The structure and function of complex networks

M. E. J. Newman

Department of Physics, University of Michigan, Ann Arbor, MI 48109, U.S.A. and
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, U.S.A.

Inspired by empirical studies of networked systems such as the Internet, social networks, and biological networks, researchers have in recent years developed a variety of techniques and models to help us understand or predict the behavior of these systems. Here we review developments in this field, including such concepts as the small-world effect, degree distributions, clustering, network correlations, random graph models, models of network growth and preferential attachment, and dynamical processes taking place on networks.

Contents

Acknowledgments

For useful feedback on early versions of this article, the author would particularly like to thank Lada Adamic, Michelle Girvan, Petter Holme, Randy LeVeque, Sidney Redner, Ricard Solé, Steve Strogatz, Alexei Vázquez, and an anonymous referee. For other helpful conversations and comments about networks thanks go to Lada Adamic, László Barabási, Stefan Bornholdt, Duncan Callaway, Peter Dodds, Jennifer Dunne, Rick Durrett, Stephanie Forrest, Michelle Girvan, Jon Kleinberg, James Moody, Cris Moore, Martina Morris, Juyong Park, Richard Rothenberg, Larry Ruzzo, Matthew Salganik, Len Sander, Steve Strogatz, Alessandro Vespignani, Chris Warren, Duncan Watts, and Barry Wellman. For providing data used in calculations and figures, thanks go to Lada Adamic, László Barabási, Jerry Davis, Jennifer Dunne, Ramón Ferrer i Cancho, Paul Ginsparg, Jerry Grossman, Oleg Khovayko, Hawoong Jeong, David Lipman, Neo Martinez, Stephen Muth, Richard Rothenberg, Ricard Solé, Grigoriy Starchenko, Duncan Watts, Geoffrey West, and Janet Wiener. Figure 2a was kindly provided by Neo Martinez and Richard Williams and Fig. 8 by James Moody. This work was supported in part by the US National Science Foundation under grants DMS–0109086 and DMS–0234188 and by the James S. McDonnell Foundation and the Santa Fe Institute.

I. INTRODUCTION

A *network* is a set of items, which we will call *vertices* or sometimes nodes, with connections between them, called *edges* (Fig. 1). Systems taking the form of networks (also called "graphs" in much of the mathematical literature) abound in the world. Examples include the Internet, the World Wide Web, social networks of acquaintance or other connections between individuals, organizational networks and networks of business relations between companies, neural networks, metabolic networks, food webs, distribution networks such as blood vessels or postal delivery routes, networks of citations between papers, and many others (Fig. 2). This paper reviews recent (and some not-so-recent) work on the structure and function of networked systems such as these.

The study of networks, in the form of mathematical graph theory, is one of the fundamental pillars of discrete mathematics. Euler's celebrated 1735 solution of the Königsberg bridge problem is often cited as the first true proof in the theory of networks, and during the twentieth century graph theory has developed into a substantial body of knowledge.

Networks have also been studied extensively in the social sciences. Typical network studies in sociology involve the circulation of questionnaires, asking respondents to detail their interactions with others. One can then use the responses to reconstruct a network in which vertices represent individuals and edges the interactions between them. Typical social network studies address issues of centrality (which individuals are best connected to others or have most influence) and connectivity (whether and how individuals are connected to one another through the network).

Recent years however have witnessed a substantial new movement in network research, with the focus shifting away from the analysis of single small graphs and the properties of individual vertices or edges within such graphs to consideration of large-scale statistical properties of graphs. This new approach has been driven largely by the availability of computers and communication networks that allow us to gather and analyze data on a scale far larger than previously possible. Where studies used to look at networks of maybe tens or in extreme cases hundreds of vertices, it is not uncommon now to see networks with millions or even billions of vertices. This change of scale forces upon us a corresponding change in our analytic approach. Many of the questions that might previously have been asked in studies of small networks are simply not useful in much larger networks. A social network analyst might have asked, "Which vertex in this network would prove most crucial to the network's connectivity if it were removed?" But such a question has little meaning in most networks of a million vertices—no single vertex in such a network will have much effect at all when removed. On the other hand, one could reasonably ask a question like, "What percentage of vertices need to be removed to substantially affect network connectivity in some given way?" and this type of statistical question has real meaning even in a very large network.

However, there is another reason why our approach to the study of networks has changed in recent years, a reason whose importance should not be underestimated, although it often is. For networks of tens or hundreds of vertices, it is a relatively straightforward matter to draw a picture of the network with actual points and lines (Fig. 2) and to answer specific questions about network structure by examining this picture. This has been one of the primary methods of network analysts since the field began. The human eye is an analytic tool of remarkable power, and eyeballing pictures of networks is an excellent way to gain an understanding of their structure. With a network of a million or a billion vertices however, this approach is useless. One simply cannot draw a meaningful picture of a million vertices, even with modern 3D computer rendering tools, and therefore direct analysis by eye is hopeless. The recent development of statistical methods for quantifying large networks is to a large extent an attempt to find something to play the part played by the eye in the network analysis of the twentieth century. Statistical methods answer the question, "How can I tell what this network looks like, when I can't actually look at it?"

The body of theory that is the primary focus of this review aims to do three things. First, it aims to find statistical properties, such as path lengths and degree distributions, that characterize the structure and behavior of networked systems, and to suggest appropriate ways to measure these properties. Second, it aims to create models of networks that can help us to understand the meaning of these properties—how they came to be as they are, and how they interact with one another. Third, it aims to predict what the behavior of networked systems will be on the basis of measured structural properties and the local rules governing individual vertices. How for example will network structure affect traffic on the Internet, or the performance of a Web search engine, or the dynamics of social or biological systems? As we will see, the scientific community has, by drawing on ideas from a broad variety of disciplines, made an excellent start on the first two of these aims, the characterization and modeling of network structure. Studies of the effects of structure on system behavior on the other hand are still in their infancy. It remains to be seen what the crucial theoretical developments will be in this area.

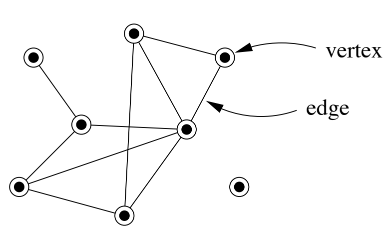

FIG. 1 A small example network with eight vertices and ten edges.

