nature physics

INSIGHT | REVIEW ARTICLES

PUBLISHED ONLINE: 22 DECEMBER 2011 | DOI: 10.1038/NPHYS2162

Communities, modules and large-scale structure in networks

M. E. J. Newman

Networks, also called graphs by mathematicians, provide a useful abstraction of the structure of many complex systems, ranging from social systems and computer networks to biological networks and the state spaces of physical systems. In the past decade there have been significant advances in experiments to determine the topological structure of networked systems, but there remain substantial challenges in extracting scientific understanding from the large quantities of data produced by the experiments. A variety of basic measures and metrics are available that can tell us about small-scale structure in networks, such as correlations, connections and recurrent patterns, but it is considerably more difficult to quantify structure on medium and large scales, to understand the 'big picture'. Important progress has been made, however, within the past few years, a selection of which is reviewed here.

A network is, in its simplest form, a collection of dots joined together in pairs by lines (Fig. 1). In the jargon of the field, a dot is called a 'node' or 'vertex' (plural 'vertices') and a line is called an 'edge'. Networks are used in many branches of science as a way to represent the patterns of connections between the components of complex systems[1–6]. Examples include the Internet[7,8], in which the nodes are computers and the edges are data connections such as optical-fibre cables, food webs in biology[9,10], in which the nodes are species in an ecosystem and the edges represent predator–prey interactions, and social networks[11,12], in which the nodes are people and the edges represent any of a variety of different types of social interaction including friendship, collaboration, business relationships or others.

In the past decade there has been a surge of interest in both empirical studies of networks[13] and development of mathematical and computational tools for extracting insight from network data[1–6]. One common approach to the study of networks is to focus on the properties of individual nodes or small groups of nodes, asking questions such as, 'Which is the most important node in this network?' or 'Which are the strongest connections?' Such approaches, however, tell us little about large-scale network structure. It is this large-scale structure that is the topic of this paper.

The best-studied form of large-scale structure in networks is modular or community structure[14,15]. A community, in this context, is a dense subnetwork within a larger network, such as a close-knit group of friends in a social network or a group of interlinked web pages on the World Wide Web (Fig. 1). Although communities are not the only interesting form of large-scale structure—there are others that we will come to—they serve as a good illustration of the nature and scope of present research in this area and will be our primary focus.

Communities are of interest for a number of reasons. They have intrinsic interest because they may correspond to functional units within a networked system, an example of the kind of link between structure and function that drives much of the present excitement about networks. In a metabolic network[16], for instance—the network of chemical reactions within a cell—a community might correspond to a circuit, pathway or motif that carries out a certain function, such as synthesizing or regulating a vital chemical product[17]. In a social network, a community might correspond to an actual community in the conventional sense of the word, a group of people brought together by a common interest, a common location or workplace or family ties[18].

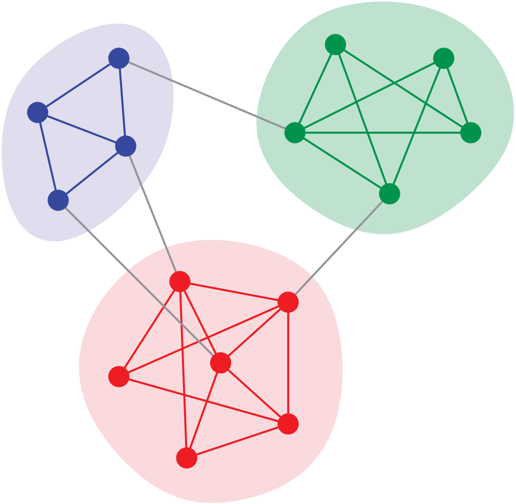

Figure 1 | Example network showing community structure. The nodes of this network are divided into three groups, with most connections falling within groups and only a few between groups.

However, there is another reason, less often emphasized, why a knowledge of community structure can be useful. In many networks it is found that the properties of individual communities can be quite different. Consider, for example, Fig. 2, which shows a network of collaborations among a group of scientists at a research institute. The network divides into distinct communities as indicated by the colours of the nodes. (We will see shortly how this division is accomplished.) In this case, the communities correspond closely to the acknowledged research groups within the institute, a demonstration that indeed the discovery of communities can point to functional divisions in a system. However, notice also that the structural features of the different communities are widely varying. The communities highlighted in red and light blue, for instance, appear to be loose-knit groups of collaborators working together in various combinations, whereas the groups in yellow and dark blue are both organized around a central hub, perhaps a group

Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109, USA. e-mail: mejn@umich.edu.

© 2012 Macmillan Publishers Limited. All rights reserved.

