# Communities and bottlenecks: Trees and treelike networks have high modularity

James P. Bagrow

Phys. Rev. E **85**, 066118 — Published 15 June 2012

DOI: 10.1103/PhysRevE.85.066118

# Communities and bottlenecks: Trees and treelike networks have high modularity

James P. Bagrow*
*Department of Engineering Sciences and Applied Mathematics,*
*Northwestern Institute on Complex Systems,*
*Northwestern University, Evanston, Illinois, USA*

Much effort has gone into understanding the modular nature of complex networks. Communities, also known as clusters or modules, are typically considered to be densely interconnected groups of nodes that are only sparsely connected to other groups in the network. Discovering high quality communities is a difficult and important problem in a number of areas. The most popular approach is the objective function known as Modularity, used to both discover communities and measure their strength. To understand the modular structure of networks it is then crucial to know how such functions evaluate different topologies, what features they account for and what implicit assumptions they may make. We show that trees and treelike networks can have unexpectedly and often arbitrarily high values of modularity. This is surprising since trees are maximally sparse connected graphs and are not typically considered to possess modular structure, yet the non-local null model used by modularity assigns low probabilities, and thus high significance, to the densities of these sparse tree communities. We further study the practical performance of popular methods on model trees and on a genealogical dataset, and find that the discovered communities also have very high modularity, often approaching its maximum value. Statistical tests reveal the communities in trees to be significant, in contrast with known results for partitions of sparse, random graphs.

PACS numbers: 89.75.Hc, 89.75.Fb, 05.10.a, 89.20.Hh

## I. INTRODUCTION

Complex networks have made an enormous impact on research in a number of disciplines [1–5]. Networks have revolutionized the study of social dynamics and human contact patterns [6–8], metabolic and protein interaction in the cell [9, 10], ecological food webs [11–13], and technological systems such as the world wide web [14, 15] and airline transportation networks [16, 17]. Seminal results include the small-world [18] and scale-free nature [14] of many real world systems.

One of the most important areas of network research has been the study of community structure [19, 20]. Communities, sometimes called modules, clusters, or groups, are typically considered to be subsets of nodes that are densely connected among themselves while being sparsely connected to the rest of the network. Networks containing such groups are said to possess modular structure. Understanding this structure is crucial for a number of applications from link prediction [21] and the flow of information [22] to a better understanding of population geography [23–25].

Much effort has been focused on finding the best possible partitioning of a network into communities. Typically this is done by optimizing an objective function that measures the community structure of a given partition. Many algorithmic approaches have been devised. Most partition the entire network while some focus on local discovery of individual groups [26–28]. Overlapping community methods, where nodes may belong to more than one group, have recently attracted much interest [29–31]. For a lengthy review of community methods see [19].

Given the reliance on objective functions, it is important to understand how the intuitive notion of communities as internally dense, externally sparse groups is encoded in the objective function. Some functions simply measure the density of links within each community, ignoring the topological features those links may display, while other functions rely upon those links forming many loops or triangles, for example. We show the importance of understanding these distinctions by revealing some surprising features of how communities are evaluated. In particular we show that the only requirement for strong communities—according to the most popular community measure—is a lack of external connections, that bottlenecks [32] leading to isolated groups can make strong communities even when those groups are internally maximally sparse. This contradicts the notion of communities as being unusually densely interconnected groups of nodes.

This paper is organized as follows. In Sec. II we present several measures of community quality and discuss their different features and purposes. In Sec. III we show analytically that trees and treelike graphs can possess partitions that display very high, often arbitrarily high values of modularity. This is our primary result. In Sec. IV we apply two successful community discovery algorithms to these trees and show that the discovered communities can have even higher modularities. We also study the community structure of a treelike network derived from genealogical data. In Sec. V we perform statistical tests on the various communities and find that most of the partitions we consider for trees are statistically significant. We finish with a discussion and conclusions in Sec. VI.

* james.bagrow@northwestern.edu; http://bagrow.com

## II. MEASURING COMMUNITIES

Given a network—represented by a graph $G$ of $N$ nodes and $M$ links whose structure is encoded in an $N \times N$ adjacency matrix $A$ where $A_{ij} = 1$ if nodes $i$ and $j$ are connected and zero otherwise—we wish to determine to what extent $G$ possesses modular structure. To put the notion of a community or module onto a firm foundation, objective functions have been introduced to quantify how "good" or "strong" a community or a partitioning into communities is. These objective functions are also often the goal of an optimization algorithm, where the algorithm attempts to find the community or communities that maximizes (or minimizes) the objective function. Here we briefly discuss three objective functions: subgraph conductance, modularity, and partition density. Due to its popularity and wide use we will focus primarily on modularity.

### A. Conductance

The conductance $\phi$ of a subgraph is a measure of how 'isolated' the subgraph is, in analogy with electrical conductance [33]. Subgraphs with many connections to the rest of the network will have high conductance, whereas a subgraph will have low conductance if it relies on a few links for external connectivity. For a given subgraph $S$ such that $|S| \leq N$, one form of conductance is

$$\phi(S) = \frac{\sum_{i,j} A_{ij} \left[i \in S\right] \left[j \notin S\right]}{\sum_{i,j} A_{ij} \left[i \in S\right] \left[j \in S\right]} = \frac{K_S - 2m_S}{2m_S}, \qquad (1)$$

where $\left[P\right] = 1$ if proposition $P$ is true and zero otherwise, $K_S = \sum_{i,j} A_{ij} \left[i \in S\right]$ is the sum of the degrees (number of neighbors) of all nodes in $S$, and $m_S$ is the total number of links in $S$. (The factor of two in the denominator is sometimes dropped.) In other words, subgraph conductance is the ratio between the number of links exiting the subgraph to the number of links within the subgraph.

