Communities in Networks

Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha

Introduction: Networks and Communities

"But although, as a matter of history, statistical mechanics owes its origin to investigations in thermodynamics, it seems eminently worthy of an independent development, both on account of the elegance and simplicity of its principles, and because it yields new results and places old truths in a new light in departments quite outside of thermodynamics."

— Josiah Willard Gibbs,
Elementary Principles in Statistical Mechanics,
1902 [47]

From an abstract perspective, the term *network* is used as a synonym for a mathematical *graph*. However, to scientists across a variety of fields, this label means so much more [13, 20, 44, 83, 88, 120, 124]. In sociology, each *node* (or vertex) of a network represents an *agent*, and a pair of nodes can be connected by a *link* (or edge) that signifies some social interaction or tie between them (see Figure 1 for an example). Each node has a *degree* given by the number of edges connected to it and a *strength* given by the total weight of those edges. Graphs can represent either man-made or natural constructs, such as the World Wide Web or neuronal synaptic networks in the brain. Agents in such networked systems are like particles in traditional statistical mechanics that we all know and (presumably) love, and the structure of interactions between agents reflects the microscopic rules that govern their behavior. The simplest types of links are binary pairwise connections, in which one only cares about the presence or absence of a tie. However, in many situations, links can also be assigned a direction and a (positive or negative) weight to designate different interaction strengths.

Traditional statistical physics is concerned with the dynamics of ensembles of interacting and noninteracting particles. Rather than tracking the motion of all of the particles simultaneously, which is an impossible task due to their tremendous number, one averages (in some appropriate manner) the microscopic rules that govern the dynamics of individual particles to make precise statements of macroscopic observables such as temperature and density [112]. It is also sometimes possible to make comments about intermediate (*mesoscopic*) structures, which lie between the microscopic and macroscopic worlds; they are large enough that it is reasonable to discuss their collective properties but small enough so that those properties are obtained through averaging over smaller numbers of constituent items. One can similarly take a collection of interacting agents, such as the nodes of a network, with some set of microscopic interaction rules and attempt to derive the resulting mesoscopic and macroscopic structures.

Mason A. Porter, Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, and CABDyN Complexity Centre, University of Oxford. His email address is porterm@maths.ox.ac.uk.

Jukka-Pekka Onnela, Harvard Kennedy School, Harvard University; Department of Physics, University of Oxford; CABDyN Complexity Centre, University of Oxford; and Department of Biomedical Engineering and Computational Science, Helsinki University of Technology. His email address is jp_onnela@ksg.harvard.edu.

Peter J. Mucha, Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, University of North Carolina, and Institute for Advanced Materials, Nanoscience and Technology, University of North Carolina. His email address is mucha@unc.edu.

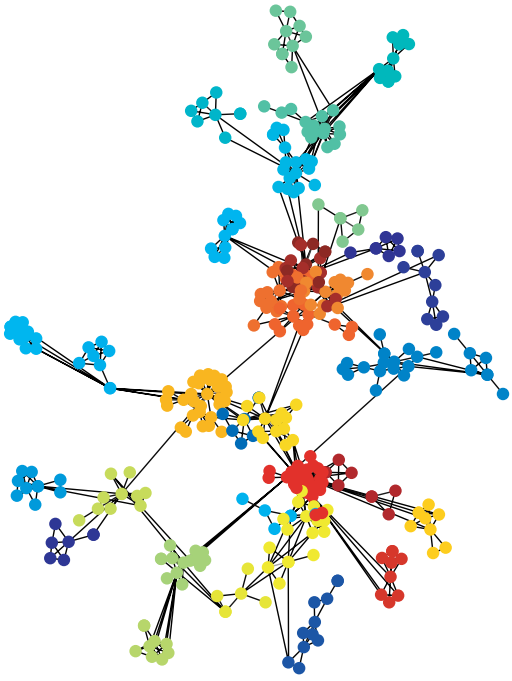

Figure 1. The largest connected component of a network of network scientists. This network was constructed based on the coauthorship of papers listed in two well-known review articles [13,83] and a small number of additional papers that were added manually [86]. Each node is colored according to community membership, which was determined using a leading-eigenvector spectral method followed by Kernighan-Lin node-swapping steps [64,86,107]. To determine community placement, we used the Fruchterman-Reingold graph visualization [45], a force-directed layout method that is related to maximizing a quality function known as "modularity" [92]. To apply this method, we treated the communities as if they were themselves the nodes of a (significantly smaller) network with connections rescaled by inter-community links. We then used the Kamada-Kawaii spring-embedding graph visualization algorithm [62] to place the nodes of each individual community (ignoring inter-community links) and then to rotate and flip the communities for optimal placement (including inter-community links). See the main text for further details on some of the ideas in this caption. (We gratefully acknowledge Amanda Traud for preparing this figure.)

One mesoscopic structure, called a *community*, consists of a group of nodes that are relatively densely connected to each other but sparsely connected to other dense groups in the network [39]. We illustrate this idea in Figure 2 using a well-known benchmark network from the sociology literature [131].

The existence of social communities is intuitively clear, and the grouping patterns of humans have been studied for a long time in both sociology [25,44,79] and social anthropology [66,113]. For example, Stuart Rice clustered data by hand to investigate political blocs in the 1920s [106], and George Homans illustrated the usefulness of rearranging the rows and columns of data matrices to reveal their underlying structure in 1950 [60]. Robert Weiss and Eugene Jacobson performed (using organizational data) what might have been the first analysis of network community structure in 1955 [126], and Herbert Simon espoused surprisingly modern views on community structure and complex systems in general in the 1960s [117]. Social communities are ubiquitous, arising in the flocking of animals and in social organizations in every type of human society: groups of hunter-gatherers, feudal structures, royal families, political and business organizations, families, villages, cities, states, nations, continents, and even virtual communities such as Facebook groups [39,88]. Indeed, the concept of community is one of everyday familiarity. We are all connected to relatives, friends, colleagues, and acquaintances who are in turn connected to each other in groups of different sizes and cohesions. The goals of studying social communities have aligned unknowingly with the statistical physics paradigm. As sociologist Mark Granovetter wrote in his seminal 1973 paper [51] on weak ties, "Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure... But how interaction in small groups aggregates to form large-scale patterns eludes us in most cases."

Sociologists recognized early that they needed powerful mathematical tools and large-scale data manipulation to address this challenging problem. An important step was taken in 2002, when Michelle Girvan and Mark Newman brought graph-partitioning problems to the broader attention of the statistical physics and mathematics communities [48]. Suddenly, community detection in networks became hip among physicists and applied mathematicians all over the world, and numerous new methods were developed to try to attack this problem. The amount of research in this area has become massive over the past seven years (with new discussions or algorithms posted on the arXiv preprint server almost every day), and the study of what has become known

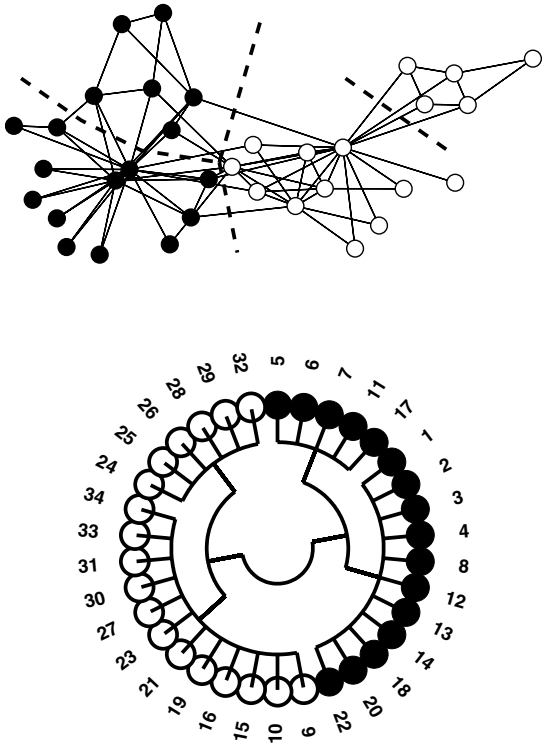

Figure 2. (Top) The Zachary Karate Club network [131], visualized using the Fruchterman-Reingold method [45]. Nodes are colored black or white depending on their later club affiliation (after a disagreement prompted the organization's breakup). The dashed lines separate different communities, which were determined using a leading-eigenvector spectral maximization of modularity [86] with subsequent Kernighan-Lin node-swapping steps (see the discussion in the main text). (Bottom) Polar coordinate dendrogram representing the results of applying this community-detection algorithm to the network. Nodes are grouped into the communities indicated in the top panel. One can see the initial split of the network into two groups (identical to the actual membership of the new clubs) by moving outward from the center of the ring. Moving farther outward culminates in the final partition of the network into four communities.

as community structure is now one of the most prominent areas of network science [26, 39, 110].

Although a rigorous notion of community structure remains elusive, one can clarify some issues through the notions of *modules* and *hierarchies*. In general, a network's community structure encompasses a potentially complicated set of hierarchical and modular components [39, 48, 117]. In this context, the term module is typically used to refer to a single cluster of nodes. Given a network that has been partitioned into nonoverlapping modules in some fashion (although some methods also allow for overlapping communities), one can continue dividing each module in an iterative fashion until each node is in its own singleton community. This hierarchical partitioning process can then be represented by a tree, or *dendrogram* (see Figure 2). Such processes can yield a hierarchy of nested modules (see Figure 3), or a collection of modules at one mesoscopic level might be obtained in an algorithm independently from those at another level. However obtained, the *community structure* of a network refers to the set of graph partitions obtained at each "reasonable" step of such procedures. Note that community structure investigations rely implicitly on using connected network components. (We will assume such connectedness in our discussion of community-detection algorithms below.) Community detection can be applied individually to separate components of networks that are not connected.

Many real-world networks possess a natural hierarchy. For example, the committee assignment network of the U.S. House of Representatives includes the House floor, groups of committees, committees, groups of subcommittees within larger committees, and individual subcommittees [100, 101]. As shown in Figure 4, different House committees are resolved into distinct modules within this network. At a different hierarchical level, small groups of committees belong to larger but less densely connected modules. To give an example closer to home, let's consider the departmental organization at a university and suppose that the network in Figure 3 represents collaborations among professors. (It actually represents grassland species interactions [23].) At one level of inspection, everybody in the mathematics department might show up in one community, such as the large one in the upper left. Zooming in, however, reveals smaller communities that might represent the department's subfields.

