Lect 2: empirical analyses of graphs

Tuesday, September 11, 2007
8:30 AM

Disclaimer

These are my personal notes from this lecture. They may be wrong or inaccurate, and have not carefully been edited. But, they are better than nothing.

I. Define a few quantities of interest in a graph
   a. n = number of vertices
   b. m = number of edges
   c. d_{ave} = average degree = 2 m / n

   d. Diameter:
      i. The distance between two nodes is least number of edges in a path connecting them. The diameter is the maximum over pairs of vertices of the distance between them. That is, the distance between the furthest pair.
   e. Bisection size: least number of edges need to remove to divide graph into parts, each having less than half the vertices. (NP complete)
   f. If disconnected, the sizes of components (but why consider a disconnected graph?) If many natural graphs, there is a very large connected component, called the giant component.

II. Before looking at real-world graphs, let's examine these quantities on some archetypical graphs.
   a. The path on n vertices:
      i. Ave degree ~ 2
      ii. Diameter n - 1
      iii. Bisection size 1
   b. The grid on n vertices (sqrt(n) x sqrt(n)):
      i. Ave degree ~ 4
      ii. Diameter ~ 2 sqrt(n)
      iii. Bisection size sqrt(n).
   c. The complete binary tree on n vertices, n = 2^k - 1.
      i. Ave degree ~ 2
      ii. Diameter 2*log(n)
      iii. Bisection size 2.
   d. The hypercube on n vertices, n = 2^d.
      i. Ave degree log(n)
      ii. Diameter log(n)
      iii. Bisection size n/2.

I hope that these examples reveal that these parameters have little to do with each other. Also note that you can glue two or more of these graphs together to get strange things.

Let me mention one more exceptional example.

e. An expander graph on n vertices in which each node has degree 3.
   (one can construct one of these by choosing a 3-regular graph at random.
   Margulis was one of the first to provide explicit constructions).
   i. Ave degree 3
   ii. Diameter 2 $\log_2$ n + O(1)
   iii. Bisection size c*n for some constant c, independent of n.

This sort of graph will be the counter-example to many conjectures about graphs, and is incredibly useful in some fields. We will encounter them again later.

The following table describes graphs was made on authors of papers in various communities. A vertex is associated with each author, and an edge is put between each pair of authors who have co-authored a paper. Below, max distance is diameter, and mean distance is the average distance between randomly chosen authors.

		Los Alamos e-Print Archive					
	MEDLINE	complete	astro-ph	cond-mat	hep-th	SPIRES	NCSTRL
total papers	2156769	98502	22029	22016	19085	66652	13169
total authors	1338969	52909	16706	16726	8361	56627	11994
first initial only	1006412	45685	14303	15451	7676	47445	10998
mean papers per author	5.5(4)	5.1(2)	4.8(2)	3.65(7)	4.8(1)	11.6(5)	2.55(5)
mean authors per paper	2.966(2)	2.530(7)	3.35(2)	2.66(1)	1.99(1)	8.96(18)	2.22(1)
collaborators per author	14.9(1.1)	9.7(2)	15.1(3)	5.86(9)	3.87(5)	179(6)	3.59(5)
cutoff z_c	7300(2700)	52.9(4.7)	49.0(4.3)	15.7(2.4)	9.4(1.3)	1200(300)	10.7(1.6)
exponent τ	2.5(1)	1.3(1)	0.91(10)	1.1(2)	1.1(2)	1.03(7)	1.3(2)
size of giant component	1190488	44327	14845	13961	5825	40902	6506
first initial only	892193	39709	12874	13324	5593	43089	6706
as a percentage	87.3(7)%	85.4(8)%	89.4(3)	84.6(8)%	71.4(8)%	88.7(1.1)%	57.2(1.9)%
2nd largest component	56	18	19	16	24	69	42
mean distance	4.4(2)	5.9(2)	4.66(7)	6.4(1)	6.91(6)	4.0(1)	9.7(4)
maximum distance	21	20	14	18	19	19	31
clustering coefficient C	0.072(8)	0.43(1)	0.414(6)	0.348(6)	0.327(2)	0.726(8)	0.496(6)

