The mathematics of networks

M. E. J. Newman

Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109–1040

In much of economic theory it is assumed that economic agents interact, directly or indirectly, with all others, or at least that they have the opportunity to do so in order to achieve a desired outcome for themselves. In reality, as common sense tells us, things are quite different. Traders in a market have preferred trading partners, perhaps because of an established history of trust, or simply for convenience. Buyers and sellers have preferred suppliers and customers. Consumers have preferred brands and outlets. And most individuals limit their interactions, economic or otherwise, to a select circle of partners or acquaintances. In many cases partners are chosen not on economic grounds but for social reasons: individuals tend overwhelmingly to deal with others who revolve in the same circles as they do, socially, intellectually or culturally.

The patterns of connections between agents form a social network (Fig. 1) and it is intuitively clear that the structure of such networks must affect the pattern of economic transactions, not to mention essentially every other type of social interaction amongst human beings. Any theory of interaction that ignores these networks is necessarily incomplete, and may in fact be missing some important and crucial phenomena. In the last few decades, therefore, a researchers have conducted extensive investigations of networks in economics, mathematics, sociology and a number of other fields, in an effort to understand and explain network effects.

The study of social (and other) networks has three primary components. First, empirical studies of networks probe network structure using a variety of techniques such as interviews, question-

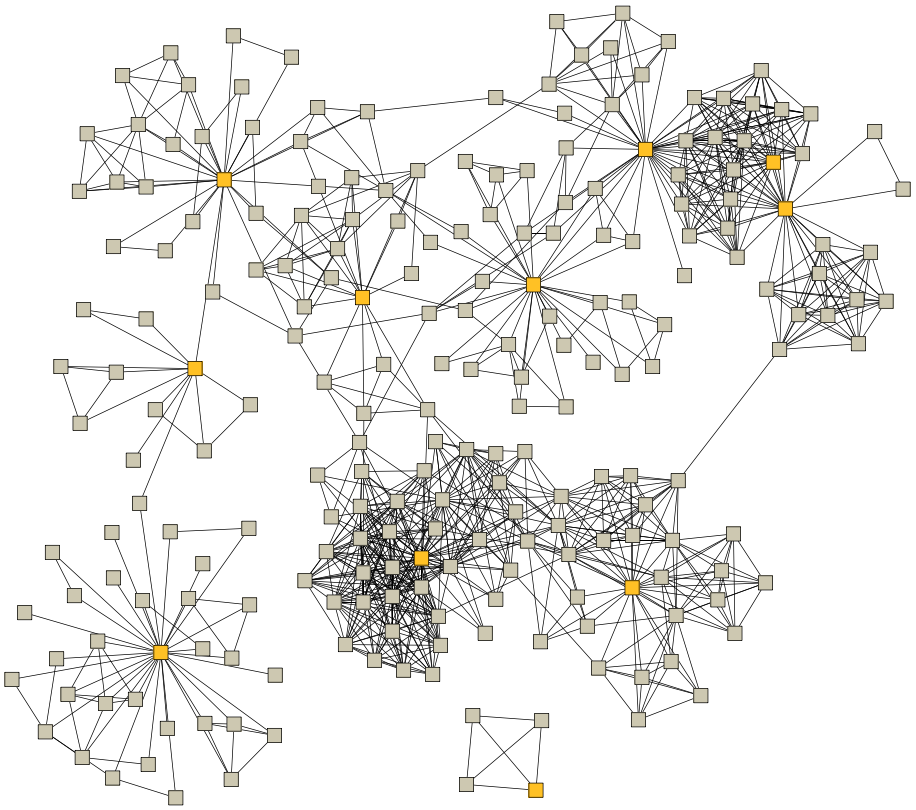

Figure 1: An example of a social network, in this case of collaborative links. The nodes (squares) represent people and the edges (lines) social ties between them.

naires, direct observation of individuals, use of archival records, and specialist tools like “snowball sampling” and “ego-centred” studies. The goal of such studies is to create a picture of the connections between individuals, of the type shown in Fig. 1. Since there are many different kinds of possible connections between people—business relationships, personal relationships, and so forth—studies must be designed appropriately to measure the particular connections of interest to the experimenter.