Although network community structure is almost always fairly complicated, several forms of it have nonetheless been observed and shown to be insightful in applications. The structures of communities and between communities are important for the demographic identification of network components and the function of dynamical processes that operate on networks (such as the spread of opinions and diseases) [39]. A community in a social network might indicate a circle of friends, a community in the World Wide Web might indicate a group of pages on closely related topics, and a community in a cellular or genetic network might be related to a functional module. In some

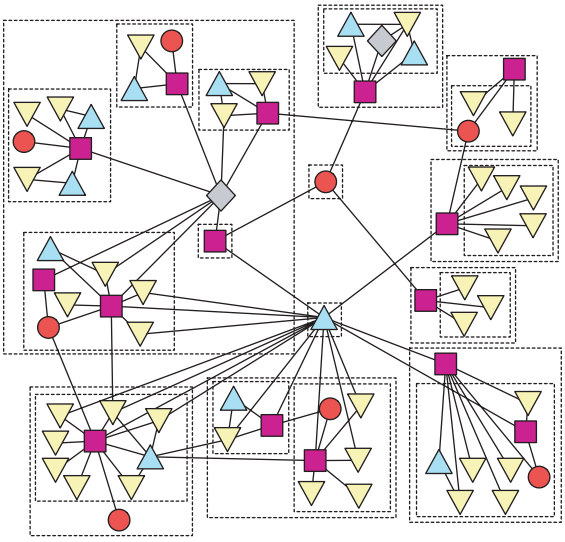

Figure 3. A network with both hierarchical and modular structure. This image, courtesy of Aaron Clauset, is an adaptation of a figure from [23].

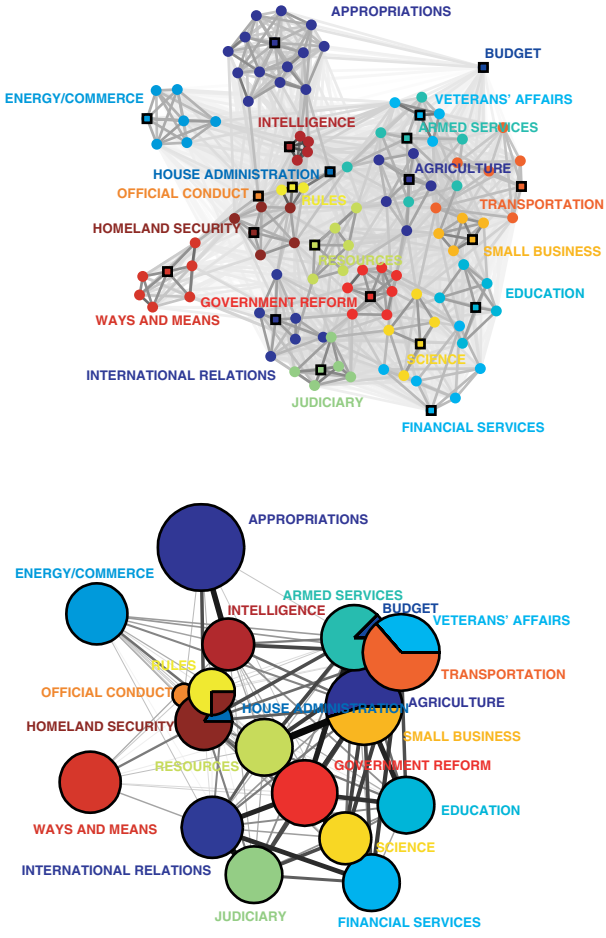

Figure 4. (Top) The network of committees (squares) and subcommittees (circles) in the 108th U.S. House of Representatives (2003–04), color-coded by the parent standing and select committees and visualized using the Kamada-Kawai method [62]. The darkness of each weighted edge between committees/subcommittees indicates how strongly they are connected. Observe that subcommittees of the same parent committee are closely connected to each other. (Bottom) Coarse-grained plot of the communities in this network. Here one can see some close connections between different committees, such as Veterans Affairs/ Transportation and Rules/Homeland Security.

cases, a network can contain several identical replicas of small communities known as motifs [75]. Consider a transcription network that controls gene expression in bacteria or yeast. The nodes represent genes or operons, and the edges represent direct transcriptional regulation. A simple motif called a "feed-forward loop" has been shown both theoretically and experimentally to perform signal-processing tasks such as pulse generation. Naturally, the situation becomes much more complicated in the case of people (doesn't it always?). However, monitoring electronically recorded behavioral data, such as mobile phone calls, allows one to study underlying social structures [49, 95]. Although these pairwise interactions (phone calls) are short in duration, they are able to uncover social groups that are persistent over time [97]. One interesting empirical finding, hypothesized by Granovetter [51], is that links within communities tend to be strong and links between them tend to be weak [95]. This structural configuration has important consequences for information flow in social systems [95] and thus affects how the underlying network channels the circulation of social and cultural resources. (See below for additional discussion.)

With methods and algorithms drawn from statistical physics, computer science, discrete mathematics, nonlinear dynamics, sociology, and other subjects, the investigation of network community structure (and more general forms of data clustering) has captured the attention of a diverse group of scientists [39, 54, 88, 110]. This breadth of interest has arisen partly because the development of community-detection methods is an inherently interdisciplinary endeavor and partly because interpreting the structure of a community and its function often requires application-specific knowledge. In fact, one aspect that makes the problem of detecting communities so challenging is that the preferred formulation of communities is often domain-specific. Moreover, after choosing a formulation, one still has to construct the desired

communities by solving an optimization problem that is known to be NP-complete in at least one family of formulations [17]. This has necessitated the adaptation of classical computational-optimization heuristics and the development of new ones.

A Simple Example

To set the stage for our survey of community-detection algorithms below, consider the ubiquitous but illustrative example of the Zachary Karate Club, in which an internal dispute led to the schism of a karate club into two smaller clubs [131]. We show a visualization of the friendships between members of the original club in Figure 2. When the club split in two, its members chose preferentially to be in the one with most of their friends. Sociologist Wayne Zachary, who was already studying the club's friendships when the schism occurred, realized that he might have been able to predict the split in advance. This makes the Zachary Karate Club a useful benchmark for community-detection algorithms, as one expects any algorithmically produced division of the network into communities to include groups that are similar to the actual memberships of the two smaller clubs.

In Figure 2, we show the communities that we obtained using a spectral-partitioning optimization of a *quality function* known as *modularity* [86]. (This method is described below.) Keeping in mind the hierarchical organization that often occurs as part of network community structure, we visualize the identified divisions using a polar coordinate dendrogram and enumerate the network's nodes around its exterior. Each distinct radius of the dendrogram corresponds to a partition of the original graph into multiple groups. That is, the community assignments at a selected level of the dendrogram are indicated by a radial cut in the bottom panel of Figure 2; one keeps only connections (of nodes to groups) that occur outside this cut. The success of the community identification is apparent in the Zachary Karate Club example, as the two main branches in the dendrogram reflect the actual memberships of the new clubs.

As shown in Figure 2, this community-detection method subsequently splits each of the two main branches. Hence, we see that the Zachary Karate Club network has a natural hierarchy of decompositions: a coarse pair of communities that correspond precisely to the observed membership split and a finer partition into four communities. In larger networks, for which algorithmic methods of investigation are especially important, the presence of multiple such partitions indicates network structures at different mesoscopic resolution levels. At each level, one can easily compare the set of communities with identifying characteristics of the nodes (e.g., the post-split Zachary Karate Club memberships) by drawing a pie chart for each community, indicating the composition of node characteristics in that community, and showing the strength of inter-community connections as ties between the pies (as in Figure 4 for Congressional committees).

Identifying Communities

Intuitively, a community is a cohesive group of nodes that are connected "more densely" to each other than to the nodes in other communities. The differences between many community-detection methods ultimately come down to the precise definition of "more densely" and the algorithmic heuristic employed to identify such sets. As different scientific fields have different needs, it is not surprising that a wide variety of community-detection methods have been developed to serve those needs [39]. These differing needs have also resulted in the deployment of different real and computer-generated benchmark networks to test community-finding algorithms [39, 70]. A 2005 review article [26] compared the performance of several of the (then-)available methods in terms of both computation time and output. A thorough, more recent discussion is available in [39].

Rather than attempting a similar comparison using every available algorithm, our aim is instead to expose a larger readership to many of the most popular methods (as well as a few of our personal favorites), while contrasting their different perspectives and revealing a few important (and sometimes surprising) similarities. Although we will highlight an extensive suite of techniques in our survey below, there are of course numerous other methods—including ones based on maximum likelihood [23], mathematical programming [1], block modeling [105, 127], link partitioning [2, 33], inference [56, 91], latent space clustering [55], and more—that we unfortunately neglect here because of space considerations. Many of them are discussed in other review articles [26, 39, 110].

Traditional Clustering Techniques

The idea of organizing data by coarse graining according to common features is a very old one [39, 119]. The original computational attempts to find clusters of similar objects are rooted in statistics and data mining. Important methods include *partitional clustering* techniques such as *k*-means clustering, *neural network clustering* techniques such as self-organizing maps, and *multidimensional scaling* (MDS) techniques such as singular value decomposition (SVD) and principal component analysis (PCA) [46]. For example, MDS algorithms of various levels of sophistication have proven to be amazingly successful at finding clusters of similar data points in myriad

applications, such as voting patterns of legislators and Supreme Court justices [99–101, 118]. Such techniques start with a matrix that indicates similarities (e.g., a tabulation of how every legislator voted on every bill) and return a coordinate matrix that minimizes an appropriate loss function. In the U.S. Congress, this allows one to see that the most important dimensions correspond to the liberal-conservative axis ("partisanship") and how well a given legislator plays with others ("bipartisanship"). During periods of heightened racial tension, such analyses have also revealed a third dimension corresponding to the division between North and South [99].