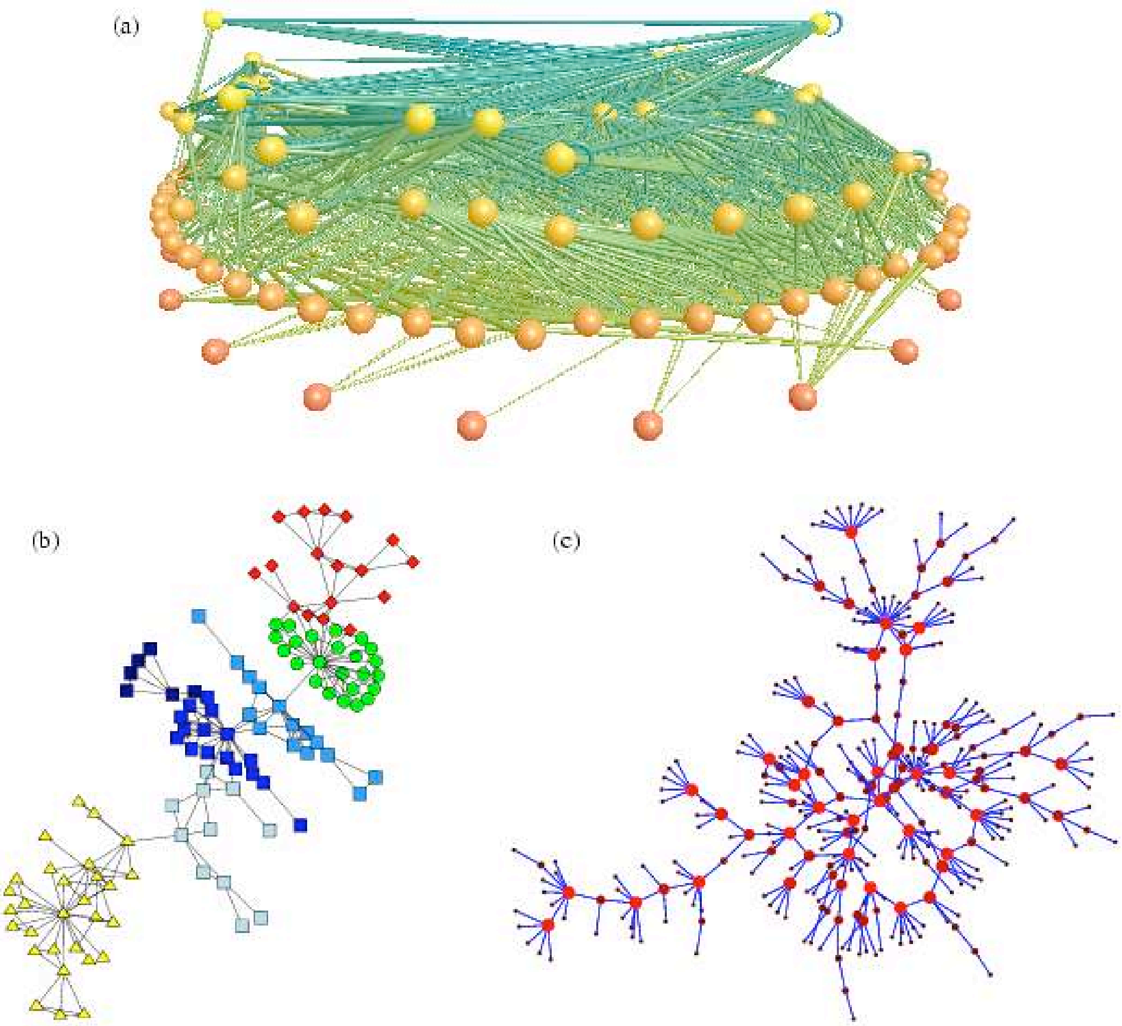

FIG. 2 Three examples of the kinds of networks that are the topic of this review. (a) A food web of predator-prey interactions between species in a freshwater lake [272]. Picture courtesy of Neo Martinez and Richard Williams. (b) The network of collaborations between scientists at a private research institution [171]. (c) A network of sexual contacts between individuals in the study by Potterat *et al.* [342].

A. Types of networks

A set of vertices joined by edges is only the simplest type of network; there are many ways in which networks may be more complex than this (Fig. 3). For instance, there may be more than one different type of vertex in a network, or more than one different type of edge. And vertices or edges may have a variety of properties, numerical or otherwise, associated with them. Taking the example of a social network of people, the vertices may represent men or women, people of different nationalities, locations, ages, incomes, or many other things. Edges may represent friendship, but they could also represent animosity, or professional acquaintance, or geographical proximity. They can carry weights, representing, say, how well two people know each other. They can also be directed, pointing in only one direction. Graphs composed of directed edges are themselves called directed graphs or sometimes *digraphs*, for short. A graph representing telephone calls or email messages between individuals would be directed, since each message goes in only one direction. Directed graphs can be either cyclic, meaning they contain closed loops of edges, or acyclic meaning they do not. Some networks, such as food webs, are approximately but not perfectly acyclic.

One can also have *hyperedges*—edges that join more than two vertices together. Graphs containing such edges are called *hypergraphs*. Hyperedges could be used to indicate family ties in a social network for example—n individuals connected to each other by virtue of belonging to the same immediate family could be represented by an n-edge joining them. Graphs may also be naturally partitioned in various ways. We will see a number of examples in this review of *bipartite graphs*: graphs that contain vertices of two distinct types, with edges running only between unlike types. So-called *affiliation networks*

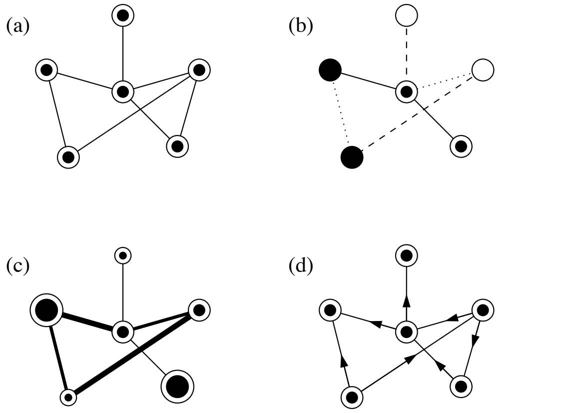

FIG. 3 Examples of various types of networks: (a) an undirected network with only a single type of vertex and a single type of edge; (b) a network with a number of discrete vertex and edge types; (c) a network with varying vertex and edge weights; (d) a directed network in which each edge has a direction.

in which people are joined together by common membership of groups take this form, the two types of vertices representing the people and the groups. Graphs may also evolve over time, with vertices or edges appearing or disappearing, or values defined on those vertices and edges changing. And there are many other levels of sophistication one can add. The study of networks is by no means a complete science yet, and many of the possibilities have yet to be explored in depth, but we will see examples of at least some of the variations described here in the work reviewed in this paper.

The jargon of the study of networks is unfortunately confused by differing usages among investigators from different fields. To avoid (or at least reduce) confusion, we give in Table I a short glossary of terms as they are used in this paper.

B. Other resources

A number of other reviews of this area have appeared recently, which the reader may wish to consult. Albert and Barabási [13] and Dorogovtsev and Mendes [120] have given extensive pedagogical reviews focusing on the physics literature. Both devote the larger part of their attention to the models of growing graphs that we describe in Sec. VII. Shorter reviews taking other viewpoints have been given by Newman [309] and Hayes [189, 190], who both concentrate on the so-called "small-world" models (see Sec. VI), and by Strogatz [387], who includes an interesting discussion of the behavior of dynamical systems on networks.