While low $\phi$ may appear to be a good indicator of community structure, we remark that it primarily measures isolation or "bottleneckedness," meaning that, e.g., a random walker moving in a subgraph with low conductance will be very few opportunities to exit the subgraph, whereas it would have many opportunities if the subgraph had high conductance. This is also true if the subgraph is a densely interconnected module. However, consider a large 2D periodic square lattice of size $L_x \times L_y$, $L_x \geq L_y$. This graph has $N = L_x L_y$ nodes and $M = 2N$ links and is generally considered to have no modular structure. The conductance of a subgraph created by cutting the lattice in half along the $y$ direction is $\phi = 2L_y/\left(L_x L_y\right) = 2/L_x$. As the lattice grows, the conductance of this subgraph decreases, despite there being no modular structure.

### B. Modularity

A key point lacking in earlier definitions of communities such as conductance is that they fail to quantify the statistical significance of the subgraph. It may be possible for a randomized null graph to contain subgraphs exhibiting comparable conductance, for example, and conductance alone does not capture this. Modularity [34, 35] was introduced to account for this in an elegant way. It has become the most common community objective function [19, 20] and possesses a number of distinct advantages over previous approaches, such as not requiring the number of communities to be known in advance. However, it has some drawbacks as well. It is known to possess a *resolution limit* where it prefers communities of a certain size that depends only on the global size of the network and not on the intrinsic quality of those communities [36, 37]. Meanwhile, sparse, uncorrelated random graphs are expected not to possess modular structure, but fluctuations may lead to partitions with high modularity [38–40]. Yet another concern is modularity's highly degenerate energy landscape [41], which may lead to very different yet equally high modularity partitions.

Modularity $Q$ can be written as:

$$\begin{aligned} Q &= \frac{1}{2M} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2M}\right] \left[c_i = c_j\right] \\ &= \sum_c \left[\frac{m_c}{M} - \left(\frac{K_c}{2M}\right)^2\right], \end{aligned} \qquad (2)$$

where $M = \frac{1}{2}\sum_{ij} A_{ij}$ is the total number of links in the network; $c_i$ is the community containing node $i$; $m_c = \frac{1}{2}\sum_{ij} A_{ij} \left[c_i = c\right] \left[c_j = c\right]$ is the total number of links inside community $c$; and $K_c = \sum_i k_i \left[c_i = c\right]$ is the total degree of all nodes in community $c$. The first definition of $Q$ illustrates the intuition of its form: For every node pair that shares a community we sum the difference between whether or not that pair is actually linked with the expected "number" of links between those same two nodes if the system was a purely random network constrained to the same degree sequence (this null model is known as the configuration model, and the loss term is approximate). This is then normalized by the total number of links in the network. By rewriting the sum over node pairs as a sum over the communities themselves, the second definition of $Q$ makes clear the resolution limit: global changes to the total number of links $M$ will disproportionately affect each community's local contribution to $Q$. This can potentially shift the maximal value of $Q$ to a different partition even when the local structure of the communities remains unchanged.

Equation 2 gives values between $-1$ and 1. When $Q \approx 0$ there is strong evidence that the discovered community structure is not significant, at least according to this null model, while the communities are considered better and more significant as $Q$ grows. In practice, researchers may assume that a network possesses modular structure when $Q > 0.25$ or $0.3$ [34]. However, since fluctuations can induce high modularity in random graphs, one must always approach the raw magnitude of $Q$ with caution: statistical testing (Sec. V) may provide stronger evidence for the presence of modules than modularity alone [38].

### C. Partition density

Yet another approach to quantifying community structure is that of partition density [30]. Partition density was introduced specifically for the case of link communities, where links instead of nodes are partitioned into groups. This allows for communities to overlap, since nodes may belong to multiple groups simultaneously. We do not consider overlapping communities here, but partition density can still be calculated for non-overlapping node communities.

The partition density $D$ is

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

Partition density measures, for each community, the number of links within that community minus the minimum number of links necessary to keep a subgraph of that size connected, $n_c - 1$, the size of its spanning tree. This is then normalized by the maximum and minimum number of links possible for that connected subgraph, $\binom{n_c}{2}$ and $n_c - 1$, respectively. The partition density is then the average of this quantity over the communities, weighted by the fraction of links within each community. For a link partition that covers an entire connected network, we have $\sum_c m_c = M$, but this does not necessarily hold for a node partition.

A crucial feature of the partition density is that it explicitly compares the link density of a subgraph to that of a tree of the corresponding size. This controls for the fact that the subgraph in question is connected, making the reasonable assumption that communities should be internally connected. The null model used by modularity on the other hand, does not make this assumption, and it may potentially assign very low probabilities to such an event. As we will show, this is a crucial aspect of modularity.

## III. COMMUNITIES IN TREES AND TREELIKE GRAPHS

We now study a model tree graph that one may consider to not possess modular structure and show that these graphs possess partitions with arbitrarily high modularity values. We also study a mixed case graph containing both modular and non-modular structures.

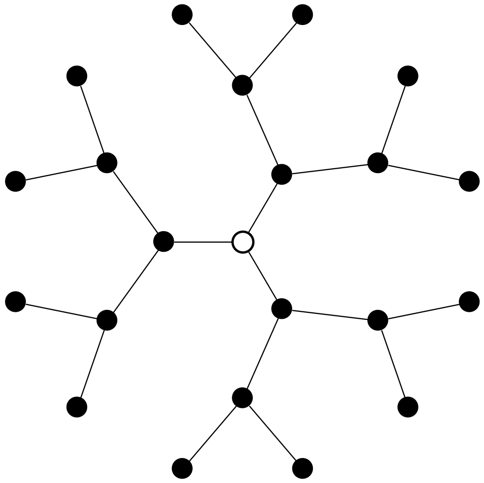

FIG. 1. Cayley tree for $z = 2$ and $g = 3$. The root node is indicated in white.