Second, once one has empirical data on a network, one can answer questions about the community the network represents using mathematical or statistical analyses. This is the domain of classical social network analysis, which focuses on issues such as: Who are the most central mem-

bers of a network and who are the most peripheral? Which people have most influence over others? Does the community break down into smaller groups and if so what are they? Which connections are most crucial to the functioning of a group?

And third, building on the insights obtained from observational data and its quantitative analysis, one can create models, such as mathematical models or computer models, of processes taking place in networked systems—the interactions of traders, for example, or the diffusion of information or innovations through a community. Modelling work of this type allows us to make predictions about the behaviour of a community as a function of the parameters affecting the system.

This article reviews the mathematical techniques involved in the second and third of these three components: the quantitative analysis of network data and the mathematical modelling of networked systems. Necessarily this review is short. Much more substantial coverage can be found in the many books and review articles in the field [1–8].

Let us begin with some simple definitions. A network—also called a *graph* in the mathematics literature—is made up of points, usually called *nodes* or *vertices*, and lines connecting them, usually called *edges*. Mathematically, a network can be represented by a matrix called the *adjacency matrix* $\mathbf{A}$, which in the simplest case is an $n \times n$ symmetric matrix, where n is the number of vertices in the network. The adjacency matrix has elements

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The matrix is symmetric since if there is an edge between i and j then clearly there is also an edge between j and i. Thus $A_{ij} = A_{ji}$.

In some networks the edges are *weighted*, meaning that some edges represent stronger connections than others, in which case the nonzero elements of the adjacency matrix can be generalized to values other than unity to represent stronger and weaker connections. Another variant is the *directed network*, in which edges point in a particular direction between two vertices. For instance,

in a network of cash sales between buyers and sellers the directions of edges might represent the direction of the flow of goods (or conversely of money) between individuals. Directed networks can be represented by an asymmetric adjacency matrix in which $A_{ij} = 1$ implies the existence (conventionally) of an edge pointing from j to i (note the direction), which will in general be independent of the existence of an edge from i to j.

Networks may also have *multiedges* (repeated edges between the same pair of vertices), *self-edges* (edges connecting a vertex to itself), *hyperedges* (edges that connect more than two vertices together) and many other features. We here concentrate however primarily on the simplest networks having undirected, unweighted single edges between pairs of vertices.

Turning to the analysis of network data, we start by looking at *centrality measures*, which are some of the most fundamental and frequently used measures of network structure. Centrality measures address the question, "Who is the most important or central person in this network?" There are many answers to this question, depending on what we mean by important. Perhaps the simplest of centrality measures is *degree centrality*, also called simply *degree*. The degree of a vertex in a network is the number of edges attached to it. In mathematical terms, the degree k_i of a vertex i is

$$k_i = \sum_{j=1}^{n} A_{ij}. \tag{2}$$

Though simple, degree is often a highly effective measure of the influence or importance of a node: in many social settings people with more connections tend to have more power.

A more sophisticated version of the same idea is the so-called *eigenvector centrality*. Where degree centrality gives a simple count of the number of connections a vertex has, eigenvector centrality acknowledges that not all connections are equal. In general, connections to people who are themselves influential will lend a person more influence than connections to less influential people. If we denote the centrality of vertex i by x_i, then we can allow for this effect by making x_i

proportional to the average of the centralities of i's network neighbours:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j, \tag{3}$$

where λ is a constant. Defining the vector of centralities $\mathbf{x} = (x_1, x_2, \ldots)$, we can rewrite this equation in matrix form as

$$\lambda \mathbf{x} = \mathbf{A} \cdot \mathbf{x}, \tag{4}$$

and hence we see that $\mathbf{x}$ is an eigenvector of the adjacency matrix with eigenvalue λ. Assuming that we wish the centralities to be non-negative, it can be shown (using the Perron–Frobenius theorem) that λ must be the largest eigenvalue of the adjacency matrix and $\mathbf{x}$ the corresponding eigenvector.