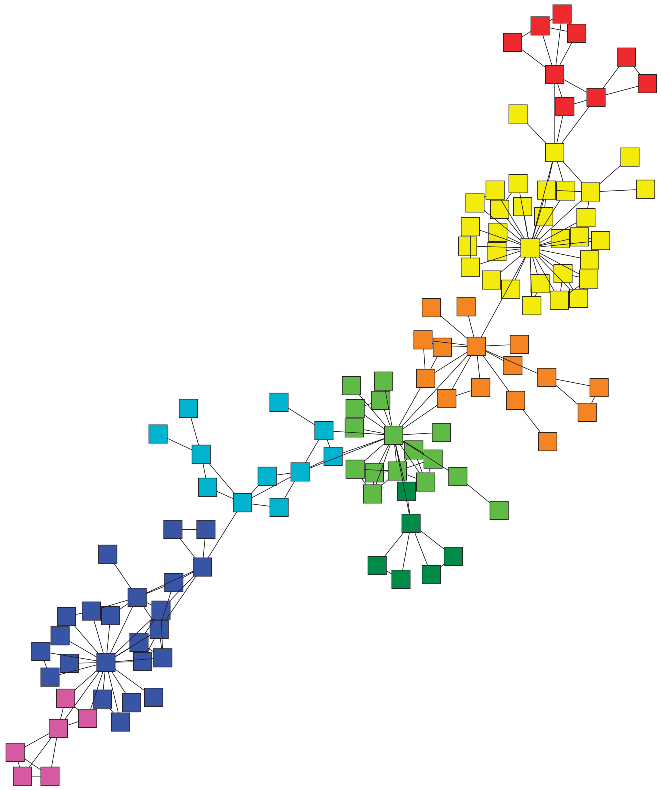

Figure 2 | A network of collaborations among scientists at a research institute. Nodes in this network represent the scientists and there is an edge between any pair of scientists who co-authored a published paper during the years of the study. Colours represent communities, as discovered using a modularity-maximization technique.

leader or principal investigator of some kind. Distinctions such as these, which may be crucial for understanding the behaviour of the system, become apparent only when one looks at structure on the community level.

The network in this particular example has the nice property that it is small enough and sparse enough to be drawn clearly on the page. One does not need any calculations to pick out the communities in this case: a good eye will do the job. However, when we are working with larger or denser networks, networks that can have thousands or even millions of nodes (or a smaller number of nodes but very many edges), clear visualization becomes impossible and we must turn instead to algorithmic methods for community detection and the development of such methods has been a highly active area of research in the past few years[15].

The community-detection problem is challenging in part because it is not very well posed. It is agreed that the basic problem is to find locally dense regions in a network, but this is not a precise formulation. If one is to create a method for detecting communities in a mechanical way, one must first define exactly what one means by a community. Researchers have been aware of this issue from the outset and have proposed a wide variety of definitions, based on counts of edges within and between communities, counts of paths across networks, spectral properties of network matrices, information-theoretic measures, random walks and many other quantities. With this array of definitions comes a corresponding array of algorithms that seek to find the communities so defined[14,15,19–31]. Unfortunately, it is no easy matter to determine which of these algorithms are the best, because the perception of good performance itself depends on how one defines a community and each algorithm is necessarily good at finding communities according to its own definition. To get around this circularity, we typically take one of two approaches. In the first, algorithms are tested against real-world networks for which there is an accepted division into communities, often based on additional measurements that are independent of the network itself, such as interviews with participants in a social network or analysis of the text of web pages. If an algorithm can reliably find the accepted structure then it is considered successful. In the second approach, algorithms are tested against computer-generated networks that have some form of community structure artificially embedded within them. A number of standard benchmark networks have been proposed for this purpose, such as the 'four groups' networks[14] or so-called the LFR benchmark networks[32]. A number of studies have been published that compare the performance of proposed algorithms in these benchmark tests[33,34]. Although these approaches do set concrete targets for performance of community-detection methods, there is room for debate over whether those targets necessarily align with good performance in broader real-world situations. If we tune our algorithms to solve specific benchmark problems we run the risk of creating algorithms that solve those problems well but other (perhaps more realistic) problems poorly.

This is a crucial issue and one that is worth bearing in mind as we take a look in the following sections at the present state of research on community detection. As we will see, however, researchers have, in spite of the difficulties, come up with a range of approaches that return real, useful information about the large-scale structure of networks, and in the process have learned much, both about individual networks that have been analysed and about mathematical methods for representing and understanding network structure.

Hierarchical clustering

Studies of communities in networks go back at least to the 1970s, when a number of techniques were developed for their detection, particularly in computer science and sociology. In computer science the problem of graph partitioning[35], which is similar but not identical to the problem of community detection, has received attention for its engineering applications, but the methods developed, such as spectral partitioning[36] and the Kernighan–Lin algorithm[37], have also been fruitfully applied in other areas. However, it is the work of sociologists that is perhaps the most direct ancestor of modern techniques of community detection.

An early, and still widely used, technique for detecting communities in social networks is hierarchical clustering[5,11]. Hierarchical clustering is in fact not a single technique but an entire family of techniques, with a single central principle: if we can derive a measure of how strongly nodes in a network are connected together, then by grouping the most strongly connected we can divide the network into communities. Specific hierarchical clustering methods differ on the particular measure of strength used and on the rules by which we group strongly connected nodes. Most common among the measures used are the so-called structural equivalence measures, which focus on the number n_{ij} of common network neighbours that two nodes i,j have. In a social network of friendships, for example, two people with many mutual friends are more likely to be close than two people with few and thus a count of mutual friends can be used as a measure of connection strength. Rather than using the raw count n_{ij}, however, one typically normalizes it in some way, leading to measures such as the Jaccard coefficient and cosine similarity. For example, the cosine similarity σ_{ij} between nodes i and j is defined by

$$\sigma_{ij} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

where k_i is the degree of node i (that is, the number of connections it has). This measure has the nice property that its

© 2012 Macmillan Publishers Limited. All rights reserved.