### A. Cayley tree

The Cayley tree is a regular graph with no loops and where every node $i$ has the same degree $k_i = z + 1$ (except for leaf nodes on the boundary who possess $k = 1$). See Fig. 1. It can be constructed by first starting from a root node at generation 0, giving that node $z + 1$ child nodes, and then repeatedly giving each new child $z$ children of its own. This continues for a fixed number of generations $g$. These trees can grow either in "width" (via $z$) or in "depth" (via $g$). The number of nodes in generation $g > 0$ is $n(g) = (z + 1)z^{g-1}$ and the total number of nodes is $N(g) = 1 + \sum_{g'=1}^{g} n(g')$. Since this is a tree, the total number of links is $M(g) = N(g) - 1 = (z + 1)(1 - z^g)/(1 - z)$. Since the bulk of the graph is regular, the Cayley tree has no density fluctuations (all connected subgraphs of the same size have the same number of links) and so it does not in an obvious way conform to our preconceived notions of communities as internally dense, externally sparse groups. In the thermodynamic limit the Cayley tree is known as the Bethe lattice. We concern ourselves here primarily with finite graphs, however, such that finite size and edge effects cannot be ignored.

We now compute the modularity of a specific partition of the Cayley tree, which we call the *analytic partition.* First place the root node into a community of its own. Then create a new community for each child of the root node, containing that child *and all of its descendants.* Thus there are $z + 2$ communities in total. Apart from the singleton community containing the root node, every community is a complete $z$-ary tree (which is not exactly a Cayley tree) with $g - 1$ generations. Partitioning the tree in this way requires cutting only $z + 1$ links. There are zero links inside the singleton community and

$$m = \frac{N(g) - 1}{z + 1} - 1 = z\frac{1 - z^{g-1}}{1 - z} \tag{4}$$

links inside the $z+1$ other communities. To compute the total degree of nodes within the community, we note that all $(N(g)-1)/(z+1)$ nodes have degree $z+1$ except the $n(g)/(z+1)$ boundary nodes that have degree 1. Thus the total degree is

$$K = (z+1)\left(\frac{N(g)-1}{z+1} - \frac{n(g)}{z+1}\right) + \frac{n(g)}{z+1}$$
$$= \frac{1+z-2z^g}{1-z}. \quad (5)$$

The final modularity is then given by substituting these expressions for $m$, $K$, and $M$ into:

$$Q_{\text{cayley}} = (z+1)\left[\frac{m}{M} - \left(\frac{K}{2M}\right)^2\right] - \left(\frac{z+1}{2M}\right)^2, \quad (6)$$

where the functional dependence on $z$ and $g$ has been suppressed. For $z=10$ and $g=4$, for example, $Q_{\text{cayley}} \approx 0.91$, an extremely high modularity. Even for $z=3$ and $g=3$ we have a high modularity of $Q_{\text{cayley}} \approx 0.7$. (Raw modularity values must be approached with caution; we will quantify these numbers in Sec. V.) In general, the limiting value of $Q_{\text{cayley}}$ for a given $z$ is

$$\lim_{g\to\infty} Q_{\text{cayley}}(z,g) = \frac{z}{z+1}. \quad (7)$$

Even for a finite $g>1$, $Q_{\text{cayley}} \to 1$ as $z\to\infty$. Thus the Cayley tree is able to achieve **arbitrarily high** modularity partitions. (We will later show these partitions to also be statistically significant.) This is not the only partition capable of achieving high modularity. We discuss another partition in the Appendix.

Meanwhile, the $z+1$ branch communities of the Cayley tree's analytic partition each have conductance

$$\phi_{\text{cayley}} = \frac{1}{m} = \frac{1-z}{z-z^g}. \quad (8)$$

For $z=4$ and $g=10$, for example, $\phi_{\text{cayley}} \approx 2.86\times10^{-6}$, a very small value. This makes sense since only a single link separates that entire branch from the rest of the graph. This also emphasizes that conductance is primarily a measure of bottlenecks and isolation and should be approached with caution when applied to community structure.

Finally, we remark that the partition density of the Cayley tree is zero since $m_c = n_c - 1$. This is true not just for the analytic partition but for all partitions of the Cayley tree where each community is connected.

### B. A clique and a tree

In practice, one may deal with networks with wide fluctuations in local density, meaning there may exist localized subgraphs of low and of high density at the same time. We analyze a simple example consisting of a single complete graph known as a clique connected by one link to the root of a $z$-ary tree of $g$ generations. See Fig. 2.

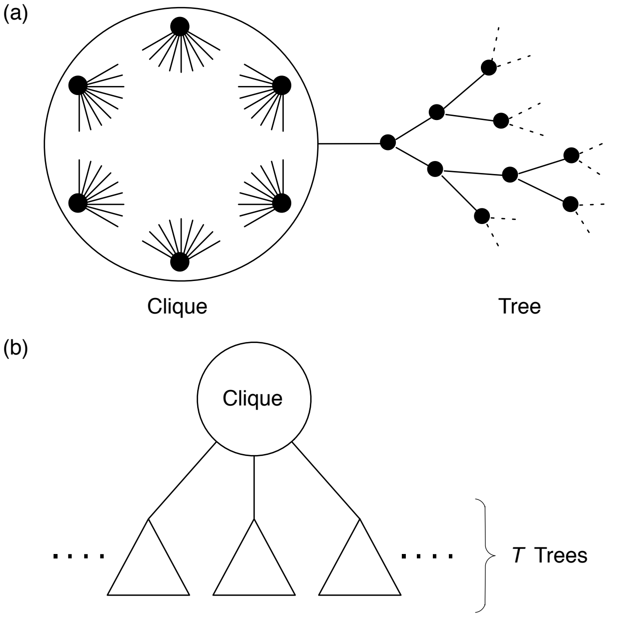

FIG. 2. (left) A mixed test case consisting of a single clique (complete subgraph) of $n_{\text{clique}}$ nodes connected by a single link to a $z$-ary tree. This is partitioned into communities by cutting the single bridging link. (right) A generalization where now $T$ trees are connected to the single clique ($T < n_{\text{clique}}$).