Another prominent set of classical techniques to detect cohesive groups in graphs are hierarchical clustering algorithms such as the *linkage clustering* methods used in phylogenetic biology [46, 61]. One starts with the complete set of N individual nodes in a weighted network, represented by an *adjacency matrix* A whose elements (links) A_{ij} indicate how closely nodes i and j are related to each other. For the purpose of our presentation, we will only consider undirected networks, which implies that A is symmetric (a few algorithms can also handle directed networks [39, 53, 72]). Linkage clustering is an example of an *agglomerative* method, as it starts from individual nodes and ultimately connects the entire graph. The nodes are conjoined sequentially into larger clusters, starting with the pair with maximal A_{ij} (i.e., the most strongly connected pair). At each step, one recomputes the similarities between the new cluster and each of the old clusters and again joins the two maximally similar clusters, and one continues iteratively until all clusters with nonzero similarity are connected. Different linkage clustering methods utilize different measures of the similarity between clusters. For instance, in *single linkage clustering*, the similarity of two clusters X and Y is defined as the greatest similarity between any pair of nodes $x \in X$ and $y \in Y$. Joining nodes using single linkage clustering essentially mirrors Joseph Kruskal's algorithm for computing minimum spanning trees (MSTs) [31, 50]. With clustering, however, the order of cluster formation is important and can be represented as a dendrogram, whose depths indicate the steps at which two clusters have been combined. More sophisticated techniques that build on these ideas are discussed in [46].

There are also a few classical *divisive* techniques, in which one starts with the full graph and breaks it up to find communities [39, 46, 110]. (As with agglomerative techniques, one can visualize the results using dendrograms.) The most prominent examples are spectral methods, which we discuss in detail below. New data clustering methods, which are applicable both to networks and to more general data structures, continue to be developed very actively [46, 110]. Scientists studying community detection and those studying data clustering are obviously looking at the same coin. The two fields are advancing in parallel, and there are numerous deep connections between them (including, we suspect, far more than are already known).

The Kernighan-Lin Algorithm

An algorithm from computer science, which can be combined with other methods, was proposed by Brian Kernighan and Shen Lin (KL) in 1970 in their study of how to partition electric circuits into boards so that the nodes in different boards can be linked to each other using the fewest number of connections [64]. To do this, they maximized a quality function $\tilde{Q}$ that relates the number of edges inside each group of nodes to the number between different groups. Starting with an initial partition of a graph into two groups of predefined size, KL steps swap subsets containing equal numbers of vertices between the two groups. To reduce the chance of getting stuck at a local maximum, the KL method permits swaps that decrease $\tilde{Q}$. After a specified number of swaps, one keeps the partition with maximal $\tilde{Q}$ to use as the initial condition for a new set of KL steps. When the number and sizes of communities are not specified, a natural generalization of the KL method is to move a single node at a time [12, 86, 87, 107]. Unsurprisingly, the partitions of networks into communities that are obtained using the KL algorithm depend strongly on one's initial partition and, therefore, it is best used as a supplement to high-quality partitions obtained using other methods [39, 86, 87]. In typical situations, both the KL swaps and the other method would seek to optimize the same $\tilde{Q}$.

Centrality-Based Community Detection

Michelle Girvan and Mark Newman generated a great deal of attention in mathematics and statistical physics for network community structure in [48] when they devised a community-finding algorithm based on the sociological notion of *betweenness centrality* [4, 42, 124]. An edge has high betweenness if it lies on a large number of short paths between vertices. (Note that betweenness can also be defined for nodes.) If one starts at a node and wants to go to some other node in the network, it is clear that some edges will experience a lot more traffic than others. The betweenness of an edge quantifies such traffic by considering strictly shortest paths (*geodesic betweenness*) or densities of random walks (*random walk betweenness*) [85] between each pair of nodes and taking into account all possible pairs. One can identify communities through a process of ranking all of the edges based on their betweenness, removing the edge with the largest value, and recalculating

betweenness for the remaining edges. The recalculation step is important because the removal of an edge can cause a previously low-traffic edge to have much higher traffic. An iterative implementation of these steps gives a divisive algorithm for detecting community structure, as it deconstructs the initial graph into progressively smaller connected chunks until one obtains a set of isolated nodes.

Betweenness-based methods have been generalized to use network components other than edges, to bipartite networks [100], and to use other sociological notions of centrality [39]. Note, however, that although centrality-based community detection is intuitively appealing, it can be too slow for many large networks (unless they are very sparse), and it tends to give relatively poor results for dense networks.

k-Clique Percolation and Other Local Methods

The method of *k-clique percolation* [97] is based on the concept of a *k-clique*, which is a complete subgraph of k nodes that are connected with all $k(k-1)/2$ possible links. The method relies on the observation that communities seem to consist of several small cliques that share many of their nodes with other cliques in the same community. A *k-clique community* is then defined as the union of all "adjacent" k-cliques, which by definition share $k-1$ nodes. One can also think about "rolling" a k-clique template from any k-clique in the graph to any adjacent k-clique by relocating one of its nodes and keeping the other $k-1$ nodes fixed [28]. A community, defined through the percolation of such a template, then consists of the union of all subgraphs that can be fully explored by rolling a k-clique template. As k becomes larger, the notion of a community becomes more stringent. Values of $k = 3, \ldots, 6$ tend to be most appropriate because larger values become unwieldy. The special case of $k = 2$ reduces to bond (link) percolation and $k = 1$ reduces to site (node) percolation.

The k-clique percolation algorithm is an example of a *local* community-finding method. One obtains a network's global community structure by considering the ensemble of communities obtained by looping over all of its k-cliques. Some nodes might not belong to any community (because they are never part of any clique), and others can belong to several communities (if they are located at the interface between two or more communities). The nested nature of communities is recovered by considering different values of k, although k-clique percolation can be too rigid because focusing on cliques typically causes one to overlook other dense modules that aren't quite as well connected. On the other hand, the advantage of k-clique percolation and other local methods is that they have to date provided some of the most successful ways to consider community overlap.

Allowing the detection of network communities that overlap is especially appealing in the social sciences, as people belong simultaneously to several communities (constructed via colleagues, family, hobbies, etc.) [78,79]. Purely agglomerative or divisive techniques do not allow communities to overlap, so it is important to consider local methods as well. Several such methods have now been developed [2, 7, 8, 14, 22, 28, 33, 39, 74, 114, 129], including one that enables the consideration of overlapping communities at multiple resolution levels [69]. We believe that further development of global clustering algorithms that take community overlap explicitly into account is essential to complement the insights from these local approaches.

Modularity Optimization

One of the most popular quality functions is *modularity*, which attempts to measure how well a given partition of a network compartmentalizes its communities [84, 86, 87, 89, 90]. The problem of optimizing modularity is equivalent to an instance of the famous MAX-CUT problem [86], so it is not surprising that it has been proven to be NP-complete [17]. There are now numerous community-finding algorithms that try to optimize modularity or similarly constructed quality functions in various ways [6, 12, 26, 39]. In the original definition of modularity, an unweighted and undirected network that has been partitioned into communities has modularity [84, 90]

$$Q = \sum_i (e_{ii} - b_i^2), \tag{1}$$

where e_{ij} denotes the fraction of ends of edges in group i for which the other end of the edge lies in group j, and $b_i = \sum_j e_{ij}$ is the fraction of all ends of edges that lie in group i. Modularity is closely related to the Freeman segregation index [43]; a key difference is that $Q = 0$ when all nodes are assigned to the same community, which enforces the existence of a nontrivial partition with $Q > 0$. Modularity explicitly takes degree heterogeneity into account, as it measures the difference between the total fraction of edges that fall within groups versus the fraction one would expect if edges were placed at random (respecting vertex degrees).[1] Thus, high values of Q indicate network partitions in which more of the edges fall within groups than expected by chance (under a specified null model, as discussed further below). This, in turn, has been

[1] *Interest in degree heterogeneity exploded in the late 1990s with the sudden wealth of empirical data and the seemingly ubiquitous manifestation of heavy-tailed degree distributions such as power laws [3, 83].*

found to be a good indicator of functional network divisions in many cases [86, 87].

For weighted networks, one counts the sums of the weights of edges instead of the number of edges, so heavily weighted edges contribute more than lightly weighted ones. Both e_{ij} and b_i are thus straightforwardly generalized, and then the modularity is again calculated from (1). The meaning of modularity remains essentially the same: It measures when a particular division of the network has more edge weight within groups than one would expect by chance.

Quality functions such as modularity provide precise measures of how to count the total strength of connections within communities versus those between communities [87, 89]. Modularity is a scaled *assortativity* measure based on whether high-strength edges are more or less likely to be adjacent to other high-strength edges [84, 89, 90]. Because communities are supposed to have high edge density relative to other parts of the graph, a high-modularity partition tends to have high edge-strength assortativity by construction. More generally, assortativity notions can be used to partition a graph into groups according to any characteristic by examining whether nodes are more likely (in *assortative* graphs) or less likely (in *disassortative* graphs) to be connected to nodes of the same type [83].

Interestingly, maximizing modularity is closely related to the energy models of pairwise attraction, such as the Fruchterman-Reingold method, that are commonly used for graph visualization [92]. While this isn't necessarily surprising given the clusters that one can typically observe with good graph visualization tools, this recent insight does suggest that such tools may also help lead to the development of better community-detection methods. Conversely, the analysis and construction of algorithms to find network communities might help lead to better graph-visualization techniques.

It is typically impossible computationally to sample a desired quality function by exhaustively enumerating the nonpolynomial number of possible partitions of a network into communities [17]. A number of different methods have thus been proposed to balance the typical quality of their identified optima with the computational costs. Some methods, such as the greedy algorithms in [24, 84], are fast heuristics intended to be applied to networks with millions of nodes or more. Other methods—such as spectral partitioning [86, 87] (discussed below), refined greedy algorithms [111], simulated annealing [52], extremal optimization [30], and others [93]—provide more sophisticated but slower means to identify high-modularity partitions.[2] We discuss the spectral partitioning method below, in part because of its interesting reformulation of the modularity scalar as a matrix, but we note that other algorithmic choices might be superior in many situations. We believe that there is considerable value in having multiple computational heuristics available, as this provides greater flexibility to compare and contrast the identified communities.

Importantly, many modularity-maximization techniques are easy to generalize for use with other related quality functions because it is far from clear that modularity is the best function to optimize. For example, modularity has a known *resolution limit* (see below) that might cause one to miss important communities [38]. A few alternatives to modularity have been considered [6, 39, 57, 69, 121], and it is ultimately desirable to optimize a quality function that includes not only information about the network structure but also other information (such as node characteristics or relevant time-dependence) that would allow one to incorporate functionality directly [116]. Such consideration of additional information is one of the most important open issues in community detection [88, 122].

