Networks

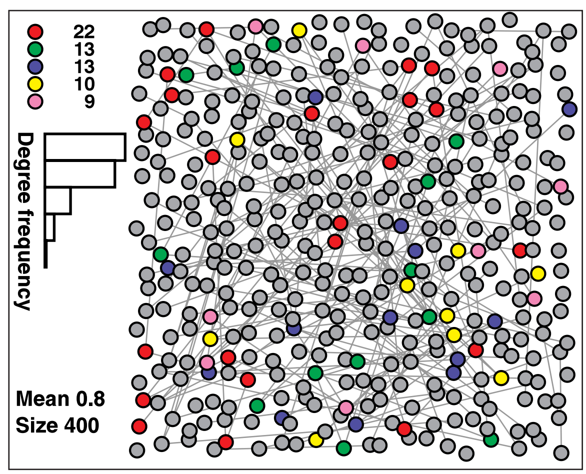
Bill Casselman

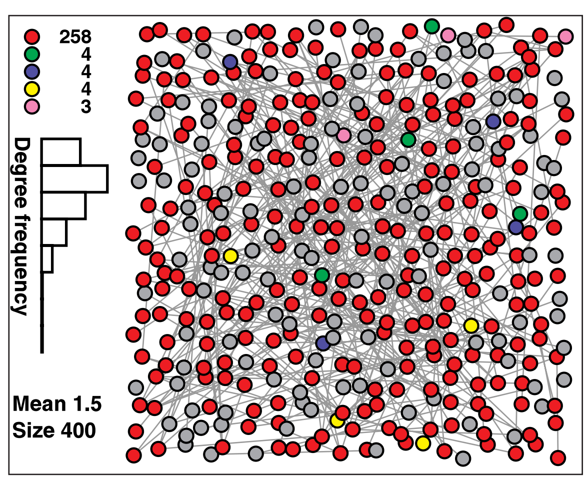
Every year a consortium of mathematics organizations chooses a theme for the month of April, designated "Math Awareness Month" (see also `http://www.ams.org/ams/mam.html`). This year's theme is **networks**. The importance of networks in mathematics at the moment can be gauged by the fact that the *Notices* has recently published two reviews of books on the topic and will publish an article by Peter Sarnak on a related topic in the near future.

"Network" is the word everybody—physicists, biologists, sociologists, engineers—except mathematicians and a few computer scientists, uses for what this minority calls "graphs", i.e., a set of nodes linked together by edges. Variants exist: the edges might be oriented or labeled, or the vertices classified. If there is a connotation to "network" as opposed to "graph", it is that networks arise from real life and are concerned with relations between real objects. Important examples include metabolic interactions of chemicals in living things, hard-wired connections among servers on the Internet, links between World Wide Web pages, citations of references in scientific papers, contagion, and electric power grids. Networks—very, very large networks—are a routine and important part of modern life. The World Wide Web has much over 1 billion live nodes. How can we make sense of something that complex?

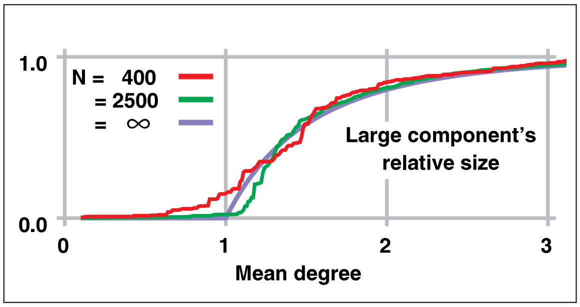
The traditional way to analyze a graph is just to draw it and look at it, but for large networks this is ridiculous. One has to condense the information available into a small package—in other words, to compile statistics.

What does a typical graph look like? In the beautiful theory originating mostly with Alfred Renyi and Pál Erdős, a **random graph** is obtained by starting with N nodes fixed in advance and then adding edges between random pairs with probability p. There are an astonishing number of interesting results known about such graphs, the most remarkable being that certain phase transitions occur as p increases. The **degree** of a node in a network is the number of edges leading to it. In the random graph, distribution of degrees is binomial, well approximated by a Poisson distribution for small p. The mean degree of nodes is $\mu = p(N-1)$. The best-known phase transition is that as μ passes through 1, the largest connected component starts to grow rapidly. For $\mu < 1$ and large N it contains a very small fraction of all nodes, but as μ increases it takes up a sizeable part; for $\mu > 1$ the fraction f taken up is the unique root of $f = 1 - e^{-\mu f}$. The sizes of the other components decrease dramatically as well. In the following two figures a sequence of randomly constructed graphs is shown for $p = 0.8$ and 1.5. The colored nodes are those in the five largest components. Also shown is the empirical distribution of nodal degrees.



