Graphs and Networks Lecture 2

Observational studies of networks

Daniel A. Spielman September 7, 2010

2.1 Introduction

There are also powerpoint slides for this lecture. Please consult them for examples.

I also recommend looking at the papers listed as being related to this lecture on the course web page.

2.2 Fundamental Properties of a Graph

When you get a graph, the first things you should compute are its number of vertices and number of edges. Throughout this course n will always denote the number of vertices and m will always denote the number of edges.

The degree of a vertex is the number of edges that are attached to it. This is well-defined for undirected graphs. For directed graphs, we will talk about the in- and out-degrees of a vertex, which are the number of edges entering and leaving the vertex respectively. If I do just say "degree", you should check whether I've made a mistake, or whether I mean the sum of the in- and out-degrees.

The average degree of a vertex is easy to compute. We could just average the degrees. Or, we could observe that the average degree is $2m/n$. To see this, note that the sum of the degrees of the vertices is $2m$. This is because we count every edge twice when we sum the degrees of the vertices.

2.3 Components

A set of vertices is said to be connected if there is a path in the graph between every pair of vertices in the set. A component in an undirected graph is a maximal connected set of vertices. Recall that a set S is maximal set having some property if S has the property and no super-set of S has the property. Whenever you have an undirected graph, you should first divide it into its connected components. Note that this decomposition is unique.

The situation is slightly more complicated for directed graphs. We say that a set of vertices in a directed graph is *strongly connected* if there is a path from every vertex in the set to every other vertex in the set. A strongly connected component of a directed graph is a maximal set of vertices that are strongly connected. The vertices of a directed graph may be partitioned into strongly connected components in exactly one way. However, there may be edges going between these components. The one thing that we do know is that these edges can only go one way. If A and B are strongly connected components in a graph, then there can be edges from A to B or from B to A. But, if both types of edges existed then $A \cup B$ would be strongly connected, and therefore A and B would not have been strongly connected components as we define strongly connected components to be maximal.

Now, consider the graph on the strongly connected components. That is, create a new graph with one vertex for each strongly connected component in the original graph. Put edges from one of these vertices to another if there are edges in the original graph from the corresponding component to the other. This graph may be obtained by *contracting* all the vertices in the strongly connected components. This graph is directed, and we can see that it has no cycles. If it did have a cycle, then the components we identified would not have been maximal. So, this graph is a *directed acyclic graph* (DAG).

2.4 Distances

In an undirected graph, we define the distance between vertices u and v to be the least number of edges one must traverse to get from u to v. In particular, if (u, v) is an edge then the distance between u and v is 1.

In directed graphs, we define the distance from u to v to be number of directed edges one must traverse to get from u to v. In general the distance from u to v may be different from the distance from v to u.

The *diameter* of a graph is maximum over vertices u and v of the distance from u to v. Of course, this is infinite if the graph is not connected (in the undirected case) or strongly connected (infinite case).

It has been found that many graphs have the "small-worlds" property: most pairs of vertices are connected by short paths. It would be more accurate to say that most pairs are connected by short paths when one can get from one to the other.

2.5 Examples

At this point, we will discuss examples from the papers

1. Graph Structure in the Web, by Broder *et. al.*, *Computer Networks* (33), 2000. pp. 309-320.
2. The Structure of Scientific Collaboration Networks, by Newman. Proc. Natl. Acad. Sci. USA 98, 404-409 (2001).

2.6 Clustering Coefficients

For a long time, sociologists have been interested in what they call transitivity or clustering in networks. I strongly dislike this use of the term "clustering". But, I'll go with it for now.

Consider a social network, in which vertices are joined by edges if the corresponding people are friends. The clustering coefficient of a vertex v, written c_v, the the probability that two friends of v are also friends. If d_v is the degree of vertex v then this may be written as

$$\frac{\text{number of edges between neighbors of v}}{\binom{d_v}{2}}.$$

For an example of how this has been used, Bearman and Moody (American Journal of Public Health, Jan 2004) found that teanage girls whose friends have low clustering coefficients are more likely to contemplate suicide.