We wish to compute the modularity of a community partition containing the entire clique in one community and the entire tree in the other, where only the link from the root of the tree to the clique was cut. We assume that there are $n_{\text{clique}}$ nodes in the clique and $n_{\text{tree}} = \left(1-z^{g+1}\right)/(1-z)$ nodes in the tree. The numbers of links in each subgraph are $m_{\text{clique}} = \binom{n_{\text{clique}}}{2}$ and $m_{\text{tree}} = z\left(1-z^g\right)/\left(1-z\right)$, respectively. The total number of links is $M = m_{\text{clique}} + m_{\text{tree}} + 1$ and the total degrees are $K_{\text{clique}} = n_{\text{clique}} + (n_{\text{clique}}-1)^2$ and $K_{\text{tree}} = z^g + (z+1)\left(1-z^g\right)/\left(1-z\right)$. The final modularity of the partition is then given by substituting these expressions into

$$Q_{\text{clique}-\text{tree}} = \frac{m_{\text{clique}} + m_{\text{tree}}}{M} - \frac{K_{\text{clique}}^2 + K_{\text{tree}}^2}{4M^2}. \quad (9)$$

We plot Eq. (9) as a function of $g$ in Fig. 3a for $n_{\text{clique}} = 100$ and several values of $z$. We see that $Q$ attains a maximum of 1/2, a value not as high as the pure Cayley tree previously analyzed despite the addition of a "perfect" community. We also see that, as $z$ increases, $Q$ becomes more sharply peaked as a function of $g$. This is due to the resolution limit: the larger $z$ is, the more quickly the tree will grow from one generation to the next and thus the tree community more quickly passes beyond the size preferred by modularity. This leads to a $Q$ that grows more rapidly and then decays more rapidly as $g$ increases.

We also study a generalization of Fig. 2a from one tree to $T$ trees (Fig. 2b), where each tree is its own community. For this model $m_{\mathrm{tree}}$ and $K_{\mathrm{tree}}$ are unchanged for each tree, while now $K_{\mathrm{clique}} = T + n_{\mathrm{clique}}(n_{\mathrm{clique}} - 1)$, $M = m_{\mathrm{clique}} + T m_{\mathrm{tree}} + T$ and

$$Q_{\mathrm{clique-trees}} = \frac{m_{\mathrm{clique}} + T m_{\mathrm{tree}}}{M} - \frac{K_{\mathrm{clique}}^2 + T K_{\mathrm{tree}}^2}{4M^2}. \tag{10}$$

We plot Eq. (10) in Fig. 3b as a function of $n_{\mathrm{clique}}$ for several values of $T$. We see that increasing $T$ raises the overall modularity of the partition, giving apparently high values of $Q_{\mathrm{clique-trees}} > 0.8$. We also see that, as $T$ increases, the curve becomes more flat, meaning that good quality partitions, according to modularity, exist for a wide range of clique sizes. We remark that this generalization may also be treated by exploiting the recursive nature of $z$-ary trees by merging all the tree roots into one node and moving that node into the clique (this is particularly simple when $T = z$).

In Sec. V we study the statistical significance of these Clique-Tree partitions.

### C. Other trees

The results above were derived for Cayley and $z$-ary trees. The regular nature of these trees allows for tractable expressions of modularity, but our results are not limited to these types of trees. The important feature in this context is that all connected subgraphs of $n$ nodes in any tree will always contain $m = n - 1$ links. Since it seems a reasonable basic requirement for a community detection method to discover communities that are connected, this density relation is a reasonable minimum baseline for a method to be compared against. This also means that, since every tree obeys this relation, bottlenecks become the primary drivers of high modularity partitions in all trees. We further explore the generality of our results in Sec. IV where we apply community detection algorithms to random trees.

## IV. REAL-WORLD EXAMPLES

The above derivations show that trees may possess arbitrarily high values of modularity. However, these calculations did not consider the resolution limit of modularity. In fact, real world optimization of modularity will result in partitions that give *even higher* values of $Q$ than those of the analytic partitions discussed in Sec. III.

To see this we apply two of the most popular and successful community discovery methods. The first is known as **Fast Unfolding** (sometimes referred to as the Louvain method) and can efficiently find very high modularity partitions [42]. The second method is called **Infomap** [43]. Infomap does not optimize modularity, instead exploiting information-theoretic arguments, but the partitions it does find are often high in modularity, especially for undirected networks.