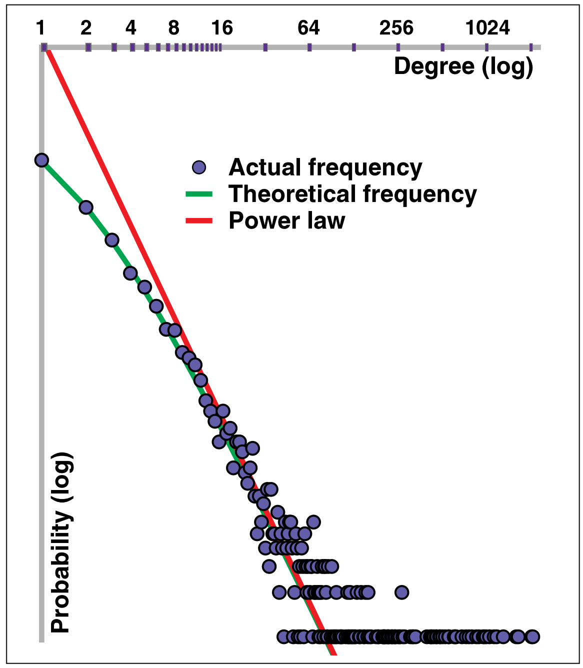


Bill Casselman is a professor of mathematics at the University of British Columbia and the graphics editor of the Notices. *His email address is* `cass@math.ubc.ca`.



The figure above shows how the largest component grows as μ increases through a sequence of values for graphs of sizes 400 and 2500 constructed randomly by a computer. The graph for $N = \infty$ shows the theoretical fraction taken up by the largest component for $p > 1$, a graph which the true fraction approximates more closely for large N.

The traditional theory of random graphs is mathematically impressive and plays an important role in estimates of the efficiency of many algorithms dealing with graphs. But its role in explaining what networks encountered in real life look like is minimal. Real networks do not spring out of nowhere; they grow in one way or another, and the structure they acquire depends on that growth. In particular, this causes them generally to look very different from one of the Renyi-Erdős random graphs. As far as I know, the first account of how real networks seem to grow dates to 1965, in a paper by the remarkable polymath Derek de Solla Price. He was interested in the statistics of the science citation network, where directed edges lead from a paper to each paper it refers to. It had recently become possible to scan efficiently through large amounts of data based on the **Science Citation Index**, and de Solla Price showed that a principle of formation he called **cumulative advantage** led to a stable degree probability distribution $p_k = (1 + 1/m)B(k + 1, 2 + 1/m)$. Here p_k is the probability of a node having degree k, m is the mean effective number of journal references in papers, and B is Legendre's beta function $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$. He also showed that there was some agreement between his theory and the nature of citations in the real world. Cumulative advantage can be succinctly summarized as "them what has, gets"—a paper that is frequently referred to will likely get more citations than one referred to less often. (A warning: de Solla Price's paper will be sobering to a naive mathematician who would like to think that a paper's intrinsic quality plays a more important role than mere popularity.) What is perhaps surprising is that the stable degree distribution is well approximated even for small networks growing in this way, at least away from the tail end of highly cited papers. Because $B(x, y) \sim \Gamma(y)/x^y$ for fixed y and increasing x, de Solla Price's distribution is reasonably approximated by a **power law** where the number of nodes of degree k is proportional to some power $1/k^\alpha$ with $\alpha > 2$, here $2 + 1/m$. The following figure shows a degree distribution arising from a sample network grown by cumulative advantage, along with the theoretical beta distribution, all plotted on a log-log graph so as to show up the power law indicated by a straight line.



De Solla Price's work seems to have been long neglected, and his results were rediscovered, apparently independently, by Albert Barabási in much cited recent work. Barabási introduced the now popular term "preferential attachment" for what de Solla Price calls cumulative advantage and the term "scale free" networks for those that satisfy a power law of degree distribution, although it seems more reasonable to apply the term only to those which, like the networks of de Solla Price, retain a stable degree distribution as they grow.

Further Reading

There is a great deal of pseudoscience as well as science in this field, which is enjoying a rapid growth at the moment. The most difficult problem is to fit theory with practice. A very good reference for mathematicians is the recent article "The structure and function of complex networks" by Mark Newman in volume 45 of the *SIAM Review* (2003), recommended also by Rick Durrett in his March *Notices* review of Barabási's book *Linked*. It contains a large and useful list of references on the topic. You can find listed there the papers of de Solla Price as well as an interesting introduction to his research on citation networks.