	Global Biogeochemical Cycles	Global Biogeochemical Sciences
Global Change Biology	Intercultural Pragmatics	Journal de Physique IV
Journal of Climate	Journal of Fluid Mechanics	Journal of Geophysical Research
Journal of Geophysical Research -Al...	Journal of Marine Research	Journal of Physical Oceanography
Journal of the Atmospheric Sciences	Molecular Phyogenetics and Evolutio...	Monthly Weather Review- Usa
Nature Geoscience	Paleoceanography	Philosophy and Literature
Quarterly Journal of the Royal Mete...	Quaternary Science Reviews	Statistics in Medicine
Systematic Biology	The Mental Lexicon	Theoretical and Applied Climatology

Natural Sciences, Community 5

Bioinformatics	Biology Letters	BMC Biochemistry
BMC Biology	BMC Ecology	Breviora
Canadian Journal of Zoology	Ecological Applications	Ecology
Ecology Letters	Evolution	Frontiers in Ecology and the Enviro...
Global Ecology and Biogeography	Herpetologica	Journal of Evolutionary Biology
Journal of Experimental Botany	Journal of Theoretical Biology	Journal of Vertebrate Paleontology
Malaria Journal	Methods in Ecology and Evolution	Molecular Ecology
Oikos	Physical Review Series e	Plos Computational Biology
Proceedings of the Royal Society B		Theoretical Population Biology

Natural Sciences, Community 6

American Journal of Human Biology	American Journal of Physical Anthro...	Cancer Causes and Control
Cancer Epidemiology Biomarkers and ...	Cancer Research	Early Human Development
European Journal of Cancer Preventi...	European Journal of Cancer Suppleme...	Evolution and Development
Evolutionary Biology	Fertility and Sterility	Hormones and Behavior
Human Reproduction	Integrative and Comparative Biology	International Journal of Andrology
Journal of Anatomy	Journal of Experimental Biology	Journal of Human Evoluion
Journal of Morphology	Medicine and Science in Sports and ...	PaleoAnthropology
Schizophrenia Research	Sports Medicine	The Anatomical Record

Natural Sciences, Community 7

- Applied and Preventive Psychology
- Behaviour Research and Therapy
- Biological Psychiatry
- Consciousness and Cognition
- Emotion
- Journal of Anxiety Disorders
- Journal of Consumer Research
- Journal of Experimental Social Psyc...
- Journal of Family Psychology
- Journal of Personality Disorders
- Journal of Psychiatric Research
- Memory
- Neuroimage
- Neuropsychologia
- Psychological Medicine
- Psychopharmacology
- Psychophysiology
- Suicide and Life-Threatening Behavi...
- The American Journal of Psychiatry

Natural Sciences, Community 8

- American Naturalist
- Auk
- Bioscience
- BMC Evolutionary Biology
- Genetical Research
- Genetics
- Genome Biology
- International Journal of Plant Scie...
- Journal of Experimental Zoology Par...
- Philosophy of Science -East Lansing...
- Plant Cell
- Plant Methods
- Quarterly Review of Biology
- Trends in Genetics
- Yeast

Natural Sciences, Community 9

- Advances in Theoretical and Mathema...
- Annals of Physics
- Annual Review of Nuclear and Partic...
- Classical and Quantum Gravity
- Fortschritte der Physik
- General Relativity and Gravitation
- Journal of High Energy Physics
- Nuclear Physics B

Natural Sciences, Community 10

- ACM Transactions on Sensor Networks
- Annual Review of Neuroscience
- Cerebral Cortex -New York- Oxford U...
- Experimental Brain Research
- Neuron
- PLoS Biology
- The Journal of Neuroscience

Natural Sciences, Community 11

- ACM SIGPLAN Notices
- Annual Symposium on Principles of P...
- International Conference on Functio...
- Journal of Functional Programming
- Proceedings of the 25th ACM SIGPLAN...
- Proceedings of the 26th ACM SIGPLAN...
- Proceedings of the ACM SIGPLAN 1996...

Natural Sciences, Community 12

- Antiquity
- Archaeological Papers of the Americ...
- Asian Perspectives
- Backdirt: Annual Review of the Cots...
- Current Anthropology
- Symbols

Natural Sciences, Community 13

- IEEE Journal of Selected Areas in C...
- IEEE Signal Processing Magazine
- IEEE Transactions on Information Th...
- IEEE Transactions on Signal Process...
- IEEE Transactions on Wireless Commu...