Spectral Partitioning

The method of spectral partitioning arose most prominently in the development of algorithms for parallel computation [35, 102]. In traditional spectral partitioning, network properties are related to the spectrum of the graph's Laplacian matrix L, which has components $L_{ij} = k_i\delta(i,j) - A_{ij}$, where k_i is the degree of node i (or, in a weighted network, its strength), and $\delta(i,j)$ is the Kronecker delta (i.e., $\delta(i,j) = 1$ if $i = j$ and 0 otherwise).

The simplest such method starts by splitting a network into two components. One then applies two-group partitioning recursively to the smaller networks one obtains as long as it is desirable to do so. (One can also partition networks into more than two groups during each step [21, 29, 86, 107].) For a single partitioning step, one defines an index vector s whose components take the value $+1$ if they belong to group 1 and -1 if they belong to group 2. The total edge weight between the two groups of nodes can then be expressed as $R = \frac{1}{4}s^T Ls$. The "best" partition of the network seemingly results from choosing s to minimize R (called the "cut size") and hence the total strength of inter-group edges. (Recall the max-flow min-cut theorem, which states that the minimum cut between any two vertices of a graph—that is, the minimum set of edges whose deletion places the

[2] *As we have discussed, one can also supplement any of these methods with KL swapping steps [12, 64, 86, 87, 93, 107].*

two vertices in disconnected components of the graph—carries the maximum flow between the two vertices [32,37].) Unfortunately, this minimization is easily accomplished by choosing the trivial (and useless) partition of a single group containing every node. The most common solution to this situation is to fix the sizes of the two groups in advance and incorporate this information in the partitioning procedure (as described in, e.g., [86]). This solution is perfectly reasonable for some applications, such as load balancing in parallel computing. However, this approach is neither appropriate nor realistic for community detection in most other contexts because one typically does not know the number or sizes of communities in advance, so choosing arbitrary sizes at the outset precludes attacking the main problem of interest.

Fortunately, one can use the idea of modularity to obtain spectral partitioning algorithms that are appropriate for a broader class of problems [87] (see also the earlier publication [29] and a spiritually similar approach based on peer influences in the sociology literature [77]). By reformulating the scalar quantity of modularity in terms of a *modularity matrix* B, with components

$$B_{ij} = A_{ij} - P_{ij}\,, \tag{2}$$

spectral partitioning can be directly applied [87] as a means of heuristically optimizing the modularity

$$Q = \frac{1}{2W}\sum_{i,j} B_{ij}\delta(C_i, C_j)\,, \tag{3}$$

where $\delta(C_i, C_j)$ indicates that the B_{ij} components are only summed over cases in which nodes i and j are assigned to the same community. The factor $W = \frac{1}{2}\sum_{ij} A_{ij}$ is the total edge strength in the network (equal to the total number of edges for unweighted networks), where k_i denotes the strength of node i. In (2), P_{ij} denotes the components of a *null model* matrix, which specifies the relative value of intra-community edges in assessing when communities are tightly connected [9, 86]. In general, one is free to specify any reasonable null model. The most popular choice, proposed by Newman and Girvan [86,87,89,90], is

$$P_{ij} = \frac{k_i k_j}{2W}\,. \tag{4}$$

This recovers the definition of modularity in (1), specified in terms of edge-weight deviations from a network chosen randomly from the set of all graphs with the same expected strength distribution as the actual network. This null model is closely related to the configuration model [76], which (as with Erdös-Renyi random graphs) yields networks that aren't expected to have a natural hierarchy [76, 83, 90]. The difference is that (4) is conditioned on the expected degree (or strength) sequence, whereas the configuration model is conditioned on the actual observed sequence.

In spectral partitioning, one can use as many eigenvectors of B as there are positive eigenvalues, but it is effective (and simplest) to recursively subdivide a network using only the "leading eigenvector" v, which is paired with the largest positive eigenvalue of B. One can then separate the network into two communities according to the signs $s_i = \mathrm{sgn}(v_i)$. The magnitude of v_i indicates the strength to which the ith node belongs to its assigned community [86]. For $v_i = 0$, one can assign node i to a community based on which choice produces the higher modularity, changing $s_i = 0$ to $+1$ or -1 as appropriate. The modularity of the resulting two-group partition of the network is $Q = \frac{1}{4W}s^T B s$. After this bipartition, one then repeats the partitioning procedure for each graph component, keeping track of the fact that they are actually part of a larger network. One continues recursively until the modularity can no longer be increased with additional subdivisions [86, 87]. The final network partition gives the community structure at a specific resolution level (e.g., committees in the U.S. House of Representatives committee assignment network). This method can be generalized by considering different quality functions [39, 107], allowing steps that decrease global quality in order to further subdivide the communities [107, 133], using more eigenvectors [107, 123], or including a resolution parameter [6, 104] that allows one to examine the network's community structure at different mesoscopic scales.

The Potts Method

Particles that possess a magnetic moment are often called *spins* [112, 130]. Such spins interact with other spins either *ferromagnetically* (they seek to align) or *antiferromagnetically* (they seek to have different orientations). A *spin glass* is a system that encompasses both disorder and competing ferromagnetic and antiferromagnetic interactions. This leads to a very large number of metastable spin configurations separated by energy barriers with long, glass-like characteristic relaxation times [36, 130]. An important recent insight, inspired by earlier work on data clustering based on the physical properties of an inhomogeneous ferromagnetic model [11], is that optimizing modularity is mathematically equivalent to minimizing the energy (i.e., finding the ground state of the Hamiltonian) of an infinite-range q-state Potts model [103, 104].

In a q-state Potts spin glass, each spin can be in one of q states. The interaction energy between spins i and j is given by $-J_{ij}$ if the spins are in the same state and zero if they are not [104, 130]. The Hamiltonian of the system is given by the sum

over all of the pairwise interaction energies:

$$H(\{\sigma\}) = -\sum_{ij} J_{ij}\delta(\sigma_i, \sigma_j)\,, \tag{5}$$

where σ_l indicates the state of spin l and $\{\sigma\}$ denotes the configuration of spins (i.e., the state of each of the system's N spins). There are a total of q^N such configurations.

We map the problem of minimizing (5) to a network community-detection problem by assigning a spin to each node and letting $q = N$. In this language, one adds the interaction energy $-J_{ij}$ if and only if nodes i and j are placed in the same community. Two nodes that are connected by an edge interact ferromagnetically ($J_{ij} > 0$) when the weight of the edge is greater than expected (given a specific null model) and interact antiferromagnetically ($J_{ij} < 0$) when it is less than expected. If $J_{ij} = 0$, spins i and j do not interact with each other. Hence, two nodes want to be in the same community if they interact ferromagnetically and in different ones if they interact antiferromagnetically. One cannot in general find any arrangement of spins (i.e., any partition of a graph into communities) that simultaneously minimizes all of the pairwise interaction energies. Regardless of this inability to satisfy all of the bonds simultaneously, a phenomenon termed "frustration" [36], one can still try to minimize H globally to find the ground state of the system. The choice of interaction matrix elements given by

$$J_{ij} = \frac{A_{ij} - P_{ij}}{W} \tag{6}$$

implies that $H = -Q$ and recovers modularity maximization [104]. (Division by W is a normalization and does not affect the optimization algorithms.) Alternative interaction models can also be used to partition networks (see, e.g., [57]).

Resolution Parameters

In 2007, Santo Fortunato and Marc Barthélemy demonstrated, using both real and computer-generated networks, that modularity suffers from a resolution limit in its original formulation [38], as it misses communities that are smaller than a certain threshold size that depends on the size of the network and the extent of interconnectedness of its communities. Communities smaller than the threshold tend to be merged into larger communities, thereby missing important structures. We have seen this in our own work on the U.S. House committee assignment network, as detecting communities by maximizing modularity can group multiple standing committees (with their subcommittees) into a single community [100].

One way to address this resolution limit is to incorporate an explicit *resolution parameter* directly into equations like (6) to obtain [104]

$$J_{ij} = \frac{A_{ij} - \lambda P_{ij}}{W}\,. \tag{7}$$

One can alternatively incorporate a resolution parameter into A_{ij} or elsewhere in the definition of a quality function (see, e.g., [6]). This allows one to zoom in and out in order to find communities of different sizes and thereby explore both the modular and the hierarchical structures of a graph. Fixing λ in (7) corresponds to setting the scale at which one is examining the network: Larger values of λ yield smaller communities (and vice versa). Resolution parameters have now been incorporated (both explicitly and implicitly) into several methods that use modularity [12, 104], other quality functions [6, 69], and other perspectives [8, 97].

Although introducing a resolution parameter using equations like (7) seems *ad hoc* at first, it can yield very interesting insights. For example, $J_{ij} = (A_{ij} - \lambda)/W$ gives a uniform null model in which a given fixed average edge weight occurs between each node. This can be useful for correlation and similarity networks, such as those produced from matrices of yea and nay votes. Nodes i and j want to be in the same community if and only if they voted the same way more than some threshold fraction of times that is specified by the value of λ.

Even more exciting, one can relate resolution parameters to the time scales of dynamical processes unfolding on a network [27, 68, 98, 108]. Just as we can learn about the behavior of a dynamical system by studying the structural properties of the network on which it is occurring, we can also learn about the network's structural properties by studying the behavior of a given dynamical process. This suggests the intuitive result that the choice of quality function should also be guided by the nature of the dynamical process of interest. In addition to revealing that resolution parameters arise naturally, this perspective shows that the Potts method arises as a special case of placing a continuous time random walk with Poisson-distributed steps on a network [27]. Freezing the dynamics at a particular point in time yields the modularity-maximizing partition. Freezing at earlier times yields smaller communities (because the random walker hasn't explored as much of the graph), and waiting until later times results in larger communities. The $t \to \infty$ limit reproduces the partitioning from Miroslav Fiedler's original spectral method [35].