From "The structure of scientific collaboration networks" by M.E.J. Newman, arXiv:cond-mat/0007214

How should we interpret these numbers? Newman's suggestion is to contrast the mean distance (or the diameter), with what one would find in a random graph with the same number of vertices and the same average degree. He finds that the average distance between vertices is approximately what one would expect to find in such a random graph: (log N / log ave-degree):

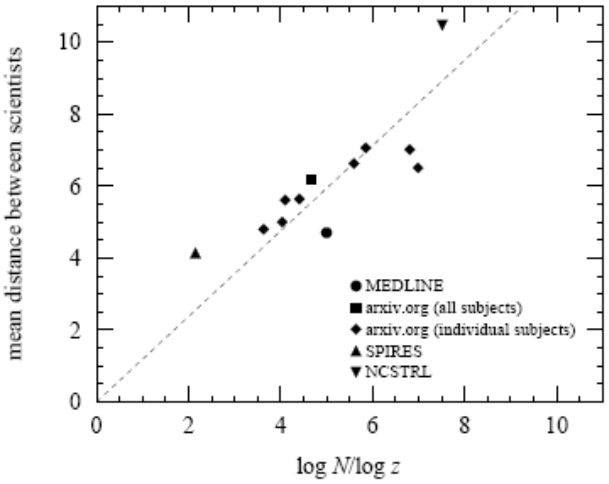


FIG. 3. Average distance between pairs of scientists in the various communities, plotted against the average distance on a random graph of the same size and average coordination number. The dotted line is the best fit to the data which also passes through the origin.

Clustering Coefficients:

The clustering coefficient is a popular thing to examine in a graph. It measures the likelihood that neighbors of a vertex are also neighbors. There are various ways of measuring it. To begin, let's consider clustering at a particular vertex. Look at all of its neighbors, and look at how many of those neighbors are also neighbors of each other. (Note that all the graphs I mentioned at the beginning have none).

Strogatz and Watts define the clustering coefficient by first considering an individual vertex. For vertex i, they define the clustering coefficient C_i to be the number of edges between its nbrs divided by the total possible number of such edges. Let U denote the neighbors of vertex i, let d be the size of U, and let E[U] denote the set of edges between vertices in U. Then

$$C_i = \frac{|E[U]|}{\binom{|U|}{2}}$$

They then define the clustering coefficient C to be the average of these:

$$C = \frac{1}{n} \sum_i C_i$$

Newman defines the clustering coefficient slightly differently. If we let U_i be the neighbors of vertex i, then his cluster coefficient is

$$C = \frac{\sum_i |E[u_i]|}{\sum_i \binom{|u_i|}{2}}$$

That is, the number of neighbors of vertices that are connected, divided by the maximum number possible.

Both of these numbers lie between 0 and 1. I don't know which is better. That is a good problem for someone. First, you would have to figure out how to tell what it means for a measure to be good, and that is quite a problem.

Assortativity

A graph is said to be assortative if vertices have some type, and edges typically appear between nodes of the same type. It is disassortative if edges typically appear between nodes of different types. For example, let's say we are doing a study of relationships among Yale students. We could divide them into types "science types" and "others".

Assortativity is a quantity that measures how assortative a network is. There are a few definitions in the literature. But, I don't know what motivates them. So, here's my proposed definition of assortativity:

> let f be the fraction of edges that go between vertices of the same type.
> let r = 2f - 1.

So, f lies between 1 and 0, and r, which we will call the assortativity, lies between 1 and -1, with 1 being very assorative, and -1 being disassortative.

Newman suggests a definition which I find less obvious than this one. Maybe he has reasons, but he doesn't say why. Can anyone figure it out?

I began by introducing assortativity this way just to motivate it. What we are actually going to be interested in is assortativty by degree. While we might not know the types of nodes in a network, we do at least know their degrees.