One way of creating a global measure of the clustering coefficient is to simply average this over all vertices. The other approach does not break it down by vertex, but rather measures over the whole graph the fraction of triples u, v, w with (u, v) and (v, w) edges for which (u, w) is also an edge. This is equal to

$$\frac{\text{6number of triangles in the graph}}{\text{the number of paths of length 2 in the graph}}.$$

The second of these is called $C^{(1)}$ and the first is called $C^{(2)}$ in the table in Newman's paper "The Structure and Function of Complex Networks". To my eye, it seems that these coefficients are much higher for social networks than for technological or biological networks.

2.7 Assortative Mixing in Networks

The coefficient of assortativity, usually denoted r, measure to what extent edges connect vertices of similar degrees. This was defined my Newman (Assortative Mixing in Networks, Phys. Rev. Lett. 89, 208701 (2002)). It may be defined as Pearson's correlation coefficient of two variables which I will now define. I will call the first d_1 and the second d_2. We sample them by choosing a random edge (u, v) and also choosing the order of u and v at random. The first variable d_1 is then the degree of u and the second variable d_2 is the degree of v. Clearly d_1 and d_2 are identically distributed. But, they are not necessarily independent.

We then have

$$r = \frac{\text{cov}(d_1, d_2)}{\text{std}(d_1)\text{std}(d_2)}.$$

Of course, d_1 and d_2 have the same standard deviation. If every node in the graph has the same degree then the standard deviation will be zero. In this case, we set $r = 1$.

Let me demonstrate how I compute this in matlab. I will begin by creating a very simple graph: a path on four vertices.

```
>> a = diag(ones(1,3),1); a = a + a'

a =

     0     1     0     0
     1     0     1     0
     0     1     0     1
     0     0     1     0

>> degs = sum(a)

degs =

     1     2     2     1

>> [ai,aj] = find(a);

>> [ai, aj]

ans =

     2     1
     1     2
     3     2
     2     3
     4     3
     3     4

>> degi = degs(ai); degj = degs(aj);

>> [degi(:), degj(:)]

ans =

     2     1
     1     2
     2     2
     2     2
     1     2
     2     1

>> mu = mean(degi);
>> std = std(degi,1);

>> r = mean( (degi - mu) .* (degj - mu) / (std^2) )
```

```
r =

   -0.5000
```

If every edge connects vertices of the same degree, then the assortativity will be 1. The minimum value of the assortativity is -1, which is achieved by star graphs. These are the graphs having one distinguished vertex and edges from that vertex to every other vertex. For example, the path on three vertices is a star graph.

2.8 Heavy-Tailed Distributions

Much has been made of the observation that the distribution of degrees in many real-world graphs have a heavy tail. This means that there is a reasonable chance of seeing vertices of very large degree. For example, it has been suggested that the vertex degrees follow a power-law. This would mean that the fraction of vertices of degree k is proportional to

$$k^{-\alpha}$$

for some fixed α.

I'm not sure that I believe all of these results, and there is good reason to be suspect of some. But, we can all agree that the distribution has a heavy-tail. For example, it looks nothing like a normal, binomial or Poisson distribution, which would give a fraction of vertices of degree k proportional to

$$e^{-k^2/\alpha} \quad \text{or} \quad 3^{-k/\beta}$$

for some constants α or β.

There were three problems with the initial studies that indicated that degrees of vertices followed a power law:

1. The way in which the graphs were generated in the traceroute studies produces power-law distributions, even for regular graphs (see the paper "Sampling Biases in IP Topology Measurements" by Lakhina *et. al.* from INFOCOM 2003).
2. Even when the data satisfies a power law, the techniques used to determine the exponents have been systematically flawed.
3. Very few of these studies tested whether other heavy-tailed distributions fit the data better. For example, one should also consider lognormal distributions.

But, the biggest problem with these observations was the hype that was generated around them. They led physicists to conjecture that there was some universal process producing power laws. Now that we've had time to reflect, we see that there is little support for this hypothesis. I'll add a bibliography of relevant papers on this topic to these notes when I get a chance.