A number of books also make worthwhile reading. Dorogovtsev and Mendes [122] have expanded their above-mentioned review into a book, which again focuses on models of growing graphs. The edited volumes by Bornholdt and Schuster [70] and by Pastor-Satorras and Rubi [330] both contain contributed essays on various topics by leading researchers. Detailed treatments of many of the topics covered in the present work can be found there. The book by Newman *et al.* [320] is a collection of previously published papers, and also contains some review material by the editors.

Three popular books on the subject of networks merit a mention. Albert-László Barabási's *Linked* [31] gives a personal account of recent developments in the study of networks, focusing particularly on Barabási's work on scale-free networks. Duncan Watts's *Six Degrees* [414] gives a sociologist's view, partly historical, of discoveries old and new. Mark Buchanan's *Nexus* [76] gives an entertaining portrait of the field from the point of view of a science journalist.

Farther afield, there are a variety of books on the study of networks in particular fields. Within graph theory the books by Harary [188] and by Bollobás [62] are widely cited and among social network theorists the books by Wasserman and Faust [409] and by Scott [363]. The book by Ahuja *et al.* [7] is a useful source for information on network algorithms.

C. Outline of the review

The outline of this paper is as follows. In Sec. II we describe empirical studies of the structure of networks, including social networks, information networks, technological networks and biological networks. In Sec. III we describe some of the common properties that are observed in many of these networks, how they are measured, and why they are believed to be important for the functioning of networked systems. Sections IV to VII form the heart of the review. They describe work on the mathematical modeling of networks, including random graph models and their generalizations, exponential random graphs, p^* models and Markov graphs, the small-world model and its variations, and models of growing graphs including preferential attachment models and their many variations. In Sec. VIII we discuss the progress, such as it is, that has been made on the study of processes taking place on networks, including epidemic processes, network failure, models displaying phase transitions, and dynamical systems like random Boolean networks and cellular automata. In Sec. IX we give our conclusions and point to directions for future research.

II. NETWORKS IN THE REAL WORLD

In this section we look at what is known about the structure of networks of different types. Recent work on the mathematics of networks has been driven largely by observations of the properties of actual networks and attempts to model them, so network data are the obvious starting point for a review such as this. It also makes sense to examine simultaneously data from dif-

Vertex (pl. vertices): The fundamental unit of a network, also called a site (physics), a node (computer science), or an actor (sociology).

Edge: The line connecting two vertices. Also called a bond (physics), a link (computer science), or a tie (sociology).

Directed/undirected: An edge is directed if it runs in only one direction (such as a one-way road between two points), and undirected if it runs in both directions. Directed edges, which are sometimes called *arcs*, can be thought of as sporting arrows indicating their orientation. A graph is directed if all of its edges are directed. An undirected graph can be represented by a directed one having two edges between each pair of connected vertices, one in each direction.

Degree: The number of edges connected to a vertex. Note that the degree is not necessarily equal to the number of vertices adjacent to a vertex, since there may be more than one edge between any two vertices. In a few recent articles, the degree is referred to as the "connectivity" of a vertex, but we avoid this usage because the word connectivity already has another meaning in graph theory. A directed graph has both an in-degree and an out-degree for each vertex, which are the numbers of in-coming and out-going edges respectively.

Component: The component to which a vertex belongs is that set of vertices that can be reached from it by paths running along edges of the graph. In a directed graph a vertex has both an in-component and an out-component, which are the sets of vertices from which the vertex can be reached and which can be reached from it.

Geodesic path: A geodesic path is the shortest path through the network from one vertex to another. Note that there may be and often is more than one geodesic path between two vertices.

Diameter: The diameter of a network is the length (in number of edges) of the longest geodesic path between any two vertices. A few authors have also used this term to mean the *average* geodesic distance in a graph, although strictly the two quantities are quite distinct.

TABLE I A short glossary of terms.

ferent kinds of networks. One of the principal thrusts of recent work in this area, inspired particularly by a groundbreaking 1998 paper by Watts and Strogatz [416], has been the comparative study of networks from different branches of science, with emphasis on properties that are common to many of them and the mathematical developments that mirror those properties. We here divide our summary into four loose categories of networks: social networks, information networks, technological networks and biological networks.

A. Social networks

A social network is a set of people or groups of people with some pattern of contacts or interactions between them [363, 409]. The patterns of friendships between individuals [296, 348], business relationships between companies [269, 286], and intermarriages between families [327] are all examples of networks that have been studied in the past.[1] Of the academic disciplines the social sciences have the longest history of the substantial quantitative study of real-world networks [162, 363]. Of particular note among the early works on the subject are: Jacob Moreno's work in the 1920s and 30s on friendship patterns within small groups [296]; the so-called "southern women study" of Davis *et al.* [103], which focused on the social circles of women in an unnamed city in the American south in 1936; the study by Elton Mayo and colleagues of social networks of factory workers in the late 1930s in Chicago [357]; the mathematical models of Anatol Rapoport [346], who was one of the first theorists, perhaps *the* first, to stress the importance of the degree distribution in networks of all kinds, not just social networks; and the studies of friendship networks of school children by Rapoport and others [149, 348]. In more recent years, studies of business communities [167, 168, 269] and of patterns of sexual contacts [45, 218, 243, 266, 303, 342] have attracted particular attention.

Another important set of experiments are the famous

[1] Occasionally social networks of animals have been investigated also, such as dolphins [96], not to mention networks of fictional characters, such as the protagonists of Tolstoy's *Anna Karenina* [244] or Marvel Comics superheroes [10].

"small-world" experiments of Milgram [283, 393]. No actual networks were reconstructed in these experiments, but nonetheless they tell us about network structure. The experiments probed the distribution of path lengths in an acquaintance network by asking participants to pass a letter[2] to one of their first-name acquaintances in an attempt to get it to an assigned target individual. Most of the letters in the experiment were lost, but about a quarter reached the target and passed on average through the hands of only about six people in doing so. This experiment was the origin of the popular concept of the "six degrees of separation," although that phrase did not appear in Milgram's writing, being coined some decades later by Guare [183]. A brief but useful early review of Milgram's work and work stemming from it was given by Garfield [169].

Traditional social network studies often suffer from problems of inaccuracy, subjectivity, and small sample size. With the exception of a few ingenious indirect studies such as Milgram's, data collection is usually carried out by querying participants directly using questionnaires or interviews. Such methods are labor-intensive and therefore limit the size of the network that can be observed. Survey data are, moreover, influenced by subjective biases on the part of respondents; how one respondent defines a friend for example could be quite different from how another does. Although much effort is put into eliminating possible sources of inconsistency, it is generally accepted that there are large and essentially uncontrolled errors in most of these studies. A review of the issues has been given by Marsden [271].