Applications

Armed with the above ideas and algorithms, we turn to selected demonstrations of their efficacy. The increasing rapidity of developments in network community detection has resulted in part

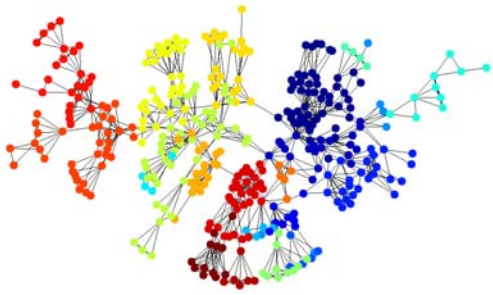

Figure 5. The largest connected component (379 nodes) of the network of network scientists (1,589 total nodes), determined by coauthorship of papers listed in two well-known review articles [13,83] and a small number of papers added manually [86]. Each of the nodes in the network, which we depict using a Kamada-Kawai visualization [62], is colored according to its community assignment using the leading-eigenvector spectral method [86].

from the ever-increasing abundance of data sets (and the ability to extract them, with user cleverness). This newfound wealth—which includes large, time-dependent data sets—has, in turn, arisen from the massive amount of information that is now routinely collected on websites and by communication companies, governmental agencies, and others. Electronic databases now provide detailed records of human activity patterns, offering novel avenues to map and explore the structure of social, communication, and collaboration networks. Biologists also have extensive data on numerous systems that can be cast into network form and which beg for additional quantitative analyses.

Because of space limitations, we restrict our discussion to five example applications in which community detection has played a prominent role: scientific coauthorship, mobile phone communication, online social networking sites, biological systems, and legislatures. We make no attempt to be exhaustive for any of these examples; we merely survey research (both by others and by ourselves) that we particularly like.

Scientific Collaboration Networks

We know from the obsessive computation of Erdös numbers that scientists can be quite narcissistic. (If you want any further evidence, just take a look at the selection of topics and citations in this section.) In this spirit, we use scientific coauthorship networks as our first example.

A bipartite (two-mode) coauthorship network—with scientists linked to papers that they authored or coauthored—can be defined by letting $\delta_i^p = 1$ if scientist i was a coauthor on paper p and 0 otherwise. Such a network was collected and examined from different databases of research papers in [80–82]. To represent the collaboration strength between scientists i and j, one can define

$$A_{ij} = \sum_p \frac{\delta_i^p \delta_j^p}{n_p - 1} \tag{8}$$

as the components of a weighted unipartite (one-mode) network, where n_p is the number of authors of paper p and the sum runs over multiple-author papers only. Applying betweenness-based community detection to a network derived from Santa Fe Institute working papers using (8) yields communities that correspond to different disciplines [48]. The statistical physics community can then be further subdivided into three smaller modules that are each centered around the research interests of one dominant member. Similar results have been found using various community-finding algorithms and numerous coauthorship networks, such as the network of network scientists [86] (see Figures 1 and 5), which has become one of the standard benchmark examples in community-detection papers.

Mobile Phone Networks

Several recent papers have attempted to uncover the large-scale characteristics of communication and social networks using mobile phone data sets [49, 95, 96]. Like many of the coauthorship data sets studied recently [96], mobile phone networks are longitudinal (time-dependent). However, in contrast to the ties in the coauthorship networks above, links in phone networks arise from instant communication events and capture relationships as they happen. This means that at any given instant, the network consists of the collection of ties connecting the people who are currently having a phone conversation. To probe longer-term social structures, one needs to aggregate the data over a time window.

In 2007 one research group used a society-wide communication network containing the mobile phone interaction patterns of millions of individuals in an anonymous European country to explore the relationship of microscopic, mesoscopic, and macroscopic network structures to the strength of ties between individuals on a societal level [95]. They obtained very interesting insights into Mark Granovetter's famous *weak tie hypothesis*, which states that the relative overlap of the friendship circles of two individuals increases with the strength

of the tie connecting them [51]. At the mesoscopic level, this leads to a structure in which individuals within communities tend to be linked via strong ties, whereas communities tend to be connected to other communities via weak ties. Because of this coupling between link strength and function, the weak ties are responsible for the structural integrity of the communication network: It is robust to the removal of the strong ties but breaks up if the weak ties are removed (see Figure 6). In fact, one can even show that the removal of weak ties leads to a (phase) transition from a regime in which the network remains globally connected to one in which the network essentially consists of insular communities. However, there is no phase transition if the strong ties are removed, so the network remains globally connected. The location of the transition also suggests a natural quantitative demarcation between weak and strong ties. This mesoscopic organization of social networks has important consequences for the flow of information. If one assumes that every tie (regardless of strength) is equally efficient in transferring information, one recovers the classical result of Granovetter that weak ties are mostly responsible for information diffusion [51]. However, if one assumes that the rate of information transfer is proportional to the strength of the tie, then neither weak nor strong ties are as effective as intermediate ties for information diffusion [95].

To help develop methods that can be applied to time-dependent networks, another research group has recently applied k-clique percolation to a large mobile phone data set to investigate community formation, evolution, and destruction [96]. They found that large communities persist longer if they are capable of dynamically altering their membership (suggesting that an ability to change the group composition results in better adaptability), whereas small groups last longer if they stay virtually unchanged. We've all seen examples of such dynamics in real life: A mathematics department at a university will last a long time and will still be identified as fundamentally the same community even though its membership can change quite drastically over time. On the other hand, an individual research group might rely on only one or two of its members for its existence.

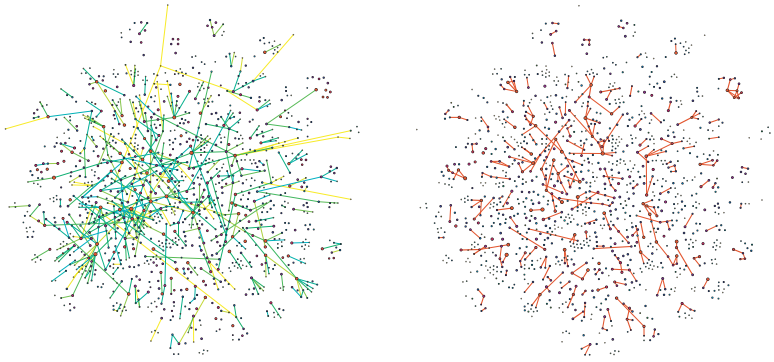

Figure 6. (Left) A sample of a mobile phone network studied in [94,95]. After the strong ties have been removed, the network still retains its global connectivity. (Right) Removal of weak ties leads, through a phase transition, to a disintegration of the network. (Figure adapted from [94].)

Online Social Networks

Social networking sites (SNSs) have become a pervasive part of everyday life. They allow users to construct a public or semi-public online profile within a bounded system, articulate a list of other users (called "friends") with whom they share a connection, and view and traverse their network of connections [16, 122]. Since their introduction, SNSs such as Facebook, LinkedIn, MySpace, and hundreds of others have collectively attracted over one billion users [128]. People have integrated SNSs into their daily lives, using them to communicate with friends, send emails, solicit opinions or votes, organize events, spread ideas, find jobs, and more.

The emergence of SNSs has also revolutionized the availability of quantitative social and demographic data, which has in turn transformed the study of social networks [16]. This has obviously been a big deal for social scientists (see [73] for an excellent recent example), but numerous mathematicians, computer scientists, physicists, and more have also had a lot of fun with this new wealth of data. This has included investigations of attachment mechanisms to determine how SNS network structure might develop [67], the formation of friends, allies, and nemeses [18, 59], and much more [16, 122].

In a recent paper [122], we used anonymous Facebook data from September 2005 to compare the friendship networks of several American universities. This data yields networks in which each node represents a person and each edge represents a self-identified, reciprocal online friendship. The institutions we considered ranged from small technical universities such as Caltech (1,089 nodes) to large state universities such as the University of Oklahoma (about 24,000 nodes). Our data also includes limited demographic information for each user: gender, high school, class year, major, and dormitory residence. Using permutation tests, we developed tools that allow one to quantitatively compare different network partitions, which can be obtained from any desired method—including, in particular, community-detection algorithms and user demographics. This enables one to investigate the demographic organization of different universities and compare the results of different community-detection algorithms. We found, for

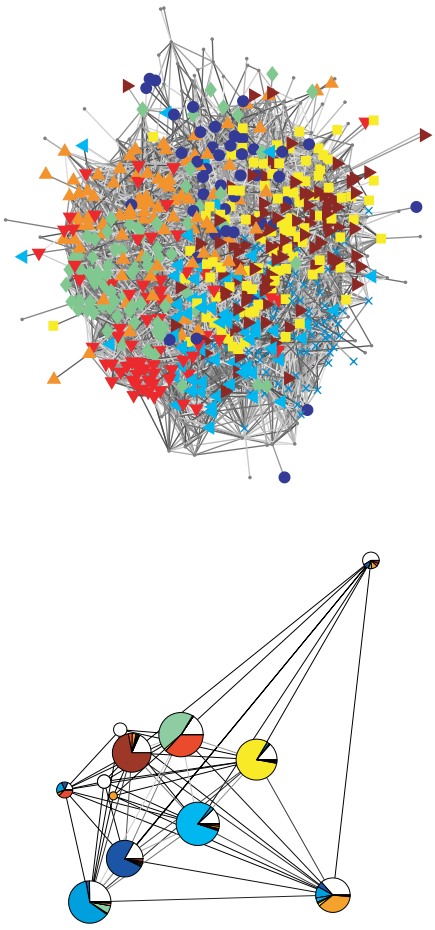

Figure 7. (Top) Caltech friendship network based on Facebook data from September 2005 using the Fruchterman-Reingold visualization method [45]. The colors (and shapes) correspond to self-identified House (dormitory) affiliation. (Bottom) Communities in the Caltech network. We obtained this community structure, which accurately matches the Caltech House structure, using a slight modification of Newman's leading-eigenvector algorithm [86] and subsequent KL node-swapping steps [107,122,133]. It also gives some indication of the most probable House affiliations of the people in white, who did not identify their House [122].

example, that communities at Princeton, Georgetown, and the University of North Carolina at Chapel Hill are organized predominantly by class year, whereas those at Caltech are based almost exclusively on House (dormitory) affiliation. As we illustrate in Figure 7, community structure can also be used to make simple yet intelligent guesses about withheld user characteristics. Naturally, this opens up significant privacy issues when data is not fully anonymous.

Biological Networks

One of the paramount goals of studying biological networks is to determine the principles governing their evolution. It is hoped that finding important network structures might give some insights into the mechanisms (and, ideally, the natural design principles) that control the flow of biological information. Thus, it is unsurprising that clustering methods form a core part of bioinformatics [15,63], and there is simply no way to do justice to this vast literature here. Accordingly, we only present a couple of our favorite examples.