Natural Sciences, Community 14

- Cell
- European Journal of Biochemistry
- Molecular Biology of the Cell
- Nature Structural and Molecular Bio...
- Structure

Mathematical Sciences, Community 1

4th Multidisciplinary Workshop on A...	AAAI Fall Symposium on Negotiation ...	AAAI Spring Symposium on Empirical ...
American Economic Journal: Microeco...	American Economic Review	American Journal of Computational L...
American Sociological Review	Annual Review of Economics	Applied Economics Research Bulletin
Artificial Intelligence	Autonomous Agents and Multi-Agent S...	Brookings Papers on Economic Activi...
Canadian Journal of Economics	Carnegie-Rochester Conference Serie...	China Economic Review
Cognitive Systems Research	Computational Management Science	Computers & Operations Research
Contributions in Macroeconomics	Decision Support Systems	Econometrica
Economia mexicana	Economic Journal	Economic Policy
Economics of Transition	European Economic Review	European Finance Review
Explorations in Economic History	Foreign Affairs	Games and Economic Behavior
Handbook of Macroeconomics	Health Services Research	IEEE Infocom
IEEE Intelligent Systems	Information Fusion	International Journal of Game Theor...
Journal of Artificial Intelligence ...	Journal of Business and Economic St...	Journal of Business -Chicago-
Journal of Economic Dynamics and Co...	Journal of Economic Literature	Journal of Economic Perspectives
Journal of Economic Theory	Journal of Empirical Finance	Journal of Epidemiology and Communi...
Journal of Finance	Journal of Financial Economics	Journal of Health Economics
Journal of International Economics	Journal of Law and Economics -Chica...	Journal of Mathematical Economics
Journal of Monetary Economics	Journal of Money, Credit and Bankin...	Journal of Policy Modeling
Journal of Political Economy	Journal of Public Economics	Journal of Public Policy and Market...
Journal of the European Economic As...	Journal of Urban Economics	Management Science
National tax journal	National Tax Journal	NBER Macroeconomics Annual
Nber Working Paper Series	NBER Working Paper Series	New York Review of Books
Operating Systems Review	Proceedings 35th International Conf...	Proceedings of Autonomous Agents an...
Proceedings of the 2006 AAAI Spring...	Proceedings of the 21st Annual Mee...	Proceedings of the Conference on Ap...
Proceedings of the DARPA Workshop o...	Proceedings of the Eighteenth Inter...	Proceedings of the Eighth National ...
Proceedings of the First Internatio...	Proceedings of the First Joint Conf...	Proceedings of the Multi-Agent Sequ...
Proceedings of the Nineteenth Annua...	Proceedings of the Sixth Internatio...	Proceedings of the Workshop on Theo...
Public Choice	Quarterly Journal of Economics	Rand Journal of Economics
Rationality and Society	Review of Economic Dynamics	Review of Economics and Statistics
Review of Economic Studies	Review of Financial Studies	Tax Policy and the Economy
The B.E. Journal of Macroeconomics	The Economic Journal	The Lancet
Theoretical Economics	University of Chicago Law Review	WIDER Research Paper

Mathematical Sciences, Community 2

ACM Transactions on Graphics	Cartographica	Communications of the Association f...
Computational Intelligence	Computational Linguistics	Computing Surveys
Electronic Commerce Research and Ap...	Formal Grammar Conferences	Future Generation Computer Systems
IEEE Computer Graphics and Applicat...	IEEE Transactions on Visualization ...	Information Technology
Journal of Biomedical Informatics	Journal of Heuristics	Journal of Linguistics
Journal of Logic Programming	Journal of Optimization Theory and ...	Journal of Psycholinguistic Researc...
Library Quarterly	Linguistics and Philosophy	Natural Language Engineering
Nos	Proceedings of ACL-08: HLT	Proceedings of Algorithms and Exper...
Proceedings of Graph Drawing	Proceedings of SIGGRAPH	Proceedings of the 10th Annual Symp...
Proceedings of the 10th Internation...	Proceedings of the 11th Conference ...	Proceedings of the 19th Internation...
Proceedings of the 2003 Internation...	Proceedings of the 2005 ACL Worksho...	Proceedings of the 7th Conference o...
Proceedings of the Eighteenth Inter...	Proceedings of the Eighth Internati...	Proceedings of the Eighth Internati...
Proceedings of the Fifth SIGdial Wo...	Proceedings of the First Workshop o...	Proceedings of the Human Language T...
Proceedings of the ICAPS-05 Worksho...	Proceedings of the Intelligent User...	Proceedings of the Ninth Internatio...
Proceedings of the Second TAG Works...	Proceedings of the Seventh Internat...	Proceedings of the Seventh Internat...
Proceedings of the Sixth Internatio...	Proceedings of the Thirteenth Annua...	Proceedings of the Twenty-First Nat...
Proceedings of the Workshop on Synt...	Proceedings of UIST	Transactions on Graphics
	Transactions on Systems, Man and Cy...	

