Networks: Spring 2007	**Graph Theory**
David Easley and Jon Kleinberg	**January 24, 2007**

A graph is simply a way of encoding the pairwise relationships among a set of objects: we will refer to the objects as *nodes*, with *edges* connecting certain pairs of them.

Edges in a graph indicate a symmetric relationship between their endpoints. Often we want to encode asymmetric relationships, and for this we use the closely related notion of a *directed graph*. A directed graph has nodes as before, but now each edge has a direction: it goes from a node u (its *tail*) to a node v (its *head*). When we want to emphasize that the graph we are considering is not directed, we will call it an *undirected graph*; by default, however, the term "graph" will mean an undirected graph.

Examples of Graphs

Graphs are very simple to define: we just take a collection of things, and join some of them by edges. But at this level of abstraction, it's hard to appreciate the typical kinds of situations in which they arise. In the first lecture, we saw a number of examples of graphs; here we summarize again some basic contexts in which graphs arise. In going through the list, it's useful to digest the meaning of the nodes and the meaning of the edges in the context of the application; in some cases, the nodes and edges both correspond to physical objects in the real world, in others the nodes are real objects while the edges are virtual, and in still others both nodes and edges are pure abstractions.

1. *Transportation networks.* The map of routes served by an airline carrier naturally forms a graph: the nodes are airports, and there is an edge from u to v if there is a non-stop flight that departs from u and arrives at v. Described this way, the graph is directed; but in practice when there is an edge from u to v, there is also almost always one from v to u. So we do not lose much by treating the airline route map as an undirected graph with edges joining pairs of airports that have non-stop flights each way. Looking at such a graph (you can generally find them depicted in the backs of in-flight airline magazines), we'd quickly notice a few things: there are a small number of hubs with a very large number of incident edges; and it's possible to get between any two nodes in the graph via a very small number of intermediate stops.

   Other transportation networks can be modeled in a similar way. For example, we could take a rail network and have a node for each terminal, and an edge joining u and v if there's a section of railway track that goes between them without stopping at any intermediate terminal. The standard depiction of the subway map in a major city is a drawing of such a graph.

Portions of the text here are reproduced from Chapter 3 of the book *Algorithm Design* (Jon Kleinberg and Éva Tardos, Addison-Wesley, 2006) [4].

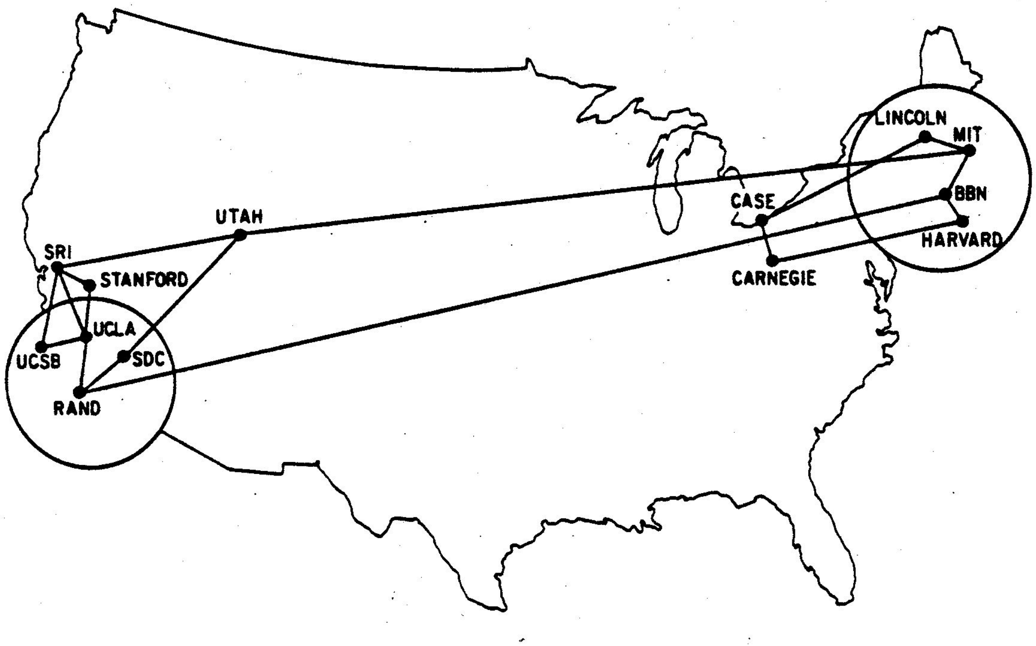

Figure 1: A network depicting the sites on the Internet, then known as the Arpanet, in December 1970. (Image from F. Heart, A. McKenzie, J. McQuillian, and D. Walden [3]; on-line at http://som.csudh.edu/cis/lpress/history/arpamaps/.)