In 2002 Ron Milo et al. investigated a plethora of directed networks to develop the idea of miniature communities known as *motifs* [75], which are used to describe patterns of nodes and edges that occur significantly more prevalently than expected in a network. Motifs can be interpreted as basic building blocks of complex networks, perhaps representing small functional modules that arose via evolutionary mechanisms. The simplest motifs—triangles (3-cliques), in which three nodes are all mutually interconnected—have long been studied in social networks [44,124]. The amazing discovery of Milo et al. is that different types of motifs are, in fact, prevalent universally in many different types of networks. Among the omnipresent motifs they observed are 3-chains in ecological networks (in which a prey node is connected by a directed edge to a predator node, which is in turn connected by a directed edge to another predator); diamonds in ecological networks, neural networks, and logic chips; feed-forward loops in gene regulation networks, neural networks, and logic chips; triangles in the World Wide Web; and more.

Numerous scientists have built on this foundation of motifs, and several investigations have provided fascinating connections between motifs and larger mesoscopic structures. For example, one team investigated three-node and four-node motifs in an integrated *Saccharomyces cerevisiae* network, which they constructed using information from protein interactions, genetic interactions, transcriptional regulation, sequence homology, and expression correlation [132]. Their primary finding was that most motifs form larger "network themes" of recurring interconnection patterns that incorporate multiple motif occurrences. They were also able to tie some of these mesoscopic themes to specific biological phenomena, such as a pair of protein complexes with many genetic interactions between them. The notion of motifs has also recently been used to develop generalizations of graph modularity [5].

One can imagine constructing a course-grained network consisting of interconnected network

themes. For example, in a seminal 2005 paper, Roger Guimerà and Luís Amaral used communities to try to construct a "functional cartography" of biological networks in order to employ nodes and modules with known roles to try to obtain interesting insights on nodes and modules with unknown functionality [52]. To understand their perspective, consider the prototypical maps of countries in which important cities are marked by special symbols, other cities are marked with dots, and smaller towns are not shown. In the network context, there is a one-to-one correspondence between markers and communities, and the symbols are determined according to structural and/or functional roles. The connections between communities are analogous to major highways. In fact, our coloring of community ties according to the demographic composition of their nodes (see Figures 4 and 7) was originally inspired by [52].

To illustrate their idea, Guimerà and Amaral considered the metabolic networks of twelve different organisms. They started by detecting communities by maximizing modularity using simulated annealing and then (following a suggestion from social scientist Brian Uzzi) calculated appropriate properties of each node to determine their role in their assigned community. After finding communities, they calculated for each node i the *within-module degree*, given by the number of its edges that connect to other nodes in the same community, and a *participation ratio* P_i, which measures the extent to which links are distributed among all network communities. Guimerà and Amaral then interpreted the role of each node based on its location in the plane determined by P_i and the z-score z_i of the within-module degree. They thereby found that "non-hub connector nodes" (which have low z_i and moderately high P_i, indicating a preferential connectivity to a subset of the network's communities) are systematically more conserved across species than "provincial hubs" (which have high z_i and low P_i). This appears to be related to the role of non-hub connectors in describing the global structure of fluxes between different network modules. Importantly, one can follow a similar procedure using other measures, such as betweenness centrality [34], as the essential insight—which, we stress, *was borrowed from ideas in the social sciences*—is to calculate network quantities relative to community assignment.

Legislative Networks

Advances in network science have also begun to uncover the ways in which social relationships shape political outcomes [40, 41, 101]. In this section, we describe our own work on legislative networks [100, 101, 125, 133], in which community detection has played a central role.

Consider a bipartite graph that is constructed using Representatives and their committee and subcommittee (henceforth called simply "committee") assignments during a single two-year term of the U.S. House of Representatives. Each edge represents a committee assignment and connects a Representative to a committee. We project each such network onto a weighted unipartite graph of committees (see Figure 4), in which the nodes are now committees and the value of each edge gives the normalized connection strength between two committees. By computing the community structure of these networks and analyzing legislator ideology, we investigated correlations between the political and organizational structure of House committees. This revealed close ties between the House Rules Committee and the Select Committee on Homeland Security in the 107th (2001–02) and 108th (2003–04) Congresses that broke the established procedures for determining the composition of select committees [100, 101]. (Figure 4 shows the 108th Congress.) We also showed that the modularity of good network partitions increased following the 1994 elections, in which the Republican party earned majority status in the House for the first time in more than forty years.

Studying networks constructed from legislation cosponsorship can help uncover social connections between politicians, as legislators who work together closely on pieces of legislation are likely to have friendly (or at least cordial) relations. Computing centrality measures in these networks gives a who's who list of American politics, as it reveals important players like Bob Dole [R-KA], John McCain [R-AZ], and Ted Kennedy [D-MA] [40]. The longitudinal study of community structure in Congressional legislation cosponsorship [133] and roll-call voting [125] networks shows that graph modularity can be used to study partisan polarization and political party realignments. This reveals patterns suggesting that political parties were not the most significant communities in Congress for certain periods of U.S. history and that the 1994 party-changing elections *followed* a rise in partisan polarization rather than themselves leading to an abrupt polarization in America.

Summary and Outlook

With origins in sociology, computer science, statistics, and other disciplines, the study of network communities is in some respects quite old. Nevertheless, it has experienced incredible growth since the seminal 2002 paper [48] that brought greater attention to the problem, particularly among statistical physicists [39]. In this survey, we have highlighted an extensive suite of techniques, and there are numerous other methods that we simply

haven't had space to discuss (see the review articles [26, 39, 110] for more information on many of them). Despite this wealth of technical advances, much work remains. As Mark Newman recently wrote [88], "The development of methods for finding communities within networks is a thriving sub-area of the field, with an enormous number of different techniques under development. Methods for understanding what the communities mean after you find them are, by contrast, still quite primitive, and much needs to be done if we are to gain real knowledge from the output of our computer programs." One of our primary purposes in writing this article is as a "call to arms" for the mathematics community to be a part of this exciting endeavor. Accordingly, we close our discussion with additional comments about important unresolved issues.

The remarkable advances of the past few years have been driven largely by a massive influx of data. Many of the fascinating networks that have been constructed using such data are enormous (with millions of nodes or more). Given that optimization procedures, such as maximizing graph modularity, have been proven to be NP-complete [17], much of the research drive has been to formulate fast methods that still find a reasonable community structure. Some of the existing algorithms scale well enough to be used on large networks, whereas others must be restricted to smaller ones. The wealth of data has also led to an increasing prevalence (and, we hope, cognizance) of privacy issues. However, although the study of network communities has become so prominent, this research area has serious flaws from both theoretical and applied perspectives: There are almost no theorems, and few methods have been developed to use or even validate the communities that we find.

We hope that some of the mathematically-minded *Notices* readers will be sufficiently excited by network community detection to contribute by developing new methods that address important graph features and make existing techniques more rigorous. When analyzing networks constructed from real-world data, the best practice right now is to use several of the available computationally tractable algorithms and trust only those structures that are similar across multiple methods in order to be confident that they are properties of the actual data rather than byproducts of the algorithms used to produce them. Numerous heuristics and analytical arguments are available, but there aren't any theorems, and even the notion of community structure is itself based on the methodology selected to compute it. There also appear to be deep but uncharacterized connections between methods that have been developed in different fields [39, 110]. Additionally, it would be wonderful if there were a clearer understanding of which notions of community and which community-detection methods might be especially appropriate for networks with specific properties and for networks belonging to a specific domain.

At the same time, the problem of how to validate and use communities once they are identified is almost completely open. Fortunately, recent work offers some hope, as new methods have been developed to infer the existence of missing edges from network data [23] and relate the composition of communities to intelligent guesses about the demographic characteristics of nodes [122]. (As with social networks more generally, sociologists have already been considering these issues for a long time [19, 44]. What we need are techniques that allow us to do this even more effectively.) In [23], Aaron Clauset, Cris Moore, and Mark Newman drew on the insight that real-world networks should not be expected to have a unique community structure (despite the output produced by almost all of the available methods) and formulated a new *hierarchical random graph model* to develop a method for inferring hierarchical structure from network data. (A different hierarchical random graph model was formulated for community detection in [109].) Their method, which shows excellent promise for future development, allowed them to make accurate predictions about missing connections in partially known networks. In our own work on Facebook networks [122], we used permutation tests to advance methods for the quantitative comparison of network partitions. Because one can obtain such partitions either from algorithmic community-detection methods or by hand (from external demographics or other properties), this provides a mechanism to compare the results of different community-finding algorithms and to try to infer node characteristics given partial demographic information.

It is also important to develop community-detection techniques that can be applied to more complicated types of graphs. As we saw in our discussion of legislative and coauthorship networks, collaboration networks have a bipartite structure. However, there has been seemingly only limited work thus far on community finding that works directly on bipartite networks rather than on their unipartite projections [10, 53, 71, 100]. Even fewer community-detection methods are able to handle directed networks (whose adjacency matrices are asymmetric) [53, 72] or signed networks (whose connections might be construed as "attracting" or "repelling") [121]. Moreover, agents in social networks are typically connected in several different manners—for example, Representatives can be connected using voting similarities, common committee assignments, common financial contributors, and more—but there are presently very

few algorithms that can be applied to such multiplex situations without constructing individual graphs for each category, and further development will likely require the application of ideas from multilinear algebra [65, 115]. It would also be desirable to detect communities in hypergraphs and to be able to consider connections between agents that are given by interval ranges rather than precise values. Finally, to be able to study interactions between dynamical processes on networks and the structural dynamics of networks themselves (e.g., if somebody spends a day at home with the flu, the network structure in the workplace is different than usual that day), a lot more work is needed on both overlap between communities and on the community structure of time-dependent and parameter-dependent networks. Analyzing time- and parameter-dependent networks currently relies on *ad hoc* amalgamation of different snapshots rather than on a systematic approach, so it is necessary to develop community-detection methods that incorporate the network structure at multiple times (or parameter values) simultaneously [58, 96, 116]. More generally, this will also have important ramifications for studies of clustering in correlated time series.