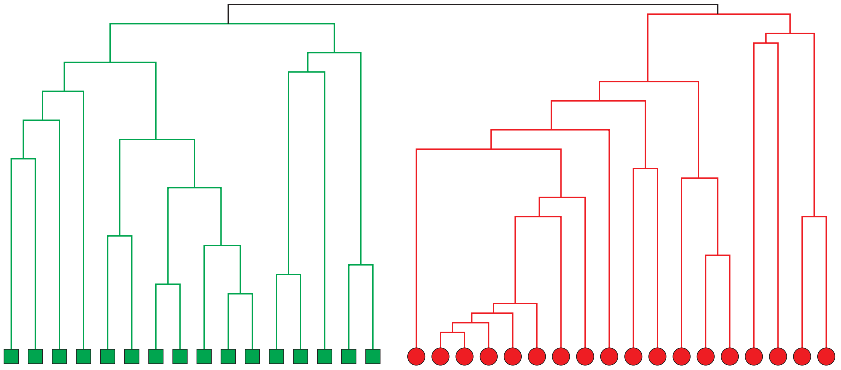

Figure 3 | Average-linkage clustering of a small social network. This tree or 'dendrogram' shows the results of the application of average-linkage hierarchical clustering using cosine similarity to the well-known karate-club network of Zachary[38], which represents friendship between members of a university sports club. The calculation finds two principal communities in this case (the left and right subtrees of the dendrogram), which correspond exactly to known factions within the club (represented by the colours).

value falls always between zero and one—zero if the nodes have no common neighbours and one if they have all their neighbours in common.

Once one has defined a measure of connection strength, one can begin to group nodes together, which is done in hierarchical fashion, first grouping single nodes into small groups, then grouping those groups into larger groups and so forth. There are a number of methods by which this grouping can be carried out, the three common ones being the methods known as single-linkage, complete-linkage and average-linkage clustering. Single-linkage clustering is the most widely used by far, primarily because it is simple to implement, but in fact average-linkage clustering generally gives superior results and is not much harder to implement.

Figure 3 shows the result of applying average-linkage hierarchical clustering based on cosine similarity to a famous network from the social networks literature, Zachary's karate-club network[38]. This network represents patterns of friendship between members of a karate club at a US university, compiled from observations and interviews of the club's 34 members. The network is of particular interest because during the study a dispute arose among the club's members over whether to raise club fees. Unable to reconcile their differences, the members of the club split into two factions, with one faction departing to start a separate club. It has been claimed repeatedly that by examining the pattern of friendships depicted in the network (which was compiled before the split happened) one can predict the membership of the two factions[14,20,26,27,38–40].

Figure 3 shows the output of the hierarchical clustering procedure in the form of a tree or 'dendrogram' representing the order in which nodes are grouped together into communities. It should be read from the bottom up: at the bottom we have individual nodes that are grouped first into pairs, and then into larger groups as we move up the tree, until we reach the top, where all nodes have been gathered into one group. In a single image, this dendrogram captures the entire hierarchical clustering process. Horizontal cuts through the figure represent the groups at intermediate stages.

As we can see, the method in this case joins the nodes together into two large groups, consisting of roughly half the network each, before finally joining those two into one group at the top of the dendrogram. It turns out that these two groups correspond precisely to the groups into which the club split in real life, which are indicated by the colours in the figure. Thus, in this case the method works well. It has effectively predicted a future social phenomenon, the split of the club, from quantitative data measured before the split occurred. It is the promise of outcomes such as this that drives much of the present interest in networks.

Hierarchical clustering is straightforward to understand and to implement, but it does not always give satisfactory results. As it exists in many variants (different strength measures and different linkage rules) and different variants give different results, it is not clear which results are the 'correct' ones. Moreover, the method has a tendency to group together those nodes with the strongest connections but leave out those with weaker connections, so that the divisions it generates may not be clean divisions into groups, but rather consist of a few dense cores surrounded by a periphery of unattached nodes. Ideally, we would like a more reliable method.