Here the way I will first define it. Consider a random edge in the graph, and treat one vertex as the "first" vertex and the other as the "second". Let X be the random variable that is the degree of the first vertex, and Y be the degree of the second. X and Y clearly have the same distribution. We will be interesteed in how correlated they are. Newman defines assortativity by degree to be the Pearson Correlation Coefficient of these variables. The easiest definition is:

$$r = \frac{Cov(X,Y)}{std(X) \cdot std(Y)}$$

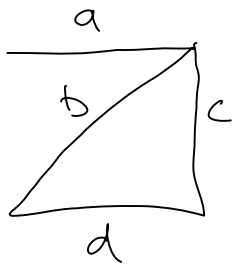
That is, the covariance of X and Y, divided by the product of their standard deviations.

To make that more explicit, let mu be the average degree, which is the mean of X and Y. Then, r is the expectation of (X-mu)*(Y-mu), divided by the variance of (X-mu).

Even more explicitly, one can treat X and Y as vectors indexed by edges of the graph (where we treat the edges as having a first and second vertex). For each such oriented edge e, we can set X_e to be the degree of the first vertex in e, and Y_e to be the degree of the second vertex in e. Then, r is the angle between the vectors (X - mu***1**) and (Y- mu***1**). Algebraically, this is given by:

$$r = \frac{(X-\mu\mathbf{1})^T(Y-\mu\mathbf{1})}{\|X-\mu\mathbf{1}\| \cdot \|Y-\mu\mathbf{1}\|}$$

Here's an example. I've labeled each edge, and give the vectors.
Note that each edge appears twice in the vectors.



	X	Y
a:	1	3
a:	3	1
b:	2	3
b:	3	2
c:	2	3
c:	3	2

c: 3 2
d: 2 2
d: 2 2

Here's how I then compute the assortativity in Matlab:

```
>> a = [1 3; 3 1; 3 2; 2 3; 3 2; 2 3; 2 2; 2 2]

a =

     1     3
     3     1
     3     2
     2     3
     3     2
     2     3
     2     2
     2     2

>> mu = mean(a(:,1));
>> r = (a(:,1)-mu)' * (a(:,2) - mu) / norm(a(:,1)-mu)^2

r =

   -0.7143
```

The reason that I get excited about assortativity is that it seems to distinguish between different types of networks. Here is the table from Newman's paper. Note that most social networks have positive assortativity, while the technological networks display negative assortativity.

	Group	Network	Type	Size n	Assortativity r	Error σ_r
	a	Physics coauthorship	undirected	52 909	0.363	0.002
	a	Biology coauthorship	undirected	1 520 251	0.127	0.0004
	b	Mathematics coauthorship	undirected	253 339	0.120	0.002
Social	c	Film actor collaborations	undirected	449 913	0.208	0.0002
	d	Company directors	undirected	7 673	0.276	0.004
	e	Student relationships	undirected	573	−0.029	0.037
	f	Email address books	directed	16 881	0.092	0.004
	g	Power grid	undirected	4 941	−0.003	0.013
Technological	h	Internet	undirected	10 697	−0.189	0.002
	i	World Wide Web	directed	269 504	−0.067	0.0002
	j	Software dependencies	directed	3 162	−0.016	0.020
	k	Protein interactions	undirected	2 115	−0.156	0.010
	l	Metabolic network	undirected	765	−0.240	0.007
Biological	m	Neural network	directed	307	−0.226	0.016
	n	Marine food web	directed	134	−0.263	0.037
	o	Freshwater food web	directed	92	−0.326	0.031

Of course, somone should try to reproduce this result, making sure that the results are not due to sampling bias (more about that later in the semester).

Caveat/Research Problem

Whenever someone tries to summarize a network by just one parameter, they are implicitly assuming that this parameter is somehow uniform throughout the network. This might be the case, but I'm not sure. It would be interesting to test this implicit hypothesis.