Because of these problems many researchers have turned to other methods for probing social networks. One source of copious and relatively reliable data is collaboration networks. These are typically affiliation networks in which participants collaborate in groups of one kind or another, and links between pairs of individuals are established by common group membership. A classic, though rather frivolous, example of such a network is the collaboration network of film actors, which is thoroughly documented in the online Internet Movie Database.[3] In this network actors collaborate in films and two actors are considered connected if they have appeared in a film together. Statistical properties of this network have been analyzed by a number of authors [4, 20, 323, 416]. Other examples of networks of this type are networks of company directors, in which two directors are linked if they belong to the same board of directors [104, 105, 269], networks of coauthorship among academics, in which individuals are linked if they have coauthored one or more papers [36, 43, 68, 107, 182, 279, 292, 311–313], and coappearance networks in which individuals are linked by mention in the same context, particularly on Web pages [3, 227] or in newspaper articles [99] (see Fig. 2b).

Another source of reliable data about personal connections between people is communication records of certain kinds. For example, one could construct a network in which each (directed) edge between two people represented a letter or package sent by mail from one to the other. No study of such a network has been published as far as we are aware, but some similar things have. Aiello *et al.* [8, 9] have analyzed a network of telephone calls made over the AT&T long-distance network on a single day. The vertices of this network represent telephone numbers and the directed edges calls from one number to another. Even for just a single day this graph is enormous, having about 50 million vertices, one of the largest graphs yet studied after the graph of the World Wide Web. Ebel *et al.* [136] have reconstructed the pattern of email communications between five thousand students at Kiel University from logs maintained by email servers. In this network the vertices represent email addresses and directed edges represent a message passing from one address to another. Email networks have also been studied by Newman *et al.* [321] and by Guimerà *et al.* [185], and similar networks have been constructed for an "instant messaging" system by Smith [371], and for an Internet community Web site by Holme *et al.* [196]. Dodds *et al.* [110] have carried out an email version of Milgram's small-world experiment in which participants were asked to forward an email message to one of their friends in an effort to get the message ultimately to some chosen target individual. Response rates for the experiment were quite low, but a few hundred completed chains of messages were recorded, enough to allow various statistical analyses.

B. Information networks

Our second network category is what we will call *information networks* (also sometimes called "knowledge networks"). The classic example of an information network is the network of citations between academic papers [138]. Most learned articles cite previous work by others on related topics. These citations form a network in which the vertices are articles and a directed edge from article A to article B indicates that A cites B. The structure of the citation network then reflects the structure of the information stored at its vertices, hence the term "information network," although certainly there are social aspects to the citation patterns of papers too [420].

Citation networks are acyclic (see Sec. I.A) because papers can only cite other papers that have already been written, not those that have yet to be written. Thus all edges in the network point backwards in time, making closed loops impossible, or at least extremely rare (see Fig. 4).

As an object of scientific study, citation networks have a great advantage in the copious and accurate data available for them. Quantitative study of publication patterns

[2] Actually a folder containing several documents.
[3] `http://www.imdb.com/`

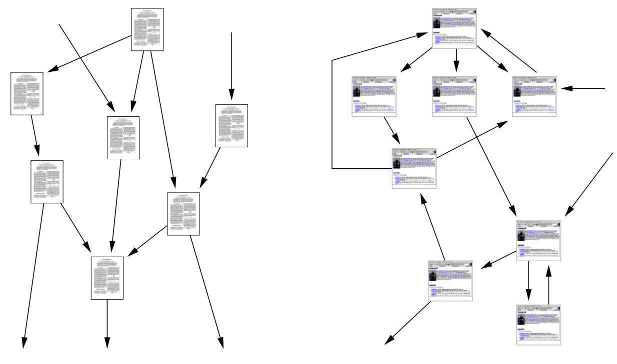

FIG. 4 The two best studied information networks. Left: the citation network of academic papers in which the vertices are papers and the directed edges are citations of one paper by another. Since papers can only cite those that came before them (lower down in the figure) the graph is acyclic—it has no closed loops. Right: the World Wide Web, a network of text pages accessible over the Internet, in which the vertices are pages and the directed edges are hyperlinks. There are no constraints on the Web that forbid cycles and hence it is in general cyclic.

stretches back at least as far as Alfred Lotka's groundbreaking 1926 discovery of the so-called Law of Scientific Productivity, which states that the distribution of the numbers of papers written by individual scientists follows a power law. That is, the number of scientists who have written k papers falls off as $k^{-\alpha}$ for some constant α. (In fact, this result extends to the arts and humanities as well.) The first serious work on citation patterns was conducted in the 1960s as large citation databases became available through the work of Eugene Garfield and other pioneers in the field of bibliometrics. The network formed by citations was discussed in an early paper by Price [343], in which among other things, the author points out for the first time that both the in- and out-degree distributions of the network follow power laws, a far-reaching discovery which we discuss further in Sec. III.C. Many other studies of citation networks have been performed since then, using the ever better resources available in citation databases. Of particular note are the studies by Seglen [364] and Redner [351].[4]

Another very important example of an information network is the World Wide Web, which is a network of Web pages containing information, linked together by hyperlinks from one page to another [203]. The Web should not be confused with the Internet, which is a physical network of computers linked together by optical fibre and other data connections.[5] Unlike a citation network, the World Wide Web is cyclic; there is no natural ordering of sites and no constraints that prevent the appearance of closed loops (Fig. 4). The Web has been very heavily studied since its first appearance in the early 1990s, with the studies by Albert *et al.* [14, 34], Kleinberg *et al.* [241], and Broder *et al.* [74] being particularly influential. The Web also appears to have power-law in- and out-degree distributions (Sec. III.C), as well as a variety of other interesting properties [2, 14, 74, 158, 241, 254].

One important point to notice about the Web is that our data about it come from "crawls" of the network, in which Web pages are found by following hyperlinks from other pages [74]. Our picture of the network structure of the World Wide Web is therefore necessarily biased. A page will only be found if another page points to it,[6] and in a crawl that covers only a part of the Web (as all crawls do at present) pages are more likely to be found the more other pages point to them [263]. This suggests for instance that our measurements of the fraction of pages with low in-degree might be an underestimate.[7] This behavior contrasts with that of a citation network. A paper can appear in the citation indices even if it has never been cited (and in fact a plurality of papers in the indices are never cited).

A few other examples of information networks have been studied to a lesser extent. Jaffe and Trajtenberg [207], for instance, have studied the network of citations between US patents, which is similar in some respects to citations between academic papers. A number of authors have looked at peer-to-peer networks [5, 6, 205], which are virtual networks of computers that allow sharing of files between computer users over local- or wide-area networks. The network of relations between word classes in a thesaurus has been studied by Knuth [244] and more recently by various other authors [234, 304, 384]. This network can be looked upon as an information network—users of a thesaurus "surf" the network from one word to another looking for the particular word that perfectly captures the idea they have in mind. However, it can also be looked at as a conceptual network representing the structure of the language, or possibly even the mental constructs used to represent the language. A number of other semantic word networks have also been investigated [119, 157, 369, 384].

Preference networks provide an example of a bipartite

4 An interesting development in the study of citation patterns has been the arrival of automatic citation "crawlers" that construct citation networks from online papers. Examples include Citeseer (`http://citeseer.nj.nec.com/`), SPIRES (`http://www.slac.stanford.edu/spires/hep/`) and Citebase (`http://citebase.eprints.org/`).