Optimization methods

Over the past decade or so, researchers in physics and applied mathematics have taken an active interest in the community-detection problem and introduced a number of fruitful approaches. Among the first proposals were approaches based on a measure known as betweenness[14,21,41], in which one calculates one of several measures of the flow of (imaginary) traffic across the edges of a network and then removes from the network those edges with the most traffic. Two other related approaches are the use of fluid-flow[19] and current-flow analogies[42] to identify edges for removal; the latter idea has been revived recently to study structure in the very largest networks[30]. A different class of methods are those based on information-theoretic ideas, such as the minimum-description-length methods of Rosvall and Bergstrom[26,43] and related methods based on statistical inference, such as the message-passing method of Hastings[25]. Another large class exploits links between community structure and processes taking place on networks, such as random walks[44,45], Potts models[46] or oscillator synchronization[47]. A contrasting set of approaches focuses on the detection of 'local communities'[23,24] and seeks to answer the question of whether we can, given a single node, identify the community to which it belongs, without first finding all communities in the network. In addition to being useful for studying limited portions of larger networks, this approach can give rise to overlapping communities, in which a node can belong to more than one community. (The generalized community-detection problem in which overlaps are allowed in this way has been an area of increasing interest within the field in recent years[22,31].)

However, the methods most heavily studied by physicists, perhaps unsurprisingly, are those that view the community-detection problem by analogy with equilibrium physical processes and treat it as an optimization task. The basic idea is to define a quantity that is high for 'good' divisions of a network and low for 'bad' ones, and then to search through possible divisions for the one with the highest score. This approach is similar to the minimization

© 2012 Macmillan Publishers Limited. All rights reserved.

of energy when finding the ground state or stable state of a physical system, and the connection has been widely exploited. A variety of different measures for assigning scores have been proposed, such as the so-called E/I ratio[48], likelihood-based measures[49] and others[50], but the most widely used is the measure known as the modularity[18,51].

Suppose you are given a network and a candidate division into communities. A simple measure of the quality of that division is the fraction of edges that fall within (rather than between) communities. If this fraction is high then you have a good division (Fig. 1). However, this measure is not ideal. It is maximized by putting all nodes in a single group together, which is a correct but trivial form of community structure and not of particular interest. A better measure is the so-called modularity, which is defined to be the fraction of edges within communities minus the expected value of that fraction if the positions of the edges are randomized[51]. If there are more edges within communities than one would find in a randomized network then the modularity will be positive and large positive values indicate good community divisions.

Let A_{ij} be equal to the number of edges between nodes i and j (normally zero or one); A_{ij} is an element of the 'adjacency matrix' of the network. It can be shown that for a network with m edges in total, the expected number that fall between nodes i and j if the positions of the edges are randomized is given by $k_i k_j/2m$, where k_i is again the degree of node i. Thus, the actual number of edges between i and j minus the expected number is $A_{ij} - k_i k_j/2m$ and the modularity Q is the sum of this quantity over all pairs of nodes that fall in the same community. If we label the communities and define s_i to be the label of the community to which node i belongs, then we can write

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta_{s_i,s_j}$$

where δ_{ij} is the Kronecker delta and the leading constant $1/2m$ is included only by convention—it normalizes Q to measure fractions of edges rather than total numbers but its presence has no effect on the position of the modularity maximum.

The modularity takes precisely the form $H = -\sum_{ij} J_{ij}\delta_{s_i,s_j}$ of the Hamiltonian of a (disordered) Potts model, apart from a minus sign, and hence its maximization is equivalent to finding the ground state of the Potts model—the community assignments s_i act similarly to spins on the nodes of the network. Unfortunately, direct optimization of the modularity by an exhaustive search through the possible spin states is intractable for any but the smallest of networks, and faster indirect (but exact) algorithms have been proved rigorously not to exist[52]. A variety of approximate techniques from physics and elsewhere, however, are applicable to the problem and seem to give good, but not perfect, solutions with relatively modest computational effort. These include simulated annealing[17,53], greedy algorithms[54,55], semidefinite programming[28], spectral methods[56] and several others[40,57]. Modularity maximization forms the basis for other more complex approaches as well, such as the method of Blondel *et al.*[27], a multiscale method in which modularity is first optimized using a greedy local algorithm, then a 'supernetwork' is formed whose nodes represent the communities so discovered and the greedy algorithm is repeated on this supernetwork. The process iterates until no further improvements in modularity are possible. This method has become widely used by virtue of its relative computational efficiency and the high quality of the results it returns. In a recent comparative study it was found to be one of the best available algorithms when tested against computer-generated benchmark problems of the type described in the introduction[34].

Figure 2, showing collaboration patterns among scientists, is an example of community detection using modularity maximization. One of the nice features of the modularity method is that one does not need to know in advance the number of communities contained in the network: a free maximization of the modularity, in which the number of communities is allowed to vary, will tell us the most advantageous number, as well as finding the exact division of the nodes among communities.

Although modularity maximization is efficient, widely used and gives informative results, it—like hierarchical clustering—has deficiencies. In particular, it has a known bias in the size of the communities it finds—it has a preference for communities of size roughly equal to the square root of the size of the network[58]. Modifications of the method have been proposed that allow one to vary this preferred size[59,60], but not to eliminate the preference altogether. The modularity method also ignores any information stored in the positions of edges that run between communities: as modularity is calculated by counting only within-group edges, one could move the between-group edges around in any way one pleased and the value of the modularity would not change at all. One might imagine that one could do a better job of detecting communities if one were to make use of the information represented by these edges.