2. *Communication networks.* A collection of computers that are connected via a communication network can be naturally modeled as a graph in a few different ways. First, we could have a node for each computer and an edge joining u and v if there is a direct physical link connecting them. Reaching back into ancient history a bit, one can find such such images of the Internet from a time when it had only a few nodes (see Figure 1).

   Alternately, for studying the large-scale structure of the Internet, people often define a node to be the set of all machines controlled by a single Internet Service Provider, with an edge joining u and v if there is a direct *peering relationship* between them — roughly, an agreement to exchange data under the standard protocol that governs global Internet routing. Note that this latter network is a more "virtual" one than the former, since the links indicate a formal agreement in addition to a physical connection.

   In studying wireless networks, one typically defines a graph where the nodes are computing devices situated at locations in physical space, and there is an edge from u to v if v is close enough to u to receive a signal from it. Note that it's often useful to view such a graph as directed, since it may be the case that v can hear u's signal but u cannot v's signal (if, for example, u has a stronger transmitter). These graphs are also interesting from a geometric perspective, since they roughly correspond to putting down points in the plane and then joining pairs that are close together.

3. *Information networks.* The World Wide Web can be naturally viewed as a directed graph, in which nodes correspond to Web pages and there is an edge from u to v if u has a hyperlink to v. The directedness of the graph is crucial here; many pages, for example, link to popular news sites, but these sites clearly do not reciprocate all these links. The structure of all these hyperlinks plays an important role in current approaches to Web search, as we will see later in the course.

   The hypertextual structure of the Web is anticipated by a number of information networks that predate the Internet by many decades; these include the network of cross-references among articles in an encyclopedia or other reference work, and the network of bibliographic citations among scientific papers.

4. *Social networks.* Given any collection of people who interact (the employees of a company, the students in a high school, or the residents of a small town), we can define a network whose nodes are people, with an edge joining u and v if they are friends with one another. We could have the edges mean a number of different things instead of friendship: the undirected edge linking u and v could mean that u and v have had a romantic relationship or a financial relationship; a directed edge from u to v could mean that u seeks advice from v, or that u lists v in his or her e-mail address book. As a supplement to the social network images from the first lecture, the image in Figure 2 shows the romantic relationships among students in a large American high school over a period of 18 months.

   At different points in the course, we will see how such networks are used by sociologists to study the dynamics of interaction among people; they can be used to identify the most "influential" people in a company or organization, to model trust relationships in a financial or political setting, and to track the spread of fads, rumors, jokes, diseases, and e-mail viruses.

5. *Dependency networks.* It is natural to define directed graphs that capture the interdependencies among a collection of objects. For example, given the list of courses offered by a college or university, we could have a node for each course and an edge from u to v if u is a pre-requisite for v. Given a list of modules in a large software system, we could have a node for each module and an edge from u to v if v requires some data being produced by u. Or given a set of species in an ecosystem, we could define a graph — a *food web* — in which the nodes are the different species and there is an edge from u to v if u consumes v.

This is far from a complete list, too far to even begin tabulating its omissions; it is meant simply to suggest some examples that are useful to have in mind as we talk in general about graphs.