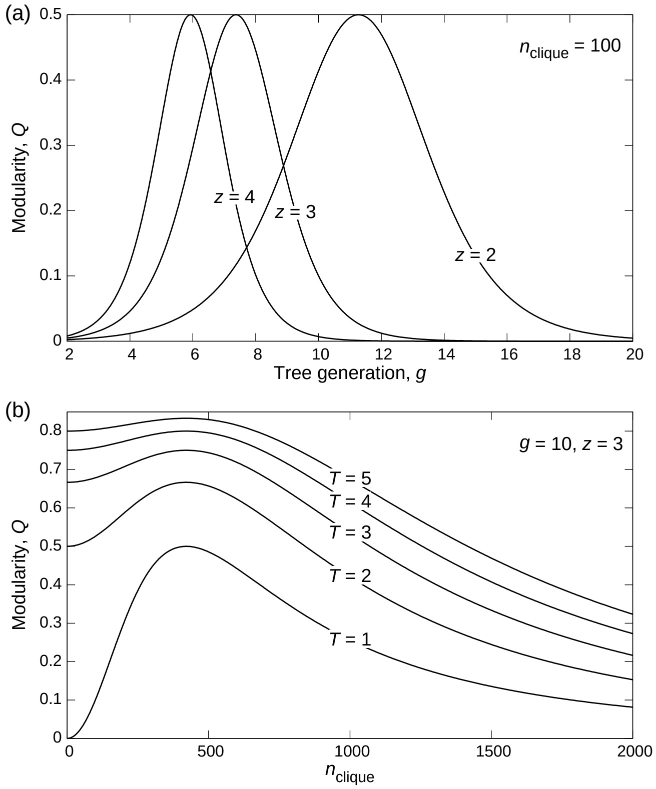

FIG. 3. (a) Modularity for the example network illustrated in Fig. 2a. As we increase the size of the tree for a fixed clique size, the modularity grows to a maximum value and then decays away. This is due to the resolution limit: there exists a specific tree size that maximizes $Q$ for each clique size. (b) For the generalization of one clique and $T$ trees, shown in Fig. 2b, we see that the analytic partition again attains high modularity, especially as more trees are added. Likewise, as $T$ increases we see that the peak of $Q$ flattens out and that the partition has high modularity for a range of clique sizes. This means that much of the resolution limit can be compensated for if the network is sufficiently treelike.

We apply these algorithms to the Cayley tree. In Fig. 4 we plot the modularities discovered by each algorithm and the modularities $Q_{\mathrm{cayley}}$ of the analytic partitions (Eq. 6). We see that the methods find communities that appear as strong as the analytic method or stronger. Fast Unfolding typically exceeds $Q_{\mathrm{cayley}}$ as the trees grow, and even approaches $Q = 1$. Infomap tends to stay closer to $Q_{\mathrm{cayley}}$ but it too can exceed these bounds, especially for trees with $z = 2$. If these methods were applied blindly to a network, such high values of modularity would suggest that these communities are extremely high quality and that the network was extremely modular.

What about trees other than the Cayley tree? Will such discovery methods find comparable values of modularity. To answer this, we now apply these methods to random trees generated from a Galton-Watson branching process [44], where each node has a random number of de-

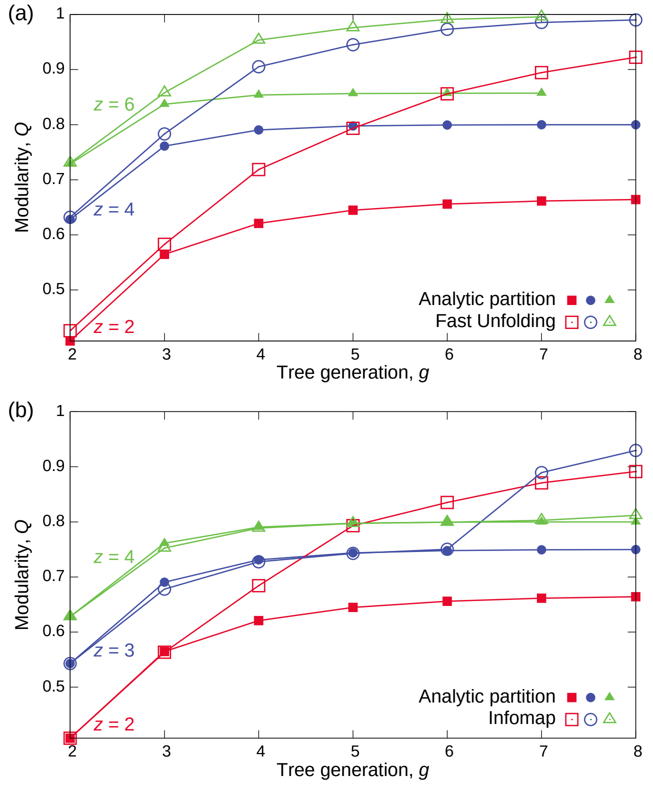

FIG. 4. (Color online) Community discovery methods find even higher values of modularity than the analytical partition of the Cayley tree. We apply two methods: (a) Fast Unfolding [42] and (b) Infomap [43] for several values of $z$. Closed symbols correspond to Eq. (6) while open symbols correspond to the modularities found by the methods. Fast Unfolding finds consistently higher modularity partitions than the analytic partition, due to the resolution limit. These partitions even approach $Q = 1$. Infomap, which does not optimize modularity, tends to find partitions comparable to the analytic partition, although it too finds higher value partitions for $z = 2$ and some values of $g$ for $z = 3$. Note that the vertical axes do not begin at $Q = 0$.

scendants drawn from a Poisson distribution with mean $\lambda$. (We also stop growing the tree at $g$ generations.) For $\lambda = 4$ and $g = 6$, for example, we find Fast Unfolding partitions with modularity $Q = 0.9814 \pm 0.0055$, while for Infomap we find partitions with $Q = 0.8594 \pm 0.0069$. This supports the generality of our results: high modularity partitions also exist in non-Cayley trees.

As a final practical example, we also apply both methods to a treelike network derived from a genealogical dataset capturing the advisor-advisee relationships between mathematicians and their students [45, 46]. (This genealogy is not exactly a tree since some students have multiple advisors.) We only consider the giant connected component of the network, capturing approximately 90% of the dataset. In total the network has $N = 133319$ nodes and $M = 148247$ links. The modularities of the partitions found by Fast Unfolding and Infomap are $Q_{\mathrm{FU}} = 0.951083$ and $Q_{\mathrm{IM}} = 0.877146$, respectively. These high values would again imply that the network is strongly modular; however, statistical testing should be performed to support this argument (see Sec. V).