5 While the Web is primarily an information network, it, like citation networks, has social aspects to its structure also [3].

6 This is not always strictly true. Some Web search engines allow the submission of pages by members of the public for inclusion in databases, and such pages need not be the target of links from any other pages. However, such pages also form a very small fraction of all Web pages, and certainly the biases discussed here remain very much present.

7 The degree distribution for the Web shown in Fig. 6 falls off slightly at low values of the in-degree, which may perhaps reflect this bias.

information network. A preference network is a network with two kinds of vertices representing individuals and the objects of their preference, such as books or films, with an edge connecting each individual to the books or films they like. (Preference networks can also be weighted to indicate strength of likes or dislikes.) A widely studied example of a preference network is the *EachMovie* database of film preferences.[8] Networks of this kind form the basis for *collaborative filtering* algorithms and *recommender systems*, which are techniques for predicting new likes or dislikes based on comparison of individuals' preferences with those of others [176, 352, 367]. Collaborative filtering has found considerable commercial success for product recommendation and targeted advertising, particularly with online retailers. Preference networks can also be thought of as social networks, linking not only people to objects, but also people to other people with similar preferences. This approach has been adopted occasionally in the literature [227].

C. Technological networks

Our third class of networks is technological networks, man-made networks designed typically for distribution of some commodity or resource, such as electricity or information. The electric power grid is a good example. This is a network of high-voltage three-phase transmission lines that spans a country or a portion of a country (as opposed to the local low-voltage a.c. power delivery lines that span individual neighborhoods). Statistical studies of power grids have been made by, for example, Watts and Strogatz [412, 416] and Amaral *et al.* [20]. Other distribution networks that have been studied include the network of airline routes [20], and networks of roads [221], railways [262, 366] and pedestrian traffic [87]. River networks could be regarded as a naturally occurring form of distribution network (actually a collection network) [111, 270, 353, 356], as could the vascular networks discussed in Sec. II.D. The telephone network and delivery networks such as those used by the post-office or parcel delivery companies also fall into this general category and are presumably studied within the relevant corporations, if not yet by academic researchers. (We distinguish here between the physical telephone network of wires and cables and the network of who calls whom, discussed in Sec. II.A.) Electronic circuits [155] fall somewhere between distribution and communication networks.

Another very widely studied technological network is the Internet, i.e., the network of physical connections between computers. Since there is a large and ever-changing number of computers on the Internet, the structure of the network is usually examined at a coarse-grained level, either the level of routers, special-purpose computers on the network that control the movement of data, or "autonomous systems," which are groups of computers within which networking is handled locally, but between which data flows over the public Internet. The computers at a single company or university would probably form a single autonomous system—autonomous systems often correspond roughly with domain names.

In fact, the network of physical connections on the Internet is not easy to discover since the infrastructure is maintained by many separate organizations. Typically therefore, researchers reconstruct the network by reasoning from large samples of point-to-point data routes. So-called "traceroute" programs can report the sequence of network nodes that a data packet passes through when traveling between two points and if we assume an edge in the network between any two consecutive nodes along such a path then a sufficiently large sample of paths will give us a fairly complete picture of the entire network. There may however be some edges that never get sampled, so the reconstruction is typically a good, but not perfect, representation of the true physical structure of the Internet. Studies of Internet structure have been carried out by, among others, Faloutsos *et al.* [148], Broida and Claffy [75] and Chen *et al.* [86].

D. Biological networks

A number of biological systems can be usefully represented as networks. Perhaps the classic example of a biological network is the network of metabolic pathways, which is a representation of metabolic substrates and products with directed edges joining them if a known metabolic reaction exists that acts on a given substrate and produces a given product. Most of us will probably have seen at some point the giant maps of metabolic pathways that many molecular biologists pin to their walls.[9] Studies of the statistical properties of metabolic networks have been performed by, for example, Jeong *et al.* [214, 340], Fell and Wagner [153, 405], and Stelling *et al.* [383]. A separate network is the network of mechanistic physical interactions between proteins (as opposed to chemical reactions among metabolites), which is usually referred to as a protein interaction network. Interaction networks have been studied by a number of authors [206, 212, 274, 376, 394].

Another important class of biological network is the genetic regulatory network. The expression of a gene, i.e., the production by transcription and translation of the protein for which the gene codes, can be controlled

[8] `http://research.compaq.com/SRC/eachmovie/`

[9] The standard chart of the metabolic network is somewhat misleading. For reasons of clarity and aesthetics, many metabolites appear in more than one place on the chart, so that some pairs of vertices are actually the same vertex.

by the presence of other proteins, both activators and inhibitors, so that the genome itself forms a switching network with vertices representing the proteins and directed edges representing dependence of protein production on the proteins at other vertices. The statistical structure of regulatory networks has been studied recently by various authors [152, 184, 368]. Genetic regulatory networks were in fact one of the first networked dynamical systems for which large-scale modeling attempts were made. The early work on random Boolean nets by Kauffman [224–226] is a classic in this field, and anticipated recent developments by several decades.

Another much studied example of a biological network is the food web, in which the vertices represent species in an ecosystem and a directed edge from species A to species B indicates that A preys on B [91, 339]—see Fig. 2a. (Sometimes the relationship is drawn the other way around, because ecologists tend to think in terms of energy or carbon flows through food webs; a predator-prey interaction is thus drawn as an arrow pointing from prey to predator, indicating energy flow from prey to predator when the prey is eaten.) Construction of complete food webs is a laborious business, but a number of quite extensive data sets have become available in recent years [27, 177, 204, 272]. Statistical studies of the topologies of food webs have been carried out by Solé and Montoya [290, 375], Camacho *et al.* [82] and Dunne *et al.* [132, 133, 423], among others. A particularly thorough study of webs of plants and herbivores has been conducted by Jordano *et al.* [219], which includes statistics for no less than 53 different networks.

Neural networks are another class of biological networks of considerable importance. Measuring the topology of real neural networks is extremely difficult, but has been done successfully in a few cases. The best known example is the reconstruction of the 282-neuron neural network of the nematode *C. Elegans* by White *et al.* [421]. The network structure of the brain at larger scales than individual neurons—functional areas and pathways—has been investigated by Sporns *et al.* [379, 380].

Blood vessels and the equivalent vascular networks in plants form the foundation for one of the most successful theoretical models of the effects of network structure on the behavior of a networked system, the theory of biological allometry [29, 417, 418], although we are not aware of any quantitative studies of their statistical structure.

Finally we mention two examples of networks from the physical sciences, the network of free energy minima and saddle points in glasses [130] and the network of conformations of polymers and the transitions between them [361], both of which appear to have some interesting structural properties.