We stress that research on network communities has focused on using exclusively structural information (i.e., node connectivity and link weights) to deduce *structural communities* as imperfect proxies for *functional communities* [39, 116, 122]. While this seems to be sufficient for some applications [39], in most situations it is not at all clear that structural communities actually map well to the organization of actors in social networks, functions in biological networks, etc. Hence, it is necessary to develop tools for the detection of functional communities that, whenever possible, incorporate node characteristics and other available information along with the network's structural information. The elephant in the literature is simply elucidated with just one question: *Now that we have all these ways of detecting communities, what do we do with them?*

Acknowledgements

Our views on network community structure have been shaped by numerous discussions with our colleagues and students over the last several years. We particularly acknowledge Aaron Clauset, Santo Fortunato, Nick Jones, Eric Kelsic, Mark Newman, Stephen Reid, and Chris Wiggins. We also thank Joe Blitzstein, Tim Elling, Santo Fortunato, James Fowler, A. J. Friend, Roger Guimerà, Nick Jones, David Kempe, Franziska Klingner, Renaud Lambiotte, David Lazer, Sune Lehmann, Jim Moody, and David Smith for useful comments on this manuscript, and Christina Frost and Amanda Traud for assistance in preparing some of the figures. We obtained data from Adam D'Angelo and Facebook, the House of Representatives Office of the Clerk (Congressional committee assignments), Mark Newman (network scientist coauthorship), James Fowler (Congressional legislation cosponsorship), and Keith Poole (Congressional roll call votes). PJM was funded by the NSF (DMS-0645369) and by start-up funds provided by the Institute for Advanced Materials, Nanoscience and Technology and the Department of Mathematics at the University of North Carolina at Chapel Hill. MAP acknowledges a research award (#220020177) from the James S. McDonnell Foundation. JPO is supported by the Fulbright Program.

References

[1] G. Agarwal and D. Kempe, *Modularity-maximizing network communities using mathematical programming*, The European Physical Journal B, 66 (2008), 409–418.

[2] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, *Communities and hierarchical organization of links in complex networks*. arXiv:0903.3178 (2009).

[3] R. Albert and A.-L. Barabási, *Statistical mechanics of complex networks*, Reviews of Modern Physics, 74 (2002), 47–97.

[4] J. M. Anthonisse, *The rush in a graph*. Technical Report BN 9/71, Stichting Mathematische Centrum, Amsterdam, 1971.

[5] A. Arenas, A. Fernández, S. Fortunato, and S. Gómez, *Motif-based communities in complex networks*, Journal of Physics A: Mathematical and Theoretical, 41, 224001 (2008).

[6] A. Arenas, A. Fernández, and S. Gómez, *Analysis of the structure of complex networks at different resolution levels*, New Journal of Physics, 10, 053039 (2008).

[7] J. P. Bagrow, *Evaluating local community methods in networks*, Journal of Statistical Mechanics: Theory and Experiment, P05001 (2008).

[8] J. P. Bagrow and E. M. Bollt, *A local method for detecting communities*, Physical Review E, 72, 046108 (2005).

[9] S. Bansal, S. Khandelwal, and L. A. Meyers, *Evolving clustered random networks*. arXiv: 0808.0509 (2008).

[10] M. J. Barber, *Modularity and community detection in bipartite networks*, Physical Review E, 76, 066102 (2007).

[11] M. Blatt, S. Wiseman, and E. Domany, *Superparamagnetic clustering of data*, Physical Review Letters, 76 (1996), 3251–3254.

[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large network*, Journal of Statistical Mechanics: Theory and Experiment, P10008 (2008).

[13] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, *Complex networks: Structure and dynamics*, Physics Reports, 424 (2006), 175–308.

References continued on page 1164.

References (Continued from page 1097)

[14] S. P. Borgatti, M. G. Everett, and P. R. Shirey, *LS Sets, Lambda Sets and other cohesive subsets*, Social Networks, 12 (1990), 337-357.

[15] J. M. Bower and H. Bolouri, eds., *Computational modeling of genetic and biochemical networks*, The MIT Press, Cambridge, Massachusetts, 2001.

[16] D. M. Boyd and N. B. Ellison, *Social network sites: Definition, history, and scholarship*, Journal of Computer-Mediated Communication, 13, 11 (2007).

[17] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, *On modularity clustering*, IEEE Transactions on Knowledge and Data Engineering, 20 (2008), 172–188.

[18] M. Brzozowski, T. Hogg, and G. Szabo, *Friends and foes: Ideological social networking*, in Proceedings of the SIGCHI Conference on Human Factors in Computing, ACM Press, New York, NY, 2008.

[19] C. T. Butts, *Network inference, error, and informant (in)accuracy: a Bayesian approach*, Social Networks, 25 (2003), 103–140.

[20] G. Caldarelli, *Scale-free networks: Complex webs in nature and technology*, Oxford University Press, Oxford, 2007.

[21] A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori, *Detecting communities in large networks*, Physica A, 352 (2004), 669–676.

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

[23] A. Clauset, C. Moore, and M. E. J. Newman, *Hierarchical structure and the prediction of missing links in networks*, Nature, 453 (2008), 98–101.

[24] A. Clauset, M. E. J. Newman, and C. Moore, *Finding community structure in very large networks*, Physical Review E, 70, 066111 (2004).

[25] J. S. Coleman, *An introduction to mathematical sociology*, Collier-Macmillan, London, 1964.

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

[27] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, *Stability of graph communities across time scales*. arXiv:0812.1811 (2008).

[28] I. Derényi, G. Palla, and T. Vicsek, *Clique percolation in random networks*, Physical Review Letters, 94, 160202 (2005).

[29] L. Donetti and M. A. Muñoz, *Detecting network communities: A new systematic and efficient algorithm*, Journal of Statistical Mechanics: Theory and Experiment, P10012 (2004).

[30] J. Duch and A. Arenas, *Community detection in complex networks using extremal optimization*, Physical Review E, 72, 027104 (2005).

[31] J. Eisner, *State-of-the-art algorithms for minimum spanning trees: A tutorial discussion*. Manuscript available at `http://cs.jhu.edu/~jason/papers/#ms97` (1997).

[32] P. Elias, A. Feinstein, and C. E. Shannon, *A note on the maximum flow through a network*, IRE Transactions on Information Theory, 2, 4 (1956), 117–119.

[33] T. S. Evans and R. Lambiotte, *Line graphs, link partitions and overlapping communities*, Physical Review E, 80, 016105 (2009).

[34] D. J. Fenn, M. A. Porter, M. McDonald, S. Williams, N. F. Johnson, and N. S. Jones, *Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007–2008 credit crisis*, Chaos, 19:3, 033119 (2009).

[35] M. Fiedler, *Algebraic connectivity of graphs*, Czechoslovak Mathematical Journal, 23 (1973), 289–305.

[36] K. H. Fischer and J. A. Hertz, *Spin glasses*, Cambridge University Press, Cambridge, 1993.

[37] L. R. Ford and D. R. Fulkerson, *Maximal flow through a network*, Canadian Journal of Mathematics, 8 (1956), 399–404.

[38] S. Fortunato and M. Barthélemy, *Resolution limit in community detection*, Proceedings of the National Academy of Sciences, 104 (2007), 36–41.

[39] S. Fortunato, *Community detection in graphs*. arXiv:0906.0612 (2009).

[40] J. H. Fowler, *Legislative cosponsorship networks in the U.S. House and Senate*, Social Networks, 28 (2006), 456–487.

[41] J. H. Fowler and S. Jeon, *The authority of Supreme Court precedent*, Social Networks, 30 (2008), 16–30.

[42] L. C. Freeman, *A set of measures of centrality based upon betweenness*, Sociometry, 40 (1977), 35–41.

[43] ———, *Segregation in social networks*, Sociological Methods Research, 6 (1978), 411–429.

[44] ———, *The development of social network analysis: A study in the sociology of science*, Empirical Press, Vancouver, Canada, 2004.

[45] T. M. J. Fruchterman and E. M. Reingold, *Graph drawing by force-directed placement*, Software–Practice and Experience, 21 (1991), 1129–1164.

[46] G. Gan, C. Ma, and J. Wu, *Data clustering: Theory, algorithms, and applications*, Society for Industrial and Applied Mathematics, Philadelphia, 2007.

[47] J. W. Gibbs, *Elementary principles in statistical mechanics*, Charles Scribner's Sons, New York, NY, 1902.

[48] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences, 99 (2002), 7821–7826.

[49] M. C. González, C. A. Hidalgo, and A.-L. Barabási, *Understanding individual human mobility patterns*, Nature, 453 (2008), 779–782.

[50] J. C. Gower and G. J. S. Ross, *Minimum spanning trees and single linkage cluster analysis*, Applied Statistics, 18 (1969), 54–64.

[51] M. Granovetter, *The strength of weak ties*, The American Journal of Sociology, 78 (1973), 1360–1380.

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

[53] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, *Module identification in bipartite and directed networks*, Physical Review E, 76, 036102 (2007).

[54] N. Gulbahce and S. Lehmann, *The art of community detection*, Bioessays, 30 (2008), 934–938.

[55] M. S. Handcock, A. E. Rafter, and J. M. Tantrum, *Model-based clustering for social networks*, Journal of the Royal Statistical Society A, 170 (2007), 301–354.

[56] M. B. Hastings, *Community detection as an inference problem*, Physical Review E, 74, 035102 (2006).