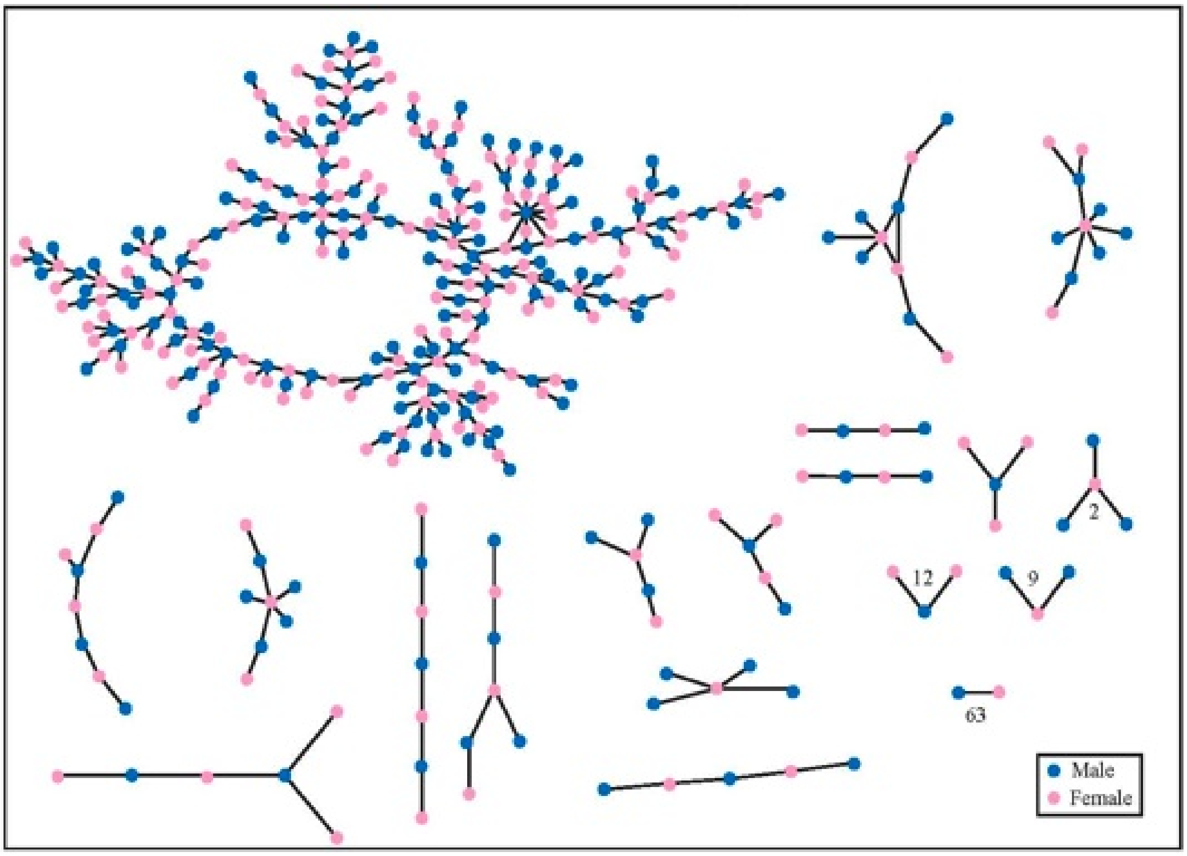

Figure 2: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted. (Image from P. Bearman, J. Moody, and K. Stovel [1].)

Paths, Cycles, Distances, and Connectivity

One of the fundamental operations in a graph is that of traversing a sequence of nodes connected by edges. In our examples above, such a traversal could correspond to a user browsing Web pages by following hyperlinks; a rumor passing by word of mouth from you to someone halfway around the world; or an airline passenger traveling from Ithaca to Palermo on a sequence of flights.

With this notion in mind, we define a *path* in an undirected graph to be a sequence of nodes $v_1, v_2, \ldots, v_{k-1}, v_k$ with the property that each consecutive pair v_i, v_{i+1} is joined by an edge in G. P is often called a path *from* v_1 *to* v_k, or a v_1-v_k path. For example, the nodes CASE, LINCOLN, MIT, UTAH form a path in Figure 1.

We call such a sequence of nodes a *cycle* if additionally we have $v_1 = v_k$ — in other words, the sequence "cycles back" to where it began. So for example, SRI, STANFORD, UCLA, UCSB, SRI is a cycle in Figure 1. In fact, every edge in the 1970 Arpanet belongs to a cycle, and this was by design: it means that if any edge were to fail (e.g. a construction crew accidentally cut through the cable), there would still be a way to get from any node to any other node.

There is an enormous cycle in the upper left of Figure 2: if you were on this cycle, then it would be possible to say that you dated someone who dated someone who ... (continue 34 more times) ... dated someone who dated you. As Bearman, Moody, and Stovel note in the paper where they analyze this network, "These structures reflect relationships that may be long over, and they link individuals together in chains far too long to be the subject of even the most intense gossip and scrutiny. Nevertheless, they are real: like social facts, they are invisible yet consequential macrostructures that arise as the product of individual agency" [1].

The definitions of paths and cycles carry over naturally to directed graphs, with the following change: In a directed path, each pair of consecutive nodes has the property that (v_i, v_{i+1}) is an edge. In other words, the sequence of nodes in the path or cycle must respect the directionality of edges.

In addition to simply knowing about the existence of a path between some pair of nodes u and v, we may also want to know whether there is a *short* path. Thus we define the *distance* between two nodes u and v to be the minimum number of edges in a u-v path. (We can designate some symbol like ∞ to denote the distance between nodes that are not connected by a path.) The term "distance" here comes from imagining G as representing a communication or transportation network; if we want to get from u to v, we may well want a route with as few "hops" as possible.