III. PROPERTIES OF NETWORKS

Perhaps the simplest useful model of a network is the random graph, first studied by Rapoport [346, 347, 378] and by Erdős and Rényi [141–143], which we describe in Sec. IV.A. In this model, undirected edges are placed at random between a fixed number n of vertices to create a network in which each of the $\frac{1}{2}n(n-1)$ possible edges is independently present with some probability p, and the number of edges connected to each vertex—the degree of the vertex—is distributed according to a binomial distribution, or a Poisson distribution in the limit of large n. The random graph has been well studied by mathematicians [63, 211, 223] and many results, both approximate and exact, have been proved rigorously. Most of the interesting features of real-world networks that have attracted the attention of researchers in the last few years however concern the ways in which networks are *not* like random graphs. Real networks are non-random in some revealing ways that suggest both possible mechanisms that could be guiding network formation, and possible ways in which we could exploit network structure to achieve certain aims. In this section we describe some features that appear to be common to networks of many different types.

A. The small-world effect

In Sec. II.A we described the famous experiments carried out by Stanley Milgram in the 1960s, in which letters passed from person to person were able to reach a designated target individual in only a small number of steps—around six in the published cases. This result is one of the first direct demonstrations of the *small-world effect*, the fact that most pairs of vertices in most networks seem to be connected by a short path through the network.

The existence of the small-world effect had been speculated upon before Milgram's work, notably in a remarkable 1929 short story by the Hungarian writer Frigyes Karinthy [222], and more rigorously in the mathematical work of Pool and Kochen [341] which, although published after Milgram's studies, was in circulation in preprint form for a decade before Milgram took up the problem. Nowadays, the small-world effect has been studied and verified directly in a large number of different networks.

Consider an undirected network, and let us define ℓ to be the mean geodesic (i.e., shortest) distance between vertex pairs in a network:

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}, \tag{1}$$

where d_{ij} is the geodesic distance from vertex i to vertex j. Notice that we have included the distance from each vertex to itself (which is zero) in this average. This is mathematically convenient for a number of reasons, but not all authors do it. In any case, its inclusion simply multiplies ℓ by $(n-1)/(n+1)$ and hence gives a correction of order n^{-1}, which is often negligible for practical purposes.

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

The quantity ℓ can be measured for a network of n vertices and m edges in time $O(mn)$ using simple breadth-first search [7], also called a "burning algorithm" in the physics literature. In Table II, we show values of ℓ taken from the literature for a variety of different networks. As the table shows, the values are in all cases quite small—much smaller than the number n of vertices, for instance.

The definition (1) of ℓ is problematic in networks that have more than one component. In such cases, there exist vertex pairs that have no connecting path. Conventionally one assigns infinite geodesic distance to such pairs, but then the value of ℓ also becomes infinite. To avoid this problem one usually defines ℓ on such networks to be the mean geodesic distance between all pairs that have a connecting path. Pairs that fall in two different components are excluded from the average. The figures in Table II were all calculated in this way. An alternative and perhaps more satisfactory approach is to define ℓ to be the "harmonic mean" geodesic distance between all pairs, i.e., the reciprocal of the average of the reciprocals:

$$\ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}. \tag{2}$$

Infinite values of d_{ij} then contribute nothing to the sum. This approach has been adopted only occasionally in network calculations [260], but perhaps should be used more often.

The small-world effect has obvious implications for the dynamics of processes taking place on networks. For example, if one considers the spread of information, or indeed anything else, across a network, the small-world effect implies that that spread will be fast on most real-world networks. If it takes only six steps for a rumor to spread from any person to any other, for instance, then the rumor will spread much faster than if it takes a hundred steps, or a million. This affects the number of "hops" a packet must make to get from one computer to another on the Internet, the number of legs of a journey for an air or train traveler, the time it takes for a disease to spread throughout a population, and so forth. The small-world effect also underlies some well-known parlor games, particularly the calculation of Erdős numbers [107] and Bacon numbers.[10]

On the other hand, the small-world effect is also mathematically obvious. If the number of vertices within a distance r of a typical central vertex grows exponentially with r—and this is true of many networks, including the random graph (Sec. IV.A)—then the value of ℓ will increase as $\log n$. In recent years the term "small-world effect" has thus taken on a more precise meaning: networks are said to show the small-world effect if the value of ℓ scales logarithmically or slower with network size for fixed mean degree. Logarithmic scaling can be proved for a variety of network models [61, 63, 88, 127, 164] and has also been observed in various real-world networks [13, 312, 313]. Some networks have mean vertex–vertex distances that increase slower than $\log n$. Bollobás and Riordan [64] have shown that networks with power-law degree distributions (Sec. III.C) have values of ℓ that increase no faster than $\log n/\log\log n$ (see also Ref. 164), and Cohen and Havlin [95] have given arguments that suggest that the actual variation may be slower even than this.

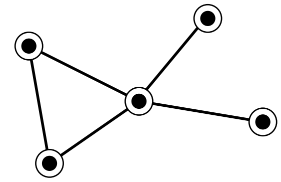

FIG. 5 Illustration of the definition of the clustering coefficient C, Eq. (3). This network has one triangle and eight connected triples, and therefore has a clustering coefficient of $3 \times 1/8 = \frac{3}{8}$. The individual vertices have local clustering coefficients, Eq. (5), of 1, 1, $\frac{1}{6}$, 0 and 0, for a mean value, Eq. (6), of $C = \frac{13}{30}$.

B. Transitivity or clustering

A clear deviation from the behavior of the random graph can be seen in the property of network transitivity, sometimes also called clustering, although the latter term also has another meaning in the study of networks (see Sec. III.G) and so can be confusing. In many networks it is found that if vertex A is connected to vertex B and vertex B to vertex C, then there is a heightened probability that vertex A will also be connected to vertex C. In the language of social networks, the friend of your friend is likely also to be your friend. In terms of network topology, transitivity means the presence of a heightened number of triangles in the network—sets of three vertices each of which is connected to each of the others. It can be quantified by defining a clustering coefficient C thus:

$$C = \frac{3\times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}}, \tag{3}$$

where a "connected triple" means a single vertex with edges running to an unordered pair of others (see Fig. 5).

In effect, C measures the fraction of triples that have their third edge filled in to complete the triangle. The factor of three in the numerator accounts for the fact that each triangle contributes to three triples and ensures that C lies in the range $0 \leq C \leq 1$. In simple terms, C is the mean probability that two vertices that are network neighbors of the same other vertex will themselves be neighbors. It can also be written in the form

$$C = \frac{6\times \text{ number of triangles in the network}}{\text{number of paths of length two}}, \tag{4}$$

[10] http://www.cs.virginia.edu/oracle/

where a path of length two refers to a directed path starting from a specified vertex. This definition shows that C is also the mean probability that the friend of your friend is also your friend.

The definition of C given here has been widely used in the sociology literature, where it is referred to as the "fraction of transitive triples."[11] In the mathematical and physical literature it seems to have been first discussed by Barrat and Weigt [40].