In the past few years, therefore, researchers have started to look for a more principled approach to community detection, and have gravitated towards the method of block modelling, a method that traces its roots back to the 1970s (refs 61,62), but which has recently enjoyed renewed popularity, with some powerful new methods and results emerging.

Block models

Block modelling[63–67] is in effect a form of statistical inference for networks. In the same way that we can gain some understanding from conventional numerical data by fitting, say, a straight line through data points, so we can gain understanding of the structure of networks by fitting them to a statistical network model. In particular, if we are interested in community structure then we can create a model of networks that contain such structure, then fit it to an observed network and in the process learn about community structure in that observed network, if it exists.

A simple example of a block model is a model network in which one has a certain number n of nodes and each node is assigned to one of several labelled groups or communities. In addition, one specifies a set of probabilities p_{rs}, which represent the probability that there will be an edge between a node in group r and a node in group s. This model can be used, for instance, in a generative process to create a random network with community structure. By making the edge probabilities higher for pairs of nodes in the same group and lower for pairs in different groups, then generating a set of edges independently with exactly those probabilities, one can produce an artificial network that has many edges within groups and few between them—the classic community structure.

However, we can also turn the experiment around and ask, 'If we observe a real network and we suppose that it was generated by this model, what would the values of the model's parameters have to be?' More precisely, what values of the parameters are most likely to have generated the network we see in real life? This leads us to a 'maximum likelihood' formulation of the community-detection problem. The probability, or likelihood, that an observed network was generated by this block model is given by

$$L = \prod_{i<j} p_{s_i s_j}^{A_{ij}} (1 - p_{s_i s_j})^{1-A_{ij}}$$

where A_{ij} is an element of the adjacency matrix, as before, and s_i is again the community to which node i belongs. Now

© 2012 Macmillan Publishers Limited. All rights reserved.

we simply maximize this quantity over the probabilities p_{rs} and the communities s_i. Again we have turned the detection of communities into an optimization problem, albeit a harder one than the modularity-maximization problem. The values of the probabilities p_{rs} are usually of lesser interest to us, but if we can find the community parameters s_i that maximize the likelihood then we have solved our community-detection problem.

Although it seems elegant and well-founded in principle, the surprising thing about this approach, at least as we have described it here, is that it does not work well. Figure 4a shows an example application of (a slight variant of) the method to a network of weblogs, or 'blogs'—personal web pages maintained by individuals or groups, on which they publish their thoughts on topics of their choosing. This particular network, which was assembled by Adamic and Glance[68], is composed of blogs about US politics that were active around the time of the US presidential election in 2004, and the edges in the network represent web hyperlinks between blogs. Adamic and Glance showed that this network was strongly divided into two communities, one of left-leaning (that is, liberal) blogs, which commonly link to one another, and the other of right-leaning (conservative) ones, which also link to one another, but that there were few links between left and right. The communities appear as roughly the left and right halves of the network as it is drawn in Fig. 4a. The colours in the figure show the division of the network into two communities found with the maximum likelihood method above, and it is clear that the method has failed to find the known division in this case. What has gone wrong?

On closer inspection, we find that the method fails in this case because it does not take into account the wide variation among the degrees of nodes in the network. In this network (and many others) degrees vary over a great range, whereas degrees in the block model are Poisson distributed and narrowly peaked about their mean. This means, in effect, that there is no choice of parameters for the model that gives a good fit to the data. Fitting this block model is similar to fitting a straight line through an inherently curved set of data points—you can do it, but it is unlikely to give you a meaningful answer.

It turns out, however, that one can fix such problems by suitably modifying the model. Figure 4b shows a different fit to the same network using now a 'degree-corrected' block model that allows for widely varying degrees[49]. As the figure shows, the model now finds a division that corresponds closely to the known division between left- and right-leaning blogs. The moral of the story is that it is not hard to come up with models so unrealistic that they will not fit the observed network for any parameter values and one must guard against this possibility if the method is to work.

Once we deal with this issue, however, the block-model method has some promising features. If we have found the parameter values for the best fit of the model to an observed network, we can then plug those values back into the model and use the model to generate further networks that are similar to the original network, but not identical. This ability to generate similar networks can be used, for instance, to guess at the locations of possible missing edges in a network. For many networks our data are incomplete or unreliable, and there may be edges missing from the recorded structure. Looking at a large selection of generated networks that are similar to the original, one can find edges that appear often in the generated networks but not in the original; such edges turn out to be reliable candidates for missing data. Guimerá and Sales-Pardo[69] have shown that this approach is at least as accurate as, and often better than, previous methods for predicting missing edges.

Another nice feature of the block-model method is that it lends itself to many variants that are suitable for particular types of problem. For instance, in some problems we can, with some effort, carry out experiments to determine the community membership of a few nodes, and the goal is to determine the rest. In recent work, Yan *et al.*[70] have devised a variant of the block-model method in which one can use the model to determine on which nodes these experiments should be done, by looking for the nodes whose membership information will be most useful, in the sense that it will tell us as much as possible not only about the measured nodes but also about the membership of other nodes in the network. They show that the accuracy of community detection can be enormously improved by carrying out just a few experiments on nodes carefully chosen using this technique.