Mathematical Sciences, Community 3

Agricultural History	American Journal of Political Scien...	American Political Science Review
Annals of the American Academy of P...	Annual Review of Psychology	BioSocieties
British Journal of Political Scienc...	Dissent	D-Lib Magazine
Economic Development and Cultural C...	Economics & Politics (Oxford, Engl...	Educational Policy
Genewatch	IMF Staff Papers	Indiana Journal of Global Legal Stu...
International Organization	International Studies Perspectives	Journal of Conflict Resolution
Journal of Economic History	Journal of Labor Economics	Journal of Legal Studies
Journal of Policy History	Journal of Politics	Journal of Social Issues
Journal of Statistical Software	Medical Anthropology	Negotiation Journal
Perspectives on Politics	PLoS Medicine	Political Analysis
Population Health Metrics	PS: Political Science and Politics	Research in Higher Education
Social Justice Research	Social Research	Social Science and Medicine
Social Science History	Social Science Research	Sociological Methods and Research
Statistical Science	Studies in American Political Devel...	The American Sociologist
The Annals of the American Academy ...	The Good Society	The Political Quarterly
World Politics		Yale Journal of International Law

Mathematical Sciences, Community 4

- Acta Mathematica -Stockholm-
- Advances in Mathematics
- American Journal of Mathematics
- Annales Academiae Scientiarum Fenni...
- Annales Scientifiques- Ecole Normal...
- Annals of Mathematics
- Applied and Computational Harmonic ...
- Biological Cybernetics
- Bulletin- American Mathematical Soc...
- Commentarii Mathematici Helvetici
- Communications on Pure and Applied ...
- Discrete and Computational Geometry
- Documenta Mathematica
- Duke Mathematical Journal
- Electronic Journal of Combinatorics
- Experimental mathematics
- Foundations and Trends in Computer ...
- Geometric and Functional Analysis
- Harvard College Mathematics Review
- IEEE Transactions on Biomedical Eng...
- IEEE Transactions on Pattern Analys...
- International Journal of Computer V...
- Inventiones mathematicae
- Journal- American Mathematical Soci...
- Journal fur die Reine und Angewandt...
- Journal of Algebra
- Journal of Number Theory
- Journal of the European Mathematica...
- Journal of the Optical Society of A...
- Journal of Topology
- Manuscripta Mathematica
- Mathematical Research Letters
- Mathematische Annalen
- Nagoya Mathematical Journal
- New York Journal of Mathematics
- Pacific Journal of Mathematics
- Periodica Mathematica Hungarica
- Proceedings- American Mathematical ...
- Proceedings of the Thirty-Fifth Ann...
- Publications Mathematiques de l'Ins...
- The Harvard College Mathematics Rev...
- Topology
- Transactions- American Mathematical...
- Vision Research -Oxford-

Mathematical Sciences, Community 5

- ACM International Conference Procee...
- Annual Symposium on Foundations of ...
- Computational Complexity
- Computer Aided Geometric Design
- Computer Graphics and Applications
- Computer Graphics Forum
- Computers and Graphics
- Eurographics/SIGGRAPH symposium on ...
- Information and Computation
- International Journal of Image and ...
- International Mathematics Research ...
- Journal of Computer & System Scien...
- Lecture Notes in Computer Science
- MM: Proceedings of the seventh ACM ...
- Proceedings of the 15th Internation...
- Proceedings of the Annual ACM Sympo...
- Proceedings of the Canadian Confere...
- SIAM Journal on Computing
- The Visual Computer: International ...

Mathematical Sciences, Community 6

- Argumentation
- ARL: A Bimonthly Report
- BMJ: British Medical Journal
- College & Research Libraries News
- Earlhamite
- Infection Control and Hospital Epid...
- Journal of Biology
- Journal of Law and Education
- Journal of Speculative Philosophy
- Legal Writing: The Journal of the L...
- New England Journal of Medicine
- Newsletter on Teaching Philosophy
- Philosophy and Rhetoric
- SPARC Open Access Newsletter
- St. John's Review

Mathematical Sciences, Community 7

- Administrative Science Quarterly
- American Journal of Sociology
- American Psychologist
- Crime and Justice
- Du Bois Review
- Journal of Organizational Behavior
- Research Evaluation
- Science, Technology and Human Value...
- Social Service Review

Mathematical Sciences, Community 8

- American Statistician
- Biometrika
- Journal – American Statistical Ass...
- Journal of Econometrics
- Journal of Educational Psychology
- NBER Technical Working Paper
- Psychological Methods
- Working paper series (National Bure...

Mathematical Sciences, Community 9

- Computer Architecture News
- Proceedings of the 2005 INFOCOM 24t...
- Systems Administration Conference
- www.eecs.harvard.edu/ margo/papers/...