An alternative definition of the clustering coefficient, also widely used, has been given by Watts and Strogatz [416], who proposed defining a local value

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}. \tag{5}$$

For vertices with degree 0 or 1, for which both numerator and denominator are zero, we put $C_i = 0$. Then the clustering coefficient for the whole network is the average

$$C = \frac{1}{n} \sum_i C_i. \tag{6}$$

This definition effectively reverses the order of the operations of taking the ratio of triangles to triples and of averaging over vertices—one here calculates the mean of the ratio, rather than the ratio of the means. It tends to weight the contributions of low-degree vertices more heavily, because such vertices have a small denominator in Eq. (5) and hence can give quite different results from Eq. (3). In Table II we give both measures for a number of networks (denoted $C^{(1)}$ and $C^{(2)}$ in the table). Normally our first definition (3) is easier to calculate analytically, but (6) is easily calculated on a computer and has found wide use in numerical studies and data analysis. It is important when reading (or writing) literature in this area to be clear about which definition of the clustering coefficient is in use. The difference between the two is illustrated in Fig. 5.

The local clustering C_i above has been used quite widely in its own right in the sociological literature, where it is referred to as the "network density" [363]. Its dependence on the degree k_i of the central vertex i has been studied by Dorogovtsev *et al.* [113] and Szabó *et al.* [389]; both groups found that C_i falls off with k_i approximately as k_i^{-1} for certain models of scale-free networks (Sec. III.C.1). Similar behavior has also been observed empirically in real-world networks [349, 350, 397].

In general, regardless of which definition of the clustering coefficient is used, the values tend to be considerably higher than for a random graph with a similar number of vertices and edges. Indeed, it is suspected that for many types of networks the probability that the friend of your friend is also your friend should tend to a non-zero limit as the network becomes large, so that $C = \mathrm{O}(1)$ as $n \to \infty$.[12] On the random graph, by contrast, $C = \mathrm{O}(n^{-1})$ for large n (either definition of C) and hence the real-world and random graph values can be expected to differ by a factor of order n. This point is discussed further in Sec. IV.A.

The clustering coefficient measures the density of triangles in a network. An obvious generalization is to ask about the density of longer loops also: loops of length four and above. A number of authors have looked at such higher order clustering coefficients [54, 79, 165, 172, 317], although there is so far no clean theory, similar to a cumulant expansion, that separates the independent contributions of the various orders from one another. If more than one edge is permitted between a pair of vertices, then there is also a lower order clustering coefficient that describes the density of loops of length two. This coefficient is particularly important in directed graphs where the two edges in question can point in opposite directions. The probability that two vertices in a directed network point to each other is called the *reciprocity* and is often measured in directed social networks [363, 409]. It has been examined occasionally in other contexts too, such as the World Wide Web [3, 137] and email networks [321].

C. Degree distributions

Recall that the degree of a vertex in a network is the number of edges incident on (i.e., connected to) that vertex. We define p_k to be the fraction of vertices in the network that have degree k. Equivalently, p_k is the probability that a vertex chosen uniformly at random has degree k. A plot of p_k for any given network can be formed by making a histogram of the degrees of vertices. This histogram is the degree distribution for the network. In a random graph of the type studied by Erdős and Rényi [141–143], each edge is present or absent with equal probability, and hence the degree distribution is, as mentioned earlier, binomial, or Poisson in the limit of large graph size. Real-world networks are mostly found to be very unlike the random graph in their degree distributions. Far from having a Poisson distribution, the degrees of the vertices in most networks are highly right-skewed, meaning that their distribution has a long right tail of values that are far above the mean.

Measuring this tail is somewhat tricky. Although in theory one just has to construct a histogram of the degrees, in practice one rarely has enough measurements to get good statistics in the tail, and direct histograms are thus usually rather noisy (see the histograms in Refs. 74,

[11] For example, the standard network analysis program UCInet includes a function to calculate this quantity for any network.

[12] An exception is scale-free networks with $C_i \sim k_i^{-1}$, as described above. For such networks Eq. (3) tends to zero as $n \to \infty$, although Eq. (6) is still finite.

148 and 343 for example). There are two accepted ways to get around this problem. One is to constructed a histogram in which the bin sizes increase exponentially with degree. For example the first few bins might cover degree ranges 1, 2–3, 4–7, 8–15, and so on. The number of samples in each bin is then divided by the width of the bin to normalize the measurement. This method of constructing a histogram is often used when the histogram is to be plotted with a logarithmic degree scale, so that the widths of the bins will appear even. Because the bins get wider as we get out into the tail, the problems with statistics are reduced, although they are still present to some extent as long as p_k falls off faster than k^{-1}, which it must if the distribution is to be integrable.

An alternative way of presenting degree data is to make a plot of the cumulative distribution function

$$P_k = \sum_{k'=k}^{\infty} p_{k'}, \tag{7}$$

which is the probability that the degree is greater than or equal to k. Such a plot has the advantage that all the original data are represented. When we make a conventional histogram by binning, any differences between the values of data points that fall in the same bin are lost. The cumulative distribution function does not suffer from this problem. The cumulative distribution also reduces the noise in the tail. On the downside, the plot doesn't give a direct visualization of the degree distribution itself, and adjacent points on the plot are not statistically independent, making correct fits to the data tricky.

In Fig. 6 we show cumulative distributions of degree for a number of the networks described in Sec. II. As the figure shows, the distributions are indeed all right-skewed. Many of them follow power laws in their tails: $p_k \sim k^{-\alpha}$ for some constant exponent α. Note that such power-law distributions show up as power laws in the cumulative distributions also, but with exponent $\alpha - 1$ rather than α:

$$P_k \sim \sum_{k'=k}^{\infty} k'^{-\alpha} \sim k^{-(\alpha-1)}. \tag{8}$$

Some of the other distributions have exponential tails: $p_k \sim \mathrm{e}^{-k/\kappa}$. These also give exponentials in the cumulative distribution, but with the *same* exponent:

$$P_k = \sum_{k'=k}^{\infty} p_k \sim \sum_{k'=k}^{\infty} \mathrm{e}^{-k'/\kappa} \sim \mathrm{e}^{-k/\kappa}. \tag{9}$$

This makes power-law and exponential distributions particularly easy to spot experimentally, by plotting the corresponding cumulative distributions on logarithmic scales (for power laws) or semi-logarithmic scales (for exponentials).

For other types of networks degree distributions can be more complicated. For bipartite graphs, for instance (Sec. I.A), there are two degree distributions, one for each type of vertex. For directed graphs each vertex has both an in-degree and an out-degree, and the degree distribution therefore becomes a function p_{jk} of two variables, representing the fraction of vertices that simultaneously have in-degree j and out-degree k. In empirical studies of directed graphs like the Web, researchers have usually given only the individual distributions of in- and out-degree [14, 34, 74], i.e., the distributions derived by summing p_{jk} over one or other of its indices. This however discards much of the information present in the joint distribution. It has been found that in- and out-degrees are quite strongly correlated in some networks [321], which suggests that there is more to be gleaned from the joint distribution than is normally appreciated.