We say that an undirected graph is *connected* if for every pair of nodes u and v, there is a path from u to v. Choosing how to define connectivity of a directed graph is a bit more subtle, since it's possible for u to have a path to v while v has no path to u. We say that a directed graph is *strongly connected* if for every two nodes u and v, there is a path from u to v and a path from v to u.

If an undirected graph is not connected, then we can talk about how it breaks up into connected pieces. For a node u, we say that the *connected component* containing u is the set of all nodes v to which u has a path. (The term *connected component* is often shortened just to *component*.) Notice that for two nodes, their connected components are either identical, or they have no nodes in common. In Figure 2, notice how there is a single very large component, and all other components are much smaller. We will find that many of the networks we study in fact have such a "giant" component that dwarfs all the others.

Summarizing a Network

As we noted in the first lecture, once a network has more than a few hundred nodes, it's hard to learn much just by looking at a picture of it. An alternative is to report some numerical properties of the network, with the goal of using these to "summarize" aspects of its structure.

Finding the most informative properties to report is in fact a question that's not entirely well-understood, and it's the topic of current research. However, a gradual consensus is emerging, and particular properties often recur in network studies. With this in mind, let's consider some of the properties used in the Kossinets-Watts study of a large e-mail network

[5], which we'll be discussing in lecture. The discussion here should also help in reading the Kossinets-Watts paper, by expanding on some of the technical notation used there.

1. *Average degree.* The *degree* of a node is the number of edges for which it is an endpoint (i.e. the number of edges attached to it in a drawing of the graph). The *average degree* of a node (denoted $\langle k \rangle$ in [5]) thus gives the "density" of the graph — the average number of edges attached to each node.

   The *degree distribution* is also of interest (see Figure 4A and 4B in [5]) — this is a plot that shows the fraction of nodes in the graph of degree k, for each value of k. The degree distribution thus tells how many nodes act as very high-degree "hubs," and how many are only weakly connected to the rest of the graph.

2. *Fraction of nodes in the largest component.* Above, we mentioned that many networks contain a single "giant component" containing a large fraction of the nodes. On the other hand, one almost never finds real networks that have two very large components — all it would take is a single edge with one end in each component, and they'd be connected together.

   Using the term "giant" a bit loosely for a moment, this means that real networks generally have either zero or one giant components, but not more than one. When you look at a large network, it's useful to know which of these two cases you're in: whether there's a single giant component, or whether all components are small. In the former case, the giant component tends to contain the interesting structure in the network, while in the latter case, there often is very little structure of interest (since the network is pulverized into small pieces). The fraction of nodes in the largest component is a simple measure that conveys this.

3. *Average Pairwise Distance.* The average distance between all pairs of nodes is a simple way to determine whether the network is very compact — with short paths linking most pairs — or "spread out," with many long paths. Of course, if two nodes are not in the same component, it's not clear how to include them in the average, so the average pairwise distance is often computed just over the nodes in the largest component. Also, one can equally well look at the median path length over all pairs of nodes, or the k^{th} percentile path length for any value of k.

4. *Clustering Coefficient.* Social networks tend to have many *triangles* — three nodes with edges between all of them. (Sociologists also refer to triangles as *closed triads.*) The reason for this abundance of triangles is the importance of *triadic closure* in the formation of social networks, as Granovetter discusses in his paper [2]: if v is friends with both u and w, then u and w are likely to become friends as well. (Note that we're talking here about friendship or collaboration networks; the social network in Figure 2, because it is restricted to high-school romantic relationships, does not have a lot of triangles, nor is triadic closure a standard mechanism by which such a network grows.)

To measure just how rich in triangles a network is, one can use the *clustering coefficient* [6]. This is defined as follows: over all sets of three nodes in the graph that form a connected set (i.e. one of the three nodes is connected to all the others), what fraction of these sets in fact form a triangle? This fraction can range from 0 (when there are no triangles) to 1 (for example, in a graph where there is an edge between each pair of nodes — such a graph is called a *clique*, or a *complete graph*).

References

[1] Peter Bearman, James Moody, and Katherine Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110(1):44–99, 2004.

[2] Mark Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[3] F. Heart, A. McKenzie, J. McQuillian, and D. Walden. *ARPANET Completion Report.* Bolt, Beranek and Newman, 1978.

[4] Jon Kleinberg and Éva Tardos. *Algorithm Design.* Addison Wesley, 2006.

[5] Gueorgi Kossinets and Duncan Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.

[6] Mark E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(026118), 2001.