However, perhaps the most promising feature of the block-model method is that it is not limited to detecting traditional community structure in networks. In principle, any type of structure that can be formulated as a probabilistic model can be detected, including overlapping communities, bipartite or k-partite

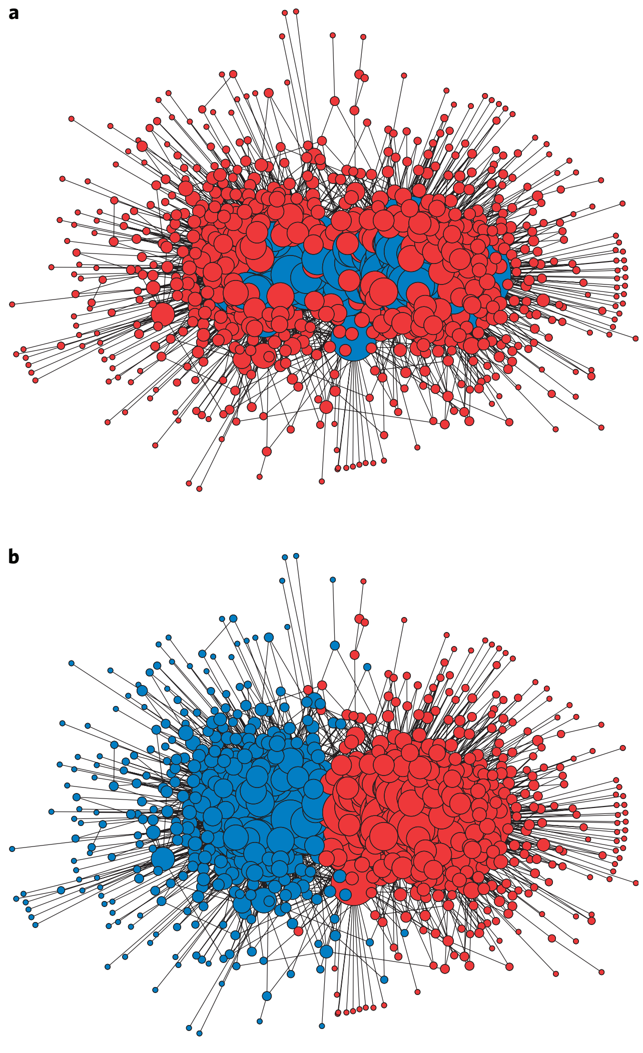

Figure 4 | Analysis of a network of links between web sites about US politics. The two panels represent the divisions found in a network of political weblogs using two different versions of the block model method. **a**, Division into two communities discovered using a fit to the basic block model described in the text, which fails to find the acknowledged division of the network into politically left- and right-leaning communities. **b**, Division using a block model that corrects for the broad distribution of node degrees in the network. This division corresponds closely to the acknowledged one. Figure reproduced with permission from ref. 49, © 2011 APS. Network data taken from ref. 68.

© 2012 Macmillan Publishers Limited. All rights reserved.

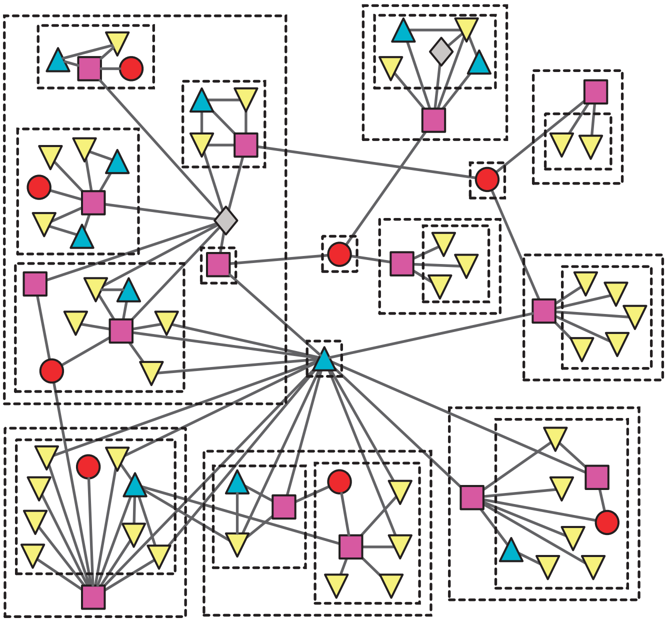

Figure 5 | Hierarchical divisions in a food web of grassland species. Outlined sets of nodes represent groups of species at different levels in the hierarchy. For clarity only two levels in the hierarchy are shown, although five levels were found in some parts of the network. Reproduced from ref. 71.

structures, communities within communities and many others. The field is only just beginning to explore the wide range of possibilities that this approach offers, but Fig. 5 shows one example, drawn from my own work[71]. In this study we examined the food web of a grassland ecosystem—the network of predator–prey interactions between species—and searched for a generalized form of hierarchical community structure in which groups divide into subgroups and subsubgroups and so on. Using a model that employs a tree structure reminiscent of the dendrogram of Fig. 3 to represent the hierarchy of groups, and edge probabilities that depend on shortest paths through the tree, we were able to discover an entire spectrum of structure within the network, spanning the range from small motifs of a few nodes to the size of the entire network. Of particular note in this example is the way in which the method groups host species (squares) with their parasites (yellow triangles), but at the next level in the hierarchy also gathers the parasites separately into their own groups. In some sense, the parasites have more in common with each other than with their host, and hence can be thought of as belonging to a separate group, even though they have no direct interactions with one another through the food web. The calculation realizes this and divides the network accordingly.