As a brief aside, another interesting aspect of a community partition is the distribution of community sizes (numbers of nodes per community). Since any discovered modular network structure depends intrinsically on the definition at the heart of the algorithm used to find that structure, it is not known for certain what the true distribution is. Nevertheless, there has been empirical evidence showing that the size distribution may exhibit a power law $Pr(s) \sim s^{-\alpha}$, for $\alpha \geq 1$ [29, 30].

Yet the distributions of community sizes found in the genealogical network, shown in Fig. 5, are not heavy-tailed. Instead both methods find approximately exponential distributions, with a small number of larger communities that would be underrepresented by a exponential distribution. The lack of very large communities may be expected in graphs without hubs, but the degree distribution for this network is heavy-tailed (Fig. 5b, inset). This relatively narrow size distribution may provide some warning that the communities found in this network differ from typical communities in some meaningful way, though this is far from certain. Further study of this distribution may prove fruitful in understanding the modular nature of complex systems.

## V. STATISTICAL TESTING

Given that there exist high modularity partitions in both the Cayley tree and the Mathematics Genealogy, a crucial question becomes, are these partitions significant in some way or could they be simply due to some random process? This is especially important since it is known that sparse, uncorrelated graphs can potentially possess high modularity partitions due to fluctuations [38–40]. To address such questions of statistical significance requires first defining an appropriate null model. Hypothesis testing then asks what is the probability that the observed phenomena (in this case the discovered communities or their properties) may have arisen within the null model. If this probability is sufficiently low, then there is evidence that the communities cannot be explained by the null model. This does not mean that the communities are "meaningful," however, since this only compares them to that particular null model. For example, a simple choice of null model is the configuration model: build uncorrelated random graphs that preserve the degree sequence of the original graph, apply community detection to these graphs, and then compare the configuration model communities to those of the original network. However, a Cayley tree is typically very different from its equivalent configuration model ensemble, being highly structurally ordered, and this alone may lead to statistically significant differences in, e.g., modularity. Thus it is crucial to choose the most appropriate

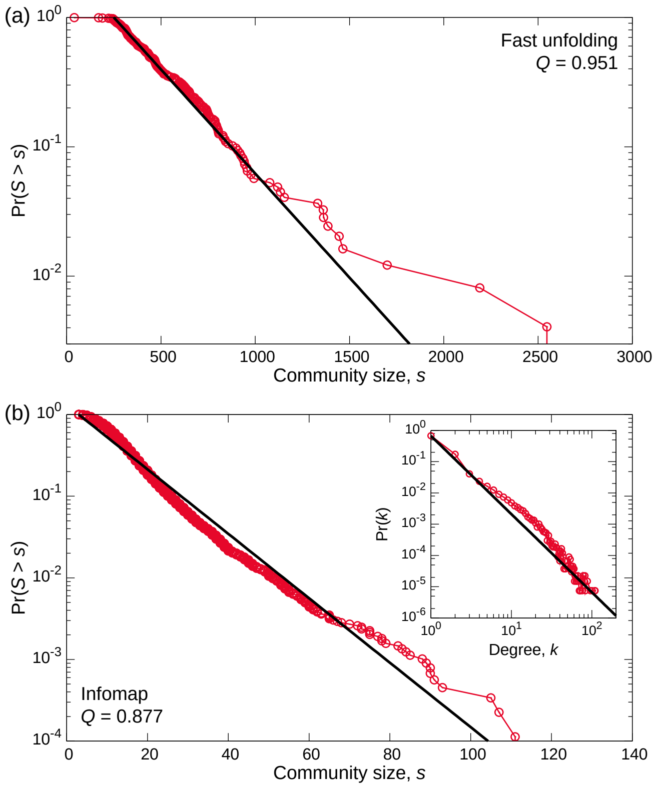

FIG. 5. The distribution of community sizes found in the genealogical network for (a) Fast Unfolding and (b) Infomap. We see that neither distribution is heavy-tailed, being instead approximately exponential (straight lines) except for a small number of the largest communities that would be underestimated by an exponential distribution. (inset) Unlike the community size distribution, the degree distribution of the network is heavy-tailed. The straight line shows a pure power law, $\Pr(k) \sim k^{-2.5}$, for comparison.

null model possible.

Defining statistical tests for community structure remains an area of research [47, 48]. For our purposes, we use the testing procedure introduced in [47]. Roughly, this test takes the worst node $w$ in a community $c$ (the node with the least neighbors also inside $c$), removes $w$ from $c$, and asks what is the probability $p$ that $w$ would have that many neighbors or more within $c$ if its links were distributed randomly over the whole graph while holding the rest of $c$ and the degrees of all other nodes fixed. Since $w$ represents the worst case, if $p$ is small then it is unlikely that $w$ or any other nodes in $c$ would have so many other neighbors also in $c$ due to chance. Therefore, they consider a community to be significant if $p < 0.05$, the standard significance level for hypothesis testing. Note that this test strictly controls for both the links inside $c$ (and therefore its density), the number of links exiting $c$, and the overall sparsity of the network. The authors in [47] used this test to show that communities in sparse Erdős-Rényi and power law graphs are not significant. For full details, see [47].

Here, we compute a $p$ for each community of interest using their method and ask what fraction of communities are significant.

For the analytic partitions of the Cayley tree (Sec. III A), we find that all $z+1$ branch communities are statistically significant. (For $z = 3$ and $g = 7$, for example, the maximum $p = 0.00144$.) This means that this particular partition cannot be explained simply by the global sparsity of the tree itself. In particular, this test considers the internal structure of a community as fixed, and yet these communities are significant even though they are internally maximally sparse trees. This contradicts the typical intuition of communities as internally dense, externally sparse subgraphs and supports the argument that bottlenecks alone are sufficient to significantly optimize modularity.

We next test the analytic partition of the Clique-Tree example (Sec. III B). We consider for $n_{\text{clique}} = 100$, all $z = 2, 3, 4$, and $g = 2, \ldots, 10$ (see also Fig. 3a). In all cases both communities were significant: the maximum $p$ observed from any combination of those parameters was $p = 0.00387$.

