36-720: Statistical Network Models

Mini-semester I, Fall 2016
Cosma Shalizi

Scribe: Momin M. Malik

7 September 2016, Lecture 3: Visualization and descriptive statistics

Agenda

Summary statistics/Centralities/Spectra

- Degree
- Distances, diameter, "closeness", "betweenness"
- Eigenvector centrality, pagerank, etc.

Today: What you should do with the data set once you have it. Before you go fitting fancy models, you should poke around the data and see what's going on (something we try to drill into people in every stats class). Do exploratory analysis. Look at some summary statistics and basic visualizations. What are some things that are useful by way of exploratory analysis, for getting a feel for what's going on in the data? Next time, actual statistical models with probability, estimation equations, etc.

Recall:

Network = real pattern of binary relations in the world. Some pattern of relationships between entities.

For mathematical purposes, we abstract as a set of vertices, with edges linking those vertices:

Graph $\equiv G = (V, E)$ where $E \subseteq V \times V$, some subset of the possible pairs of nodes. Typically we do not all self-edges; edges are possibly asymmetric/directed.

Other thing we talked about last time: data collection. Have in the back of your head, what you have might be the complete network carefully surveyed from a careful census, or the result of a weird sampling process, or produced by a powerful and malicious system administrator trying to fool you. Should think about where it comes from and how this might influence anything you are doing further downstream.

Jargon:

Adjacency matrix:

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases}$$

Degree of i = # of edges to $i = \sum_{j=1}^{n} A_{ij}$ or A_{ji} (sum of all edges in row or column)

If directed, have to specify row or column for in-degree and out-degree.

Connected components: largest group of nodes you can have which are connected to each other.

Counts of subgraphs or motifs (e.g., # of triangles)

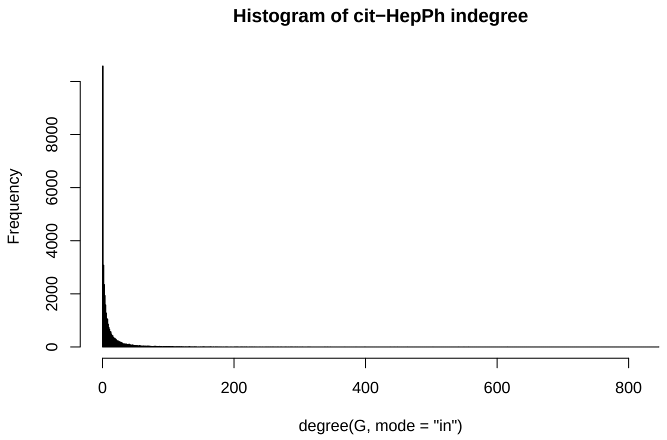
Distance between two nodes: number of edges on the shortest path between nodes. Also called a geodesic path (minimizing distance as though it were going over the surface of the earth).

Summary statistics

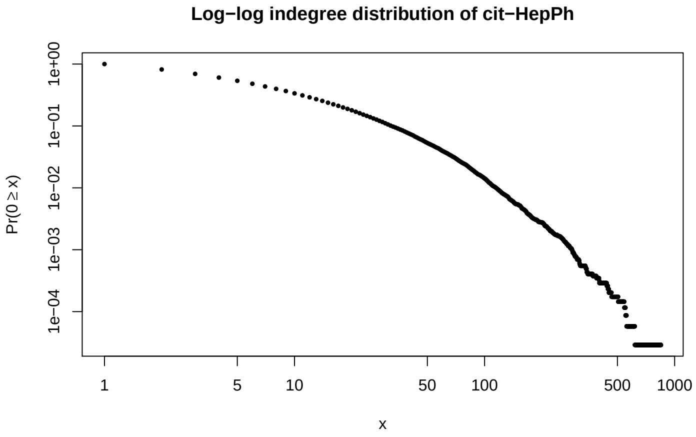
Summaries: statistics (i.e., functions of the data alone) which we hope are useful for a sense of the graph. Generally not going to be sufficient: won't tell you everything that is going on with the network, usually will be losing a lot of information. The hope of using them is exploratory: after you've calculated them for enough graphs, you get a sense of what is a reasonable deviation, and when there is something funny worth investigating.

- *Degree* of each node is a perfectly good summary statistic. There are various things you can calculate around degree.
  - First: the vector of degrees.
  - Can reduce that to a *degree sequence* (e.g., "7 nodes of degree 5," rather than keeping track of which individual node has which degree).
  - A degree distribution. At this point you have a univariate quantity for each node, and can use any of the summary statistics you learned in baby stats. Mean, median, sd, interquartile range, 90% percentile, kurtosis if you really have a reason to do so, anything which you can calculate for a univariate quantity. Can pretend momentarily that is a distribution, do any reduction you like. When we come later in the course to look at different processes going on in networks (diffusion, contagion), we see the behavior of these processes can depend a lot on the degree distribution (median degree, skew of degrees, etc.).