Mathematical Sciences, Community 10

- Addictive Behaviors
- American Journal of Drug and Alcoho...
- Behavior Research Methods
- Learning and Motivation

Mathematical Sciences, Community 11

- Acta Crystallographica Section C: C...
- Chemistry & Biology
- Tetrahedron Letters

History / Philosophy, Community 1

- American Historical Review
- American Scientist
- Annals of Science
- Annual Review of Sociology
- Architectural History
- British Journal for the History of ...
- Bulletin of the History of Medicine
- Central European History
- Common Knowledge
- Configurations
- Contemporary European History
- Historical Journal
- History of Science
- International Migration Review
- Isis
- Journal of British Studies
- Journal of the History of Biology
- Journal of the History of Ideas
- Journal of Visual Culture
- Modern Language Notes
- Oxford Review of Education
- Past and Present
- Perspectives on Science
- Public Understanding of Science
- Science in Context
- Science Studies
- Shakespeare Survey
- Social Forces
- Social Studies of Science
- The British Journal for the History...
- The British Medical Journal
- The International Migration Review
- Transactions of the Institute of Br...
- Transactions of the Royal Historica...

History / Philosophy, Community 2

- Australasian Journal of Philosophy
- Biological Theory
- Bulletin of Symbolic Logic
- European Journal of Philosophy
- Journal of Aesthetics and Art Criti...
- Journal of Symbolic Logic
- Nous
- Philosophers' Imprint
- Philosophical Quarterly
- Philosophical Review
- Philosophical Studies
- Philosophical Topics
- Philosophy and Phenomenological Res...
- Proceedings of the Aristotelian Soc...
- Southern Journal of Philosophy
- The Cambridge Companion to the Phil...
- Theoria : Revista de Teoria, Histor...

History / Philosophy, Community 3

- Contemporary Readings in Law and So...
- Economics and Philosophy
- Ethics
- Harvard Divinity Bulletin
- Journal of Philosophy
- Pacific Philosophical Quarterly
- Philosophical Perspectives
- Philosophy and Public Affairs
- Proceedings and Addresses of the Am...
- Ratio
- Social Philosophy and Policy
- The Constitution of Agency
- The Monist
- The Quality of Life
- The Tanner Lectures on Human Values
- Women, Culture, and Development: A ...

History / Philosophy, Community 4

- American Literary History
- American Literary Scholarship
- American literature
- Critical Inquiry
- Prospects

Linguistics, Community 1

- riu
- Acta Poetica
- Amsterdam Studies in the Theory and...
- Annual of Armenian Linguistics
- Baltistica
- Brain Research
- Die Sprache
- Euskalingua
- Harvard Ukrainian Studies
- Harvard Working Papers in Linguisti...
- Heritage Language Journal
- Historische Sprachforschung
- Indo-European Studies
- Innsbrucker Beitrge zur Spr...
- Journal of Cuneiform Studies
- Journal of Indo-European Studies
- Journal of the American Oriental So...
- Journal of the Cork Historical and ...
- Language
- Language and Linguistics Compass
- Language Research
- Lingua
- Linguistic Variation Yearbook
- Mnchener Studien zur Sprach...
- MIT Working Papers in Linguistics
- Natural Language and Linguistic The...
- Oceanic Linguistics
- Proceedings of the North East Lingu...
- Syntax
- The Crane Bag
- Tocharian and Indo-European Studies
- Transactions of the Philological So...

Linguistics, Community 2

- Brain and Language
- Journal of East Asian Linguistics
- Journal of Memory and Language
- Linguistic Inquiry
- Synthese

Miscellaneous, Community 1

- Book History
- Contributions to the History of Con...
- European Review
- French Historical Studies
- Journal of Modern History
- Modern Intellectual History
- Pmla
- Princeton University Library Chroni...
- Proceedings of the British Academy
- Representations

Miscellaneous, Community 2

- American Journal of Public Health
- Annals of Internal Medicine
- Daedalus
- International Anesthesiology Clinic...
- Journal of Social Work and Human Se...
- Journal of the American Medical Ass...
- Journal of the American Medical Wom...
- Milbank Quarterly
- The Hastings Center Report

Miscellaneous, Community 3

- American Music
- Black Music Research Journal
- Early Music History
- Journal of Musicology
- Journal of the Society for American...
- Musical Quarterly
- Tempo

Law, Community 1

- Emory Law Journal
- Georgetown Law Journal
- Harvard Civil Rights-Civil Libertie...
- Journal of Legal Analysis
- Loyola of Los Angeles Law Review
- Maryland Law Review
- Michigan Law Review
- New York University Law Review
- Roger Williams University Law Revie...
- Southern California Law Review

Law, Community 2		
Arizona Law Review	Global Policy	Harvard BlackLetter Journal
Harvard International Law Journal	Harvard Journal on Legislation	Harvard Law Review
Harvard Women's Law Journal	Lewis and Clark Law Review	Minnesota Law Review