Directed Graphs

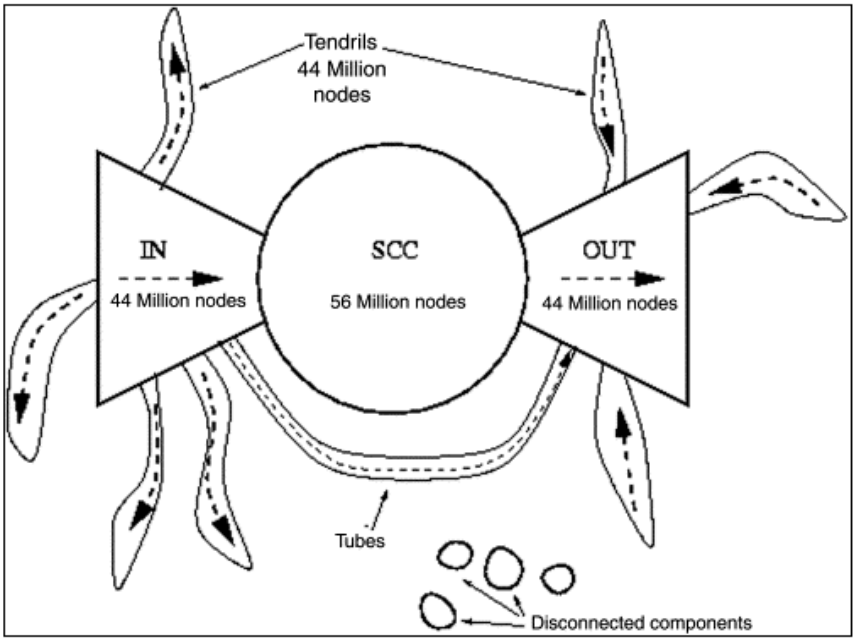
For directed graphs, there are a few other things to consider. First of all, you can't necessarily get from any node to any other node by a directed path. We call a set of nodes S strongly connected if for each pair of nodes in S there is a path from one to the other.

We say that S is a strongly connected component if it is strongly connected, and there is no other vertex that can both reach S and is reachable from S.

In a directed graph, it is interesting to investigate the sizes of the strongly connected components.

There are also additional structures attached to the strongly connected components: there are vertices that can reach them, and are vertices reachable from them. Let's look at an example.

The following figure is a sketch of the web graph generated by researchers at Altavista in 1999. It appeared in the paper "Graph Structure in the Web", Broder et. al., Computer Networks 33, (2000) pp.309-320.



You will see that they found that the graph had a very large strongly connected component (with around 1/4 of the vertices), and that around 1/2 of the vertices could either reach this component, or could be reached from it.

They also investigated the average distance between randomly chosen pairs of vertices. They found that about 1/4th of the pairs had a directed path from one vertex to the other. The average lenghts of the paths are given in the following table.

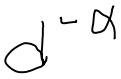
Edge type	In-links (directed)	Out-links (directed)	Undirected
Average connected distance	16.12	16.18	6.83

They also tried to make some measurements of the diameter of the graph. They do this by first choosing a vertex, and then computing the distance of it to every other vertex. (this is done by Breadth-First-Search, the algorithm in which you first identify all neighbors of a vertex, then all their neighbors, etc.). The distance of the furthest node from a given vertex v is the depth of the graph with respect to v.

They did this both by considering links in the directions given, and by reversing the directions of the links. Starting from vertices in the large strongly connected component, they computed the depth of the graph from those vertices. The minimum depth that they found was 475, and the maximum was 503. This is a lot more than "6 degrees of separation". It also tells them that the diameter of the strongly connected component is at least 503 - 475 = 28.

Degree distributions

I ended class by talking a little bit about degree distributions. Many papers have been published that demonstrate that the degrees of vertices in various networks satisfy a power law. That is, that the proportion of vertices of degree d is proportional to



For some non-negative constant alpha.

Much of this work is now viewed as suspect, with some conclusions being attributed to bad statistics, and some being attributed to sampling error. In a later lecture, I will talk about problems that come from sampling. In the meantime, I suggest that you read some of Newman's big survey or the Broder et. al. paper for a credulous analysis, Evelyn Fox Keller's essay for a critique, and Robinson's article for a short description of the problems that come from sampling. (all linked to under this lecture)