Conclusion

The study of network structure and its links with the function and behaviour of complex systems is a large and active field of endeavor, with new results appearing daily and an energetic community of researchers working on both methods and applications. Some of the ideas discussed here are now well established and widely used, whereas others, such as the block-model methods, are being actively researched and developed, and there are many others still that there is not room to describe in this article. The pace of developments is, if anything, accelerating, and the field offers substantial promise for those in physics, biology, the social sciences and elsewhere, for whom the ability to make sense of the structures, large and small, found in networks can open a new window on the behaviour of systems of many kinds.

References

1. Albert, R. & Barabási, A-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74,** 47–97 (2002).
2. Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of networks. *Adv. Phys.* **51,** 1079–1187 (2002).
3. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45,** 167–256 (2003).
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D-U. Complex networks: Structure and dynamics. *Phys. Rep.* **424,** 175–308 (2006).
5. Newman, M. E. J. *Networks: An Introduction* (Oxford Univ. Press, 2010).
6. Cohen, R. & Havlin, S. *Complex Networks: Structure, Stability and Function* (Cambridge Univ. Press, 2010).
7. Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the internet topology. *Comput. Commun. Rev.* **29,** 251–262 (1999).
8. Pastor-Satorras, R. & Vespignani, A. *Evolution and Structure of the Internet* (Cambridge Univ. Press, 2004).
9. Pimm, S. L. *Food Webs* 2nd edn (Univ. Chicago Press, 2002).
10. Pascual, M. & Dunne, J. A. (eds) *Ecological Networks: Linking Structure to Dynamics in Food Webs* (Oxford Univ. Press, 2006).
11. Wasserman, S. & Faust, K. *Social Network Analysis* (Cambridge Univ. Press, 1994).
12. Scott, J. *Social Network Analysis: A Handbook* 2nd edn (Sage, 2000).
13. Costa, L. da F., Rodrigues, F. A., Travieso, G. & Boas, P. R. V. Characterization of complex networks: A survey of measurements. *Adv. Phys.* **56,** 167–242 (2007).
14. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99,** 7821–7826 (2002).
15. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486,** 75–174 (2010).
16. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A-L. The large-scale organization of metabolic networks. *Nature* **407,** 651–654 (2000).
17. Guimerà, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433,** 895–900 (2005).
18. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69,** 026113 (2004).
19. Flake, G. W., Lawrence, S. R., Giles, C. L. & Coetzee, F. M. Self-organization and identification of Web communities. *IEEE Comput.* **35,** 66–71 (2002).
20. Zhou, H. Distance, dissimilarity index, and network community structure. *Phys. Rev. E* **67,** 061901 (2003).
21. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. Natl Acad. Sci. USA* **101,** 2658–2663 (2004).
22. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435,** 814–818 (2005).
23. Bagrow, J. P. & Bollt, E. M. Local method for detecting communities. *Phys. Rev. E* **72,** 046108 (2005).
24. Clauset, A. Finding local community structure in networks. *Phys. Rev. E* **72,** 026132 (2005).
25. Hastings, M. B. Community detection as an inference problem. *Phys. Rev. E* **74,** 035102 (2006).
26. Rosvall, M. & Bergstrom, C. T. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl Acad. Sci. USA* **104,** 7327–7331 (2007).
27. Blondel, V. D., Guillaume, J-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008,** P10008 (2008).
28. Agrawal, G. & Kempe, D. Modularity-maximizing network communities via mathematical programming. *Eur. Phys. J. B* **66,** 409–418 (2008).
29. Hofman, J. M. & Wiggins, C. H. Bayesian approach to network modularity. *Phys. Rev. Lett.* **100,** 258701 (2008).
30. Leskovec, J., Lang, K., Dasgupta, A. & Mahoney, M. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Math.* **6,** 29–123 (2009).
31. Ahn, Y-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466,** 761–764 (2010).
32. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78,** 046110 (2008).
33. Danon, L., Duch, J., Diaz-Guilera, A. & Arenas, A. Comparing community structure identification. *J. Stat. Mech.* P09008 (2005).
34. Lancichinetti, A. & Fortunato, S. Community detection algorithms: A comparative analysis. *Phys. Rev. E* **80,** 056117 (2009).
35. Schaeffer, S. E. Graph clustering. *Comput. Sci. Rev.* **1,** 27–64 (2007).
36. Pothen, A., Simon, H. & Liou, K-P. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.* **11,** 430–452 (1990).
37. Kernighan, B. W. & Lin, S. An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.* **49,** 291–307 (1970).
38. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33,** 452–473 (1977).