1. Scale-free networks

Networks with power-law degree distributions have been the focus of a great deal of attention in the literature [13, 120, 387]. They are sometimes referred to as *scale-free networks* [32], although it is only their degree distributions that are scale-free;[13] one can and usually does have scales present in other network properties. The earliest published example of a scale-free network is probably Price's network of citations between scientific papers [343] (see Sec. II.B). He quoted a value of $\alpha = 2.5$ to 3 for the exponent of his network. In a later paper he quoted a more accurate figure of $\alpha = 3.04$ [344]. He also found a power-law distribution for the out-degree of the network (number of bibliography entries in each paper), although later work has called this into question [396]. More recently, power-law degree distributions have been observed in a host of other networks, including notably other citation networks [351, 364], the World Wide Web [14, 34, 74], the Internet [86, 148, 401], metabolic networks [212, 214], telephone call graphs [8, 9], and the network of human sexual contacts [218, 266]. The degree distributions of some of these networks are shown in Fig. 6.

Other common functional forms for the degree distribution are exponentials, such as those seen in the power grid [20] and railway networks [366], and power laws with exponential cutoffs, such as those seen in the network of movie actors [20] and some collaboration networks [313]. Note also that while a particular form may be seen in the degree distribution for the network as a whole, specific subnetworks within the network can have other forms. The World Wide Web, for instance, shows a power-law degree distribution overall but unimodal distributions

[13] The term "scale-free" refers to any functional form $f(x)$ that remains unchanged to within a multiplicative factor under a rescaling of the independent variable x. In effect this means power-law forms, since these are the only solutions to $f(ax) = bf(x)$, and hence "power-law" and "scale-free" are, for our purposes, synonymous.

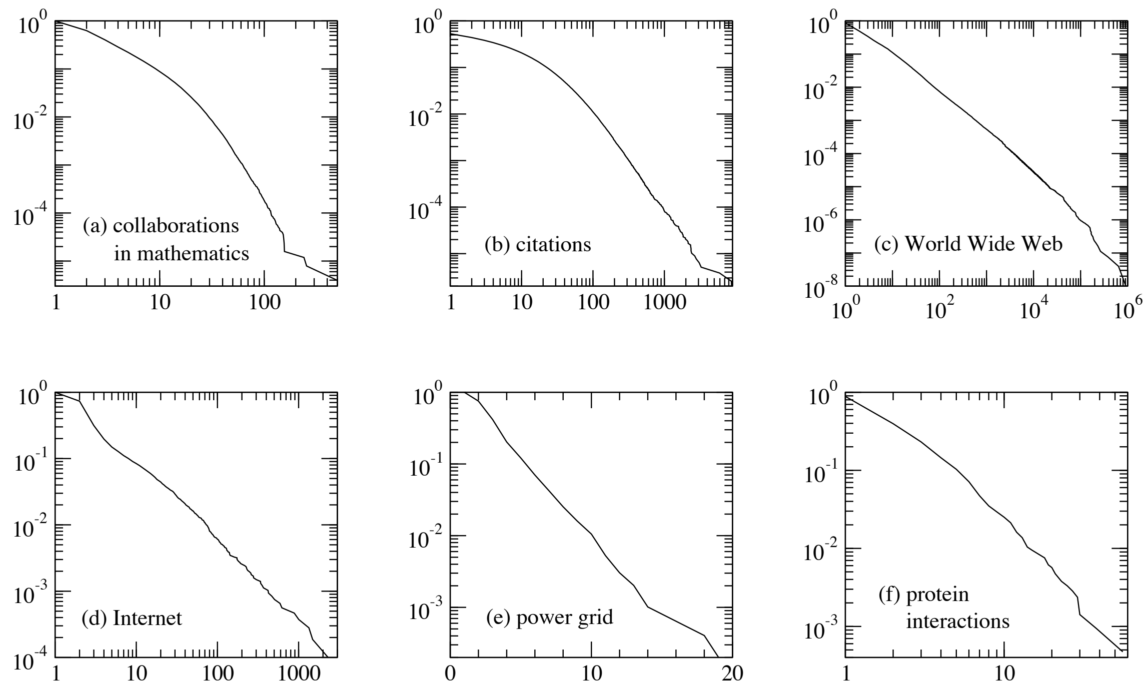

FIG. 6 Cumulative degree distributions for six different networks. The horizontal axis for each panel is vertex degree k (or in-degree for the citation and Web networks, which are directed) and the vertical axis is the cumulative probability distribution of degrees, i.e., the fraction of vertices that have degree greater than or equal to k. The networks shown are: (a) the collaboration network of mathematicians [182]; (b) citations between 1981 and 1997 to all papers cataloged by the Institute for Scientific Information [351]; (c) a 300 million vertex subset of the World Wide Web, *circa* 1999 [74]; (d) the Internet at the level of autonomous systems, April 1999 [86]; (e) the power grid of the western United States [416]; (f) the interaction network of proteins in the metabolism of the yeast *S. Cerevisiae* [212]. Of these networks, three of them, (c), (d) and (f), appear to have power-law degree distributions, as indicated by their approximately straight-line forms on the doubly logarithmic scales, and one (b) has a power-law tail but deviates markedly from power-law behavior for small degree. Network (e) has an exponential degree distribution (note the log-linear scales used in this panel) and network (a) appears to have a truncated power-law degree distribution of some type, or possibly two separate power-law regimes with different exponents.

within domains [338].

2. Maximum degree

The maximum degree $k_{\max}$ of a vertex in a network will in general depend on the size of the network. For some calculations on networks the value of this maximum degree matters (see, for example, Sec. VIII.C.2). In work on scale-free networks, Aiello *et al.* [8] assumed that the maximum degree was approximately the value above which there is less than one vertex of that degree in the graph on average, i.e., the point where $np_k = 1$. This means, for instance, that $k_{\max} \sim n^{1/\alpha}$ for the power-law degree distribution $p_k \sim k^{-\alpha}$. This assumption however can give misleading results; in many cases there will be vertices in the network with significantly higher degree than this, as discussed by Adamic *et al.* [6].

Given a particular degree distribution (and assuming all degrees to be sampled independently from it, which may not be true for networks in the real world), the probability of there being exactly m vertices of degree k and no vertices of higher degree is $\binom{n}{m} p_k^m (1-P_k)^{n-m}$, where P_k is the cumulative probability distribution, Eq. (7). Hence the probability h_k that the highest degree on the graph is k is

$$\begin{aligned} h_k &= \sum_{m=1}^{n} \binom{n}{m} p_k^m (1-P_k)^{n-m} \\ &= (p_k + 1 - P_k)^n - (1-P_k)^n, \end{aligned} \tag{10}$$

and the expected value of the highest degree is $k_{\max} = \sum_k k h_k$.

For both small and large values of k, h_k tends to zero, and the sum over k is dominated by the terms close to the maximum. Thus, in most cases, a good approximation to the expected value of the maximum degree is given by the modal value. Differentiating and observing that $\mathrm{d}P_k/\mathrm{d}k = p_k$, we find that the maximum of h_k occurs when

$$\left(\frac{\mathrm{d}p_k}{\mathrm{d}k} - p_k\right)(p_k + 1 - P_k)^{n-1} + p_k(1-P_k)^{n-1} = 0, \tag{11}$$

or $k_{\max}$ is a solution of

$$\frac{\mathrm{d}p_k}{\mathrm{d}k} \simeq -np_k^2, \tag{12}$$