The eigenvector centrality defined in this way accords each vertex a centrality that depends both on the number and the quality of its connections: having a large number of connections still counts for something, but a vertex with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts. Eigenvector centrality turns out to be a revealing measure in many situations. For example, a variant of eigenvector centrality is employed by the well-known Web search engine Google to rank Web pages, and works well in that context.

Two other useful centrality measures are *closeness centrality* and *betweenness centrality*. Both are based upon on the concept of network paths. A path in a network is a sequence of vertices traversed by following edges from one to another across the network. A *geodesic path* is the shortest path, in terms of number of edges traversed, between a specified pair of vertices. (Geodesic paths need not be unique; there is no reason why there should not be two paths that tie for the title of shortest.) The closeness centrality of vertex i is the mean geodesic distance (i.e., the mean length of a geodesic path) from vertex i to every other vertex. Closeness centrality is *lower* for vertices that are more central in the sense of having a shorter network distance on average to other vertices. (Some writers define closeness centrality to be the reciprocal of the average so that higher numbers indicate greater centrality. Also, some vertices may not be reachable from vertex i—two vertices

can lie in separate “components” of a network, with no connection between the components at all. In this case closeness as above is not well defined. The usual solution to this problem is simply to define closeness to be the average geodesic distance to all *reachable* vertices, excluding those to which no path exists.)

The betweenness centrality of vertex *i* is the fraction of geodesic paths between other vertices that *i* falls on. That is, we find the shortest path (or paths) between every pair of vertices, and ask on what fraction of those paths vertex *i* lies. Betweenness is a crude measure of the control *i* exerts over the flow of information (or any other commodity) between others. If we imagine information flowing between individuals in the network and always taking the shortest possible path, then betweenness centrality measures the fraction of that information that will flow through *i* on its way to wherever it is going. In many social contexts a vertex with high betweenness will exert substantial influence by virtue not of being in the middle of the network (although it may be) but of lying “between” other vertices in this way. It is in most cases only an approximation to assume that information flows along geodesic paths; normally it will not, and variations of betweenness centrality such as “flow betweenness” and “random walk betweenness” have been proposed to allow for this. In many practical cases however, the simple (geodesic path) betweenness centrality gives quite informative answers.

The study of shortest paths on networks also leads to another interesting network concept, the *small-world effect*. It is found that in most networks the mean geodesic distance between vertex pairs is small compared to the size of the network as a whole. In a famous experiment conducted in the 1960s, the psychologist Stanley Milgram asked participants to get a message to a specified target person elsewhere in the country by passing it from one acquaintance to another, stepwise through the population. Milgram’s remarkable finding that the typical message passed though just six people on its journey between (roughly) randomly chosen initial and final individuals has been immortalized in popular culture in the phrase “six degrees of separation”, which was the title of

a 1990 Broadway play by John Guare in which one of the characters discusses the small-world effect. Since Milgram's experiment, the small-world effect has been confirmed experimentally in many other networks, both social and nonsocial.

Other network properties that have attracted the attention of researchers in recent years include network *transitivity* or *clustering* (the tendency for triangles of connections to appear frequently in networks—in common parlance, "the friend of my friend is also my friend"), vertex similarity (the extent to which two given vertices do or do not occupy similar positions in the network), communities or groups within networks and methods for their detection, and crucially, the distribution of vertex degrees, a topic discussed in more detail below.