[57] J. M. Hofman and C. H. Wiggins, *Bayesian approach to network modularity*, Physical Review Letters, 100, 258701 (2008).
[58] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, *Tracking evolving communities in large linked networks*, Proceedings of the National Academy of Sciences, 101 (2004), 5249–5253.
[59] T. Hogg, D. Wilkinson, G. Szabo, and M. Brzozowski, *Multiple relationship types in online communities and social networks*, in Proceedings of the AAAI Spring Symposium on Social Information Processing, AAAI Press (2008).
[60] G. C. Homans, *The human group*, Harcourt, Brace, New York, 1950.
[61] S. C. Johnson, *Hierarchical clustering schemes*, Psychometrica, 32 (1967), 241–254.
[62] T. Kamada and S. Kawai, *An algorithm for drawing general undirected graphs*, Information Processing Letters, 31 (1989), 7–15.
[63] F. Képès, ed., *Biological networks*, Complex Systems and Interdisciplinary Science, Vol. 3, World Scientific, Hackensack, NJ, 2007.
[64] B. W. Kernighan and S. Lin, *An efficient heuristic procedure for partitioning graphs*, The Bell System Technical Journal, 49 (1970), 291–307.
[65] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Review, 51:3 (2009), in press.
[66] C. P. Kottak, *Cultural anthropology*, McGraw-Hill, New York, 5th ed., 1991.
[67] R. Kumar, J. Novak, and A. Tomkins, *Structure and evolution of online social networks*. 12th International Conference on Knowledge Discovery and Data Mining (2006).
[68] R. Lambiotte, J.-C. Delvenne, and M. Barahona, *Dynamics and modular structure in networks*. arXiv:0812.1770 (2008).
[69] A. Lancichinetti, S. Fortunato, and J. Kertesz, *Detecting the overlapping and hierarchical community structure of complex networks*, New Journal of Physics, 11, 033015 (2009).
[70] A. Lancichinetti, S. Fortunato, and F. Radicchi, *Benchmark graphs for testing community detection algorithms*, Physical Review E, 93, 046110 (2008).
[71] S. Lehmann, M. Schwartz, and L. K. Hansen, *Biclique communities*, Physical Review E, 78, 016108 (2008).
[72] E. A. Leicht and M. E. J. Newman, *Community structure in directed networks*, Physical Review Letters, 100, 118703 (2008).
[73] K. Lewis, J. Kaufman, M. Gonzalez, M. Wimmer, and N. A. Christakis, *Tastes, ties, and time: A new (cultural, multiplex, and longitudinal) social network dataset using Facebook.com*, Social Networks, 30 (2008), 330–342.
[74] R. Luce and A. Perry, *A method of matrix analysis of group structure*, Psychometrika, 14 (1949), 95–116.
[75] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network motifs: Simple building blocks of complex networks*, Science, 298 (2002), 824–827.
[76] M. Molloy and B. Reed, *A critical point for random graphs with a given degree sequence*, Random Structures and Algorithms, 6 (1995), 161-180.
[77] J. Moody, *Peer influence groups: Identifying dense clusters in large networks*, Social Networks, 23 (2001), 261–283.
[78] J. Moody and D. R. White, *The cohesiveness of blocks in social networks*, Sociological Methodology, 31 (2001), 305–359.
[79] ———, *Structural cohesion and embeddedness: A hierarchical concept of social groups*, American Sociological Review, 68 (2003), 103–127.
[80] M. E. J. Newman, *Scientific collaboration networks: I. Network construction and fundamental results*, Physical Review E, 64, 016131 (2001).
[81] ———, *Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality*, Physical Review E, 64, 016132 (2001).
[82] ———, *The structure of scientific collaborations*, Proceedings of the National Academy of Sciences, 98 (2001), 404–409.
[83] ———, *The structure and function of complex networks*, SIAM Review, 45 (2003), 167–256.
[84] ———, *Fast algorithm for detecting community structure in networks*, Physical Review E, 69, 066133 (2004).
[85] ———, *A measure of betweeenness centrality based on random walks*, Social Networks, 27 (2005), 39–54.
[86] ———, *Finding community structure in networks using the eigenvectors of matrices*, Physical Review E, 74, 036104 (2006).
[87] ———, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences, 103 (2006), 8577–8582.
[88] ———, *The physics of networks*, Physics Today, 61:11 (2008), pp. 33–38.
[89] M. E. J. Newman and M. Girvan, *Mixing patterns and community structure in networks*, in Statistical Mechanics of Complex Networks (R. Pastor-Satorras, J. Rubi, and A. Diaz-Guilera, eds.), Springer-Verlag, Berlin, 2003.
[90] ———, *Finding and evaluating community structure in networks*, Physical Review E, 69, 026113 (2004).
[91] M. E. J. Newman and E. A. Leicht, *Mixture models and exploratory analysis in networks*, Proceedings of the National Academy of Sciences, 104 (2007), 9564–9569.
[92] A. Noack, *Modularity clustering is force-directed layout*, Physical Review E, 79, 026102 (2009).
[93] A. Noack and R. Rotta, *Multi-level algorithms for modularity clustering*, in Experimental Algorithms: 8th International Symposium, SEA 2009, Dortmund, Germany, June 4–6, 2009, Proceedings, Springer-Verlag, New York, 2009, pp. 257–268.
[94] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész, *Analysis of a large-scale weighted network of one-to-one human communication*, New Journal of Physics, 9 (2007), 179.
[95] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, *Structure and tie strengths in mobile communication networks*, Proceedings of the National Academy of Sciences, 104 (2007), 7332–7336.
[96] G. Palla, A.-L. Barabási, and T. Vicsek, *Quantifying social group evolution*, Nature, 446 (2007), 664–667.
[97] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Uncovering the overlapping community structure of*

complex networks in nature and society, Nature, 435 (2005), 814–818.

[98] P. Pons and M. Latapy, *Computing communities in large networks using random walks*, Journal of Graph Algorithms and Applications, 10 (2006), 191–218.

[99] K. T. Poole and H. Rosenthal, *Congress: A political-economic history of roll call voting*, Oxford University Press, Oxford, 1997.

[100] M. A. Porter, P. J. Mucha, M. E. J. Newman, and A. J. Friend, *Community structure in the United States House of Representatives*, Physica A, 386 (2007), 414–438.

[101] M. A. Porter, P. J. Mucha, M. E. J. Newman, and C. M. Warmbrand, *A network analysis of committees in the United States House of Representatives*, Proceedings of the National Academy of Sciences, 102 (2005), 7057–7062.

[102] A. Pothen, H. D. Simon, and K.-P. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM Journal on Matrix Analysis and Applications, 11 (1990), 430–452.

[103] J. Reichardt and S. Bornholdt, *Detecting fuzzy community structures in complex networks with a Potts model*, Physical Review Letters, 93, 218701 (2004).

[104] ———, *Statistical mechanics of community detection*, Physical Review E, 74, 016110 (2006).

[105] J. Reichardt and D. R. White, *Role models for complex networks*, European Physical Journal B, 60 (2007), 217–224.

[106] S. A. Rice, *The identification of blocs in small political bodies*, American Political Science Review, 21 (1927), 619–627.

[107] T. Richardson, P. J. Mucha, and M. A. Porter, *Spectral tripartitioning of networks*, Physical Review Letters E, in press (2009), arXiv:0812.2852.

[108] M. Rosvall and C. T. Bergstrom, *Maps of random walks on complex networks reveal community structure*, Proceedings of the National Academy of Sciences, 105 (2008), 1118–1123.

[109] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral, *Extracting the hierarchical organization of complex systems*, Proceedings of the National Academy of Sciences, 104, 39 (2007), 15224–15229.

[110] S. E. Schaeffer, *Graph clustering*, Computer Science Review, 1 (2007), 27–64.

[111] P. Schuetz and A. Caflisch, *Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement*, Physical Review E, 77, 046112 (2008).

[112] F. Schwabl, *Statistical mechanics*, Springer-Verlag, New York, NY, 2002.

[113] R. Scupin, *Cultural anthropology—a global perspective*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[114] S. Seidman and B. Foster, *A graph theoretic generalization of the clique concept*, Journal of Mathematical Sociology, 6 (1978), 139–154.

[115] T. M. Selee, T. G. Kolda, W. P. Kegelmeyer, and J. D. Griffin, *Extracting clusters from large datasets with multiple similarity measures using IMSCAND*, in CSRI Summer Proceedings 2007, Technical Report SAND2007-7977, Sandia National Laboratories, Albuquerque, NM and Livermore, CA (M. L. Parks and S. S. Collis, eds.), 2007, pp. 87–103.

[116] C. R. Shalizi, M. F. Camperi, and K. L. Klinkner, *Discovering functional communities in dynamical networks*, in Statistical Network Analysis: Models, Issues, and New Directions, Springer-Verlag, New York, NY (E. M. Airoldi, D. M. Blei, S. E. Feinberg, A. Goldenberg, E. P. Xing, and A. X. Zheng, eds.), 2007, pp. 140–157.

[117] H. Simon, *The architecture of complexity*, Proceedings of the American Philosophical Society, 106 (1962), 467–482.

[118] L. Sirovich, *A pattern analysis of the second Renquist U.S. Supreme Court*, Proceedings of the National Academy of Sciences, 100 (2003), 7432–7437.

[119] P. B. Slater, *Established clustering procedures for network analysis*. arXiv:0806.4168 (2008).

[120] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), 268–276.

[121] V. A. Traag and J. Bruggeman, *Community detection in networks with positive and negative links*. arXiv:0811.2329 (2008).

[122] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, *Community structure in online collegiate social networks*. arXiv:0809.0960 (2008).

[123] G. Wang, Y. Shen, and M. Ouyang, *A vector partitioning approach to detecting community structure in complex networks*, Computers & Mathematics with Applications, 55 (2007), 2746–2752.

[124] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, Structural Analysis in the Social Sciences, Cambridge University Press, Cambridge, 1994.

[125] A. S. Waugh, L. Pei, J. H. Fowler, P. J. Mucha, and M. A. Porter, *Party polarization in Congress: A social networks approach*. arXiv:0907.3509 (2009).

[126] R. S. Weiss and E. Jacobson, *A method for the analysis of the structure of complex organizations*, American Sociological Review, 20 (1955), 661–668.

[127] H. C. White, S. A. Boorman, and R. L. Breiger, *Social structure from multiple networks. I. Blockmodels of roles and positions*, The American Journal of Sociology, 81, 4 (1976), 730–780.

[128] Wikipedia, *List of social networking websites*. `http://en.wikipedia.org/wiki/List_of_social_networking_websites`, accessed 13 February 2009.

[129] F. Wu and B. A. Huberman, *Finding communities in linear time: A physics approach*, The European Physical Journal B, 38 (2004), 331–338.

[130] F. Y. Wu, *The Potts model*, Reviews of Modern Physics, 54 (1982), 235–268.

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

[132] L. V. Zhang, O. D. King, S. L. Wong, D. S. Goldberg, A. H. Y. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F. P. Roth, *Motifs, themes and thematic maps of an integrated* Saccharomyces cerevisiae *interaction network*, Journal of Biology, 4, 6 (2005).

[133] Y. Zhang, A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha, *Community structure in Congressional cosponsorship networks*, Physica A, 387 (2008), 1705–1712.