We now turn to the community discovery methods Fast Unfolding and Infomap. For Fast Unfolding on the Cayley tree ($z = 3$, $g = 7$) we find that approximately 91% of the discovered communities were significant. This shows that even practical methods can find statistically significant, high modularity partitions solely through the discovery of bottlenecks. For Infomap, which does not optimize modularity, we find that no communities are statistically significant according to this test. However, we remark that a less strict test also introduced in [47] shows that approximately 48% of the Infomap communities are significant. The truth likely lies between these extremes, but we can safely conclude that most Infomap communities could be explained by this test's null model.

Next, we test Fast Unfolding and Infomap on the finite Galton-Watson trees discussed in Sec. IV. We find comparable results to the partitions of the Cayley tree: for $\lambda = 4$ and $g = 6$, we find that $96.9\% \pm 0.808\%$ Fast Unfolding communities were significant while almost no ($0.0311\% \pm 0.0350\%$) Infomap communities were significant.

Finally, we consider the practical example of the mathematics genealogy. Both Fast Unfolding and Infomap found apparently high values of modularity, but are these results significant? Applying this test shows that they are not: for Fast Unfolding and Infomap only approximately 2.4% and 2.6% of the communities were significant, respectively. Although we again caution that this does not necessarily prove these communities to be meaningless, it further underlines the potential danger of relying upon raw modularity values as a quantifier of modular structure.

## VI. DISCUSSION AND CONCLUSIONS

We have shown that trees appear very modular. Yet connected trees are maximally sparse and possess no density fluctuations, going against the tenet that communities are unusually dense subgraphs. Thus, counter to our intuition, measures such as modularity, while ostensibly rewarding densely interconnected groups, can actually be optimized solely through the discovery of bottlenecks and it is not necessary for the discovered groups to be internally dense. In particular, we do not claim that trees lack communities, nor do we claim that these communities are not meaningful in some way. Instead we argue only that it is sufficient to discover bottlenecks to optimize modularity and conductance. This disconnect between intuition and practice has not been well discussed in the literature and in fact most work has overlooked the out-sized role that bottlenecks play in the existence of modular structure.

So is our definition of modular structure correct? Equation 2 depends so strongly on its null model that we must judiciously understand all facets of it. We have shown that communities do not need significantly high internal density to lead to high quality (according to modularity). Therefore, if researchers want to consider modules according to their intuition, they may need to introduce measures that specifically account for internal density in some way beyond that of Eq. 2. Taken together with modularity's other issues such as its resolution limit, it appears that rigorously and unambiguously quantifying modular network structure is difficult and remains an open question.

Researchers have shown that sparse graphs will have high modularity, yet the statistical tests applied here show that the sparsity of trees alone is not sufficient to explain these results. By controlling for tree sparsity, we have shown that bottlenecks lead not only to high modularity but to statistically significantly high modularity. Our results on trees further differ from sparse random graphs in that the expected high modularity partitions do not need to be equipartitions (see Appendix), and the derivations here do not invoke features of *ensembles* of random graphs.

One may suspect that the addition of non-tree components to a network may destroy the observed phenomena—that it is somehow fragile—yet our results in Sec. III B show that this is not the case and that merely the presence of trees may lead to modular structure. A crucial consequence of this is that, since trees are the limiting structure as networks become sparse, sampled and missing data [49] may boost modularity, at least in some regions of the network, even though the network remains globally connected. Incomplete data remains an issue in high-throughput biological assays for example [50], and thus one should consider both sparsity and bottlenecks when approaching graph partitioning in these problems.

Finally, the statistical tests we used in Sec. V show that many of the communities found in trees are significant, whereas the communities found in the mathematics genealogy, while very high in modularity, are typically not significant. However, this test does not verify that the tree communities are "meaningful," only that they differ from the test's null model. Likewise, the discovered genealogical communities could still be meaningful in other ways, perhaps revealing important schools of mathematicians or mathematics research. For networks that possess additional data annotating the properties or roles of network elements—for example Gene Ontology terms describing proteins in protein-protein interaction networks [51–54]—these discovered groups may in fact be highly *enriched*, meaning that their nodes or links share many annotations [30, 55], even though *structurally* the community is not significant. Further study of the interplay between these different validation mechanisms may be crucial to increasing our understanding of modular networks and complex systems in general.

## ACKNOWLEDGMENTS

We thank Gino Biondini for providing the genealogical data, the anonymous referees for their comments, and Dirk Brockmann and the Volkswagen Foundation for support. The authors' research was also supported by grants from the National Science Foundation (OCI-0838564-VOSS) and the U.S. Army Research Laboratory's Network Science Collaborative Technology Alliance (W911NF-09-2-0053).

### Appendix: Another high modularity partition of the Cayley tree

Consider the analytic partition of the Cayley tree, where the root node occupies a singleton community alongside the $z+1$ branch communities. Neglecting the singleton community, which has a vanishing contribution to $Q$ anyway, the communities are all the same size. We showed in Sec. III that the modularity of this partition $Q_{\mathrm{cayley}}$ can become arbitrarily close to one. This is not the only arbitrarily high modularity partition present in the Cayley Tree.