Turning to models of networks and of the behaviour of networked systems, perhaps the simplest useful model of a network (and one of the oldest) is the *Bernoulli random graph*, often called just the *random graph* for short [9–11]. In this model one takes a certain number of vertices n and creates edges between them with independent probability p for each vertex pair. When p is small there are only a few edges in the network, and most vertices exist in isolation or in small groups of connected vertices. Conversely, for large p almost every possible edge is present between the $\binom{n}{2}$ possible vertex pairs, and all or almost all of the vertices join together in a single large connected group. One might imagine that for intermediate values of p the sizes of groups would just grow smoothly from small to large, but this is not the case. It is found instead that there is a *phase transition* at the special value $p = 1/n$ above which a *giant component* forms, a group of connected vertices occupying a fixed fraction of the whole network, i.e., with size varying as n. For values of p less than this, only small groups of vertices exist of a typical size that is independent of n. Many real-world networks show behaviour reminiscent of this model, with a large component of connected vertices filling a sizable fraction of the entire network, the remaining vertices falling in much smaller components that are unconnected to the rest of the network.

The random graph has a major shortcoming however: the distribution of the degrees of the

vertices is quite unlike that seen in most real-world networks. The fraction p_k of vertices in a random graph having degree k is given by the binomial distribution, which becomes Poisson in the limit of large n:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \simeq \frac{z^k e^{-z}}{k!}, \tag{5}$$

where $z = (n-1)p$ is the mean degree. Empirical observations of real networks, social and otherwise, show that most have highly non-Poisson distributions of degree, often heavily right-skewed with a fat tail of vertices having unusually high degree [6, 7]. These high-degree nodes or "hubs" in the tail can, it turns out, have a substantial effect on the behaviour of a networked system.

To allow for these non-Poisson degree distributions, one can generalize the random graph, specifying a particular, arbitrary degree distribution p_k and then forming a graph that has that distribution but is otherwise random. A simple algorithm for doing this is to choose the degrees of the n vertices from the specified distribution, draw each vertex with the appropriate number of "stubs" of edges emerging from it, and then pick stubs in pairs uniformly at random and connect them to create complete edges. The resulting model network (or more properly the ensemble of such networks) is called the *configuration model*.

The configuration model also shows a phase transition, similar to that of the Bernoulli random graph, at which a giant component forms. To see this, consider a set of connected vertices and consider the "boundary vertices" that are immediate neighbours of that set. Let us grow our set by adding the boundary vertices to it one by one. When we add one boundary vertex to our set the number of boundary vertices goes down by 1. However, the number of boundary vertices also increases by the number of new neighbours of the vertex added, which is one less than the degree k of that vertex. Thus the total change in the number of boundary vertices is $-1+(k-1) = k-2$. However, the probability of a particular vertex being a boundary vertex is proportional to k, since there are k times as many edges by which a vertex of degree k could be connected to our set than a vertex of degree 1. Thus the average change in the number of boundary vertices when we add one

vertex to our set is a weighted average $\sum_i k_i(k_i-2)/\sum_j k_j = \sum_i k_i(k_i-2)/(nz)$, where z is again the mean degree. If this quantity is less than zero, then the number of boundary vertices dwindles as our set grows bigger and will in the end reach zero, so that the set will stop growing. Thus in this regime all connected sets of vertices are of finite size. If on the other hand this number is greater than zero then the number of boundary vertices will grow without limit, and hence the size of our set of connected vertices is limited only by the size of the network.

Thus a giant component exists in the network if and only if

$$\langle k^2\rangle - 2\langle k\rangle > 0, \tag{6}$$

where $\langle k\rangle = z = n^{-1}\sum_i k_i$ is the mean degree and $\langle k^2\rangle = n^{-1}\sum_i k_i^2$ is the mean-square degree.