© 2012 Macmillan Publishers Limited. All rights reserved.

39. White, D. R. & Harary, F. The cohesiveness of blocks in social networks: Connectivity and conditional density. *Sociol. Methodol.* **31,** 305–359 (2001).
40. Duch, J. & Arenas, A. Community detection in complex networks using extremal optimization. *Phys. Rev. E* **72,** 027104 (2005).
41. Wilkinson, D. M. & Huberman, B. A. A method for finding communities of related genes. *Proc. Natl Acad. Sci. USA* **101,** 5241–5248 (2004).
42. Wu, F. & Huberman, B. A. Finding communities in linear time: A physics approach. *Eur. Phys. J. B* **38,** 331–338 (2004).
43. Rosvall, M. & Bergstrom, C. T. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS One* **6,** e18209 (2011).
44. Zhou, H. & Lipowsky, R. *Network Brownian Motion: A New Method to Measure Vertex–Vertex Proximity and to Identify Communities and Subcommunities* 1062–1069 (Lecture Notes in Computer Science, Vol. 3038, Springer, 2004).
45. Pons, P. & Latapy, M. *Proc. 20th International Symposium on Computer and Information Sciences* 284–293 (Lecture Notes in Computer Science, Vol. 3733, Springer, 2005).
46. Reichardt, J. & Bornholdt, S. Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.* **93,** 218701 (2004).
47. Boccaletti, S., Ivanchenko, M., Latora, V., Pluchino, A. & Rapisarda, A. Detection of complex networks modularity by dynamical clustering. *Phys. Rev. E* **75,** 045102 (2007).
48. Karckhardt, D. & Stern, R. Informal networks and organizational crises: An experimental simulation. *Soc. Psychol. Q.* **51,** 123–140 (1988).
49. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83,** 016107 (2011).
50. Li, Z., Zhang, S., Wang, R-S., Zhang, X-S. & Chen, L. Quantitative function for community detection. *Phys. Rev. E* **77,** 036109 (2008).
51. Newman, M. E. J. Mixing patterns in networks. *Phys. Rev. E* **67,** 026126 (2003).
52. Brandes, U. *et al. Proc. 33rd International Workshop on Graph-Theoretic Concepts in Computer Science* (Lecture Notes in Computer Science,Vol. 4769, Springer, 2007).
53. Medus, A., Acuña, G. & Dorso, C. O. Detection of community structures in networks via global optimization. *Physica A* **358,** 593–604 (2005).
54. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70,** 066111 (2004).
55. Wakita, K. & Tsurumi, T. in *Proc. IADIS International Conference, WWW/Internet 2007* (eds Isaías, P., Nunes, M. B. & Barroso, J.) 153–162 (IADIS Press, 2007).
56. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA* **103,** 8577–8582 (2006).
57. Shuzhuo, L., Yinghui, C., Haifeng, D. & Feldman, M. W. A genetic algorithm with local search strategy for improved detection of community structure. *Complexity* **15,** 53–60 (2010).
58. Fortunato, S. & Barthélémy, M. Resolution limit in community detection. *Proc. Natl Acad. Sci. USA* **104,** 36–41 (2007).
59. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74,** 016110 (2006).
60. Arenas, A., Fernandez, A. & Gomez, S. Analysis of the structure of complex networks at different resolution levels. *New J. Phys.* **10,** 053039 (2008).
61. Breiger, R. L., Boorman, S. A. & Arabie, P. An algorithm for clustering relations data with applications to social network analysis and comparison with multidimensional scaling. *J. Math. Psychol.* **12,** 328–383 (1975).
62. Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: Some first steps. *Soc. Networks* **5,** 109–137 (1983).
63. Snijders, T. A. B. & Nowicki, K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14,** 75–100 (1997).
64. Nowicki, K. & Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.* **96,** 1077–1087 (2001).
65. Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. Mixed membership stochastic blockmodels. *J. Mach. Learning Res.* **9,** 1981–2014 (2008).
66. Goldenberg, A., Zheng, A. X., Feinberg, S. E. & Airoldi, E. M. A survey of statistical network structures. *Found. Trends Mach. Learning* **2,** 1–117 (2009).
67. Bickel, P. J. & Chen, A. A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl Acad. Sci. USA* **106,** 21068–21073 (2009).
68. Adamic, L. A. & Glance, N. *Proc. WWW-2005 Workshop on the Weblogging Ecosystem* (2005).
69. Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl Acad. Sci. USA* **106,** 22073–22078 (2009).
70. Yan, X., Zhu, Y., Rouquier, J-B. & Moore, C. in *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association of Computing Machinery, 2011).
71. Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453,** 98–101 (2008).

Acknowledgements

Some of the work described here was financially supported by the US National Science Foundation under grants DMS–0405348 and DMS–0804778.

Additional information

The author declares no competing financial interests. Reprints and permissions information is available online at http://www.nature.com/reprints.

© 2012 Macmillan Publishers Limited. All rights reserved.