To see this, take one of the $z+1$ branch communities and "shatter" it such that all nodes in that branch now form singleton communities of their own. The modularity $Q_{\mathrm{shattered}}$ of this partition is

$$\begin{aligned} Q_{\mathrm{shattered}} = z &\left[\frac{m}{M} - \left(\frac{K}{2M}\right)^2\right] \\ &- \left(\frac{N(g)-1}{z+1} - \frac{n(g)}{z+1} + 1\right)\left(\frac{z+1}{2M}\right)^2 \\ &- \frac{n(g)}{z+1}\left(\frac{1}{2M}\right)^2 . \end{aligned} \tag{A.1}$$

Here the first term is the contribution of the remaining $z$ "unshattered" branches; the second term accounts for the losses due to singleton interior nodes, both the root node and the shattered branch interior; and the last term accounts for losses due to the singleton leaf or interface nodes of the shattered branch. The quantities $m$, $M$, $K$, $n(g)$, and $N(g)$ correspond to those derived in Sec. III A. Substituting these into Eq. A.1 gives

$$\lim_{g\to\infty} Q_{\text{shattered}} = \left(\frac{z}{z+1}\right)^2 . \tag{A.2}$$

As expected, this value is smaller than the limit $Q_{\text{cayley}} \to z/(z+1)$ as $g \to \infty$ but it shows that this partition still achieves arbitrarily high values of modularity.

As mentioned in the main text, it has been shown that sparse, uncorrelated random graphs may possess high modularity partitions [38, 39]. Under these conditions all communities should be equivalent in expectation and thus one expects all communities to be roughly comparable in size, so that the community structure must be an **equipartition** of the network. The high value of modularity displayed in Eq. A.2 shows that trees, while also being sparse, can contain high modularity partitions that are very far from equipartitions, in contrast with the results of [38, 39].

[1] M. E. J. Newman, *Networks: an introduction* (Oxford University Press, 2010).

[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47. 54 p (2001).

[3] M. Newman, SIAM Rev. **45**, 167 (2003).

[4] A. Barrat, M. Bathélemy, and A. Vespignani, *Dynamical processes on complex networks* (Cambridge University Press, 2008).

[5] A. Vespignani, Nature Physics **8**, 32 (2011).

[6] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. d. Menezes, K. Kaski, A.-L. Barabási, and J. Kertész, New J. Phys. **9**, 179 (2007).

[7] M. C. González, C. A. Hidalgo, and A.-L. Barabási, Nature **453**, 779 (2008).

[8] J. P. Bagrow, D. Wang, and A.-L. Barabási, PLoS ONE **6**, e17680 (2011).

[9] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabási, Nature **407**, 651 (2000).

[10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).

[11] J. Dunne, R. Williams, and N. Martinez, Ecology Letters **5**, 558 (2002).

[12] A. E. Krause, K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor, Nature **426**, 282 (2003).

[13] L. S. Schulman, J. P. Bagrow, and B. Gaveau, Advs. Compl. Syst. **14**, 829 (2011).

[14] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[15] J. Kleinberg, Nature **406**, 845 (2000).

[16] V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani, Proc. Natl. Acad. Sci. U.S.A. **103**, 2015 (2006).

[17] D. Brockmann, L. Hufnagel, and T. Geisel, Nature **439**, 462 (2006).

[18] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).

[19] S. Fortunato, Physics Reports **486**, 75 (2010).

[20] M. E. J. Newman, Nature Physics **8**, 25 (2011).

[21] A. Clauset, C. Moore, and M. E. J. Newman, Nature **453**, 98 (2008).

[22] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Proc. Natl. Acad. Sci. U.S.A. **104**, 7332 (2007).

[23] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, PLoS ONE **5**, e14248 (2010).

[24] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, Proc. Natl. Acad. Sci. U.S.A. **108**, 7663 (2011).

[25] C. Thiemann, F. Theis, D. Grady, R. Brune, and D. Brockmann, PLoS ONE **5**, e15422 (2010).

[26] J. P. Bagrow and E. M. Bollt, Phys. Rev. E **72**, 046108 (2005).

[27] A. Clauset, Phys. Rev. E **72**, 026132 (2005).

[28] J. P. Bagrow, J. Stat. Mech. **05**, 001 (2008).

[29] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature **435**, 814 (2005).

[30] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature **466**, 761 (2010).

[31] B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **84**, 36103 (2011).

[32] S. Sreenivasan, R. Cohen, E. López, Z. Toroczkai, and H. Stanley, Phys. Rev. E **75**, 036105 (2007).

[33] B. Bollobás, *Modern graph theory* (Springer Verlag, 1998).

[34] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, (2004).

[35] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).

[36] S. Fortunato and M. Barthelemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).

[37] A. Lancichinetti and S. Fortunato, Phys. Rev. E **84**, 066122 (2011).

[38] J. Reichardt and S. Bornholdt, Physica D: Nonlinear Phenomena **224**, 20 (2006).

[39] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).

[40] R. Guimera, M. Sales-Pardo, and L. Amaral, Phys. Rev. E **70**, 025101 (2004).

[41] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).

[42] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. **2008**, P10008 (2008).

[43] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. U.S.A. **105**, 1118 (2008).

[44] T. Harris, *The Theory of Branching Processes*, Phoenix Edition Series (Dover Publications, 2002).

[45] North Dakota State University, "The mathematics genealogy project," `http://genealogy.math.ndsu.nodak.edu`.

[46] R. D. Malmgren, J. M. Ottino, and L. A. N. Amaral, Nature **465**, 622 (2010).

[47] A. Lancichinetti, F. Radicchi, and J. J. Ramasco, Phys.

Rev. E **81**, 046110 (2010).

[48] A. Mirshahvalad, J. Lindholm, M. Derlén, and M. Rosvall, PLoS ONE **7**, e33721 (2012).

[49] J. P. Bagrow, S. Lehmann, and Y.-Y. Ahn, Arxiv preprint arXiv:1102.5085 (2011).

[50] H. Yu, P. Braun, M. Yldrm, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, and N. Simonis, Science **322**, 104 (2008).

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

[52] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga, Nature **440**, 631 (2006).

[53] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. S. Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt, Nature **440**, 637 (2006).

[54] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal, Science **322**, 104 (2008).

[55] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, SIAM Review **53**, 526 (2011).