The occurrence here of the mean-square degree is a phenomenon that appears over and over in the mathematics of networks. Another context in which it appears is in the spread of information (or anything else) over a network. Taking a simple model of the spread of an idea (or a rumour or a disease), imagine that each person who has heard the idea communicates it with independent probability r to each of his or her friends. If the person's degree is k then there are $k-1$ friends to communicate the idea to, not counting the one from whom they heard it in the first place, so the expected number who hear it is $r(k-1)$. Performing the weighted average over vertices again, the average number of people a person passes the idea on to, also called the *basic reproductive number* R_0, is

$$R_0 = r\,\frac{\sum_i k_i(k_i-1)}{\sum_i k_i} = r\,\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}. \tag{7}$$

If R_0 is greater than 1, then the number of people hearing the idea grows as it gets passed around and it will take off exponentially. If R_0 is less than 1 then the idea will die. Again, we have a phase transition, or *tipping point*, for the spread of the idea: it spreads if and only if

$$r > \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle}. \tag{8}$$

The simple understanding behind the appearance of the mean-square degree in this expression is the following. If a person with high degree hears this idea they can spread it to many others, because they have so many friends. However, such a person is also *more likely* to hear the idea in the first place, because they have so many friends to hear it from. Thus, the degree enters twice into the process: a person with degree 10 is $10 \times 10 = 100$ times more efficacious at spreading the idea than a person with degree 1.

The appearance of the mean-square degree in expressions like (6) and (8) can have substantial effects. Of particular interest are networks whose degree distributions have fat tails. It is possible for such networks to have very large values of $\langle k^2 \rangle$—in the hundreds or thousands—so that, for example, the right-hand side of Eq. (8) is very small. This means that the probability of each individual person spreading an idea (or rumour or disease) need not be large for it still to spread through the whole community.

Another important class of network models is the class of generative models, models that posit a quantitative mechanism or mechanisms by which a network forms, usually in an effort to explain how the observed structure of the network arises. The best known example of such a model is the "cumulative advantage" or "preferential attachment" model [12, 13], which aims to explain the fat-tailed degree distributions mentioned above. In its simplest form this model envisages a network that grows by the steady addition of vertices, one at a time. Many networks, such as the World Wide Web and citation networks grow this way; it is a matter of current debate whether the model applies to social networks as well. Each vertex is added with a certain number m of edges emerging from it, whose other ends connect to preexisting vertices with probability proportional to those vertices' current degree. That is, the higher the current degree of a vertex, the more likely that vertex is to acquire new edges when the graph grows. This kind of rich-get-richer phenomenon is plausible in many network contexts and is known to generate Pareto degree distributions. Using a rate-equation method [12, 14, 15] it is straightforward to show that in the limit of large network

size the degree distribution obeys:

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}. \tag{9}$$

This distribution has a tail going as $p_k \sim k^{-3}$ in the large-k limit, which is strongly reminiscent of the degree distributions seen particularly in citation networks and also in the World Wide Web. Generative models of this type have been a source of considerable interest in recent years and have been much extended beyond the simple ideas described here by a number of authors [6, 7].

Concepts such as those appearing in this article can be developed a great deal further and lead to a variety of useful, and in some cases surprising, results about the function of networked systems. More details can be found in the references.

References

[1] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge University Press, Cambridge (1994).

[2] J. Scott, *Social Network Analysis: A Handbook*. Sage, London, 2nd edition (2000).

[3] D. B. West, *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ (1996).

[4] F. Harary, *Graph Theory*. Perseus, Cambridge, MA (1995).

[5] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Upper Saddle River, NJ (1993).

[6] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford (2003).

[7] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

[8] M. E. J. Newman, The structure and function of complex networks. *SIAM Review* **45**, 167–256 (2003).

[9] R. Solomonoff and A. Rapoport, Connectivity of random nets. *Bulletin of Mathematical Biophysics* **13**, 107–117 (1951).

[10] P. Erdős and A. Rényi, On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–61 (1960).

[11] B. Bollobás, *Random Graphs*. Academic Press, New York, 2nd edition (2001).

[12] D. J. de S. Price, A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Inform. Sci.* **27**, 292–306 (1976).

[13] A.-L. Barabási and R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

[14] H. A. Simon, On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955).

[15] P. L. Krapivsky, S. Redner, and F. Leyvraz, Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4629–4632 (2000).