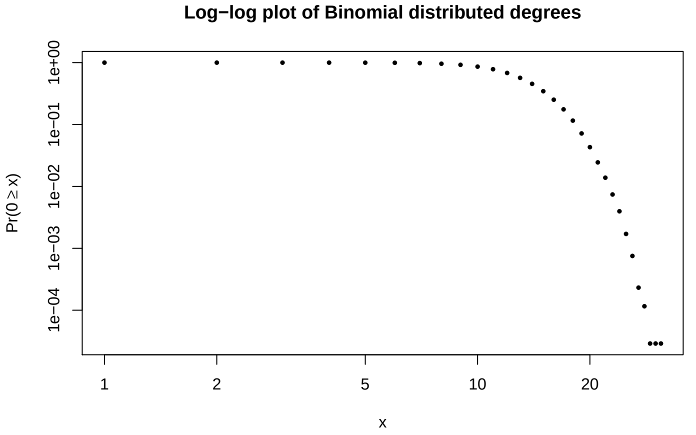
You should worry if the degree distribution isn't right-skewed and heavy-tailed. Be suspicious, double-check data. Certainly can find networks with symmetric distributions with light tails, but if your practice is anything like Cosma's has been, will have a strongly right-skewed and heavy tailed distribution. If you plotted on an ordinary scale, you would see it falls off like this:



If you plot on log-log scale, $\log(Pr(0 \geq d))$, it looks like:



Whereas a Gaussian or exponential cumulative distribution function looks like:



Beyond a point, it is exponentially rare to be 1, 2 standard deviations above the mean. Degree distributions fall off very slowly for very large values of degree. We will come back later in the course to whether this is a power law, exactly algebraic. This has been controversial in the literature.

- *Degree centrality* of of a node = degree. People have devoted a lot of attention to measuring how central, or important, nodes are. Something that makes sense intuitively is that high-degree nodes are more important. If you want to be very formal, you can say there is a thing called the degree centrality of a node, which is just its degree. You can imagine wanting to draw the graph, and wanting to put more central nodes to the center of the picture, and do that by putting the high-degree nodes to the center. For a directed network, you have indegree centrality and outdegree centrality. People will sometimes sum them even in a directed network, but they should make it clear from the writing and math what they are using.

- *Distances between nodes*: almost always means the geodesic distance unless otherwise specified. Write d_{ij} for distance from i to j. Define as $d_{ij} = \infty$ if they are unconnected.

- Average distance between nodes (or between nodes in the same component, if there are multiple connected components)

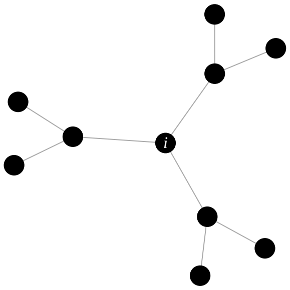
- Maximum distance, has a special name: *diameter* of graph. Typically this is small. Pretty argument for why it is hard to make graphs with large diameter. $\mathrm{diam} G = \max_{i,j} \left[\min_{\text{paths}(i,j)} \text{length of path} \right]$.

Why low diameter is typical: Suppose n nodes, and average degree $\overline{d}$.

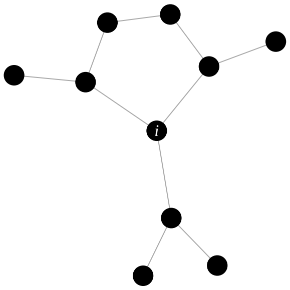
Pick an arbitrary starting node, i. Say that $N(i,r) = \#$ of nodes within r steps of i.

$N(i,1) \approx \overline{d}$. Expect $\overline{d}$ nodes in its immediate neighborhood.

Claim: $N(i,2) \approx \overline{d}(\overline{d}-1)$. The -1 is because need to exclude i. Presumes no overlap in next-nearest neighbors. Picture you should have in mind:



There might be connections within this.



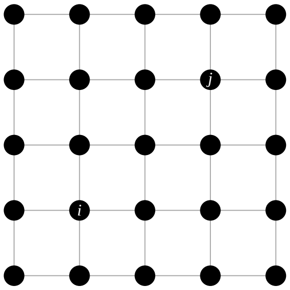
After r steps, by iterating, reach this many nodes:

$$N(i,r) \approx \overline{d}(\overline{d}-1)^{r-1} \approx \overline{d}^r$$

But, there are only n nodes in the entire graph. So $\max r \approx \log n$.

Unless there is a *lot* of overlap in neighborhoods between nodes with a common friend, diameter has to be $O(\log n)$. The number of nodes you can reach at each step increases geometrically, so the maximum number of steps you can take is order $\log n$. This is what is sometimes called the 'small world phenomenon'. For the claim that people in the US can be reached through six steps of separation: log 300m ≈ 6, so this is not too crazy.

For example, low-dimenstional lattices (1, 2, 3 dimensions) break this and have high diameters since there is a lot overlap between neighborhoods of friends of friends. So, not the same sort of exponential growth in the number of friends you can reach. So you get large diameters. Think about what a 2d lattice looks like:



You can get lots and lots of short paths to the same point, because lots of overlap in neighborhoods of neighbors. This is what keeps this from having a diameter that is log of the number of nodes. Instead, the diameter is of order square root of the number of nodes. A p-dimensional lattice typically has diameter $O(n^{1/p})$.

Yet another way to say this is that most networks live in a very high-dimensional space. If you try and think about the graph geometrically, how you have to locate points in space so you have one point for each node in the network and connected them to their nearest neighbor in the space to recover the graph: would have to be done in a very high-dimensional space.

There are graphs where you could do this in a reasonably low number of dimensions (can do lattices in two dimensions), but you are not going to be able to do this for most graphs.

(In graph theory: 'embedding dimension', where you want to connect nodes to their nearest neighbors. Could also try to reproduce the distances of the graph and ask, to do that, how many dimensions do I need? Or enforce, maybe distances are messed up but don't want lines to cross. Graphs that can be drawn in 2d without cross lines are 'planar', but there are very few planar graphs.)

Curse of dimensionality applies: if you try and reproduce the points in space, to have it such that you can get to a large part of the graph in a small number of steps, you need many dimensions. So, network data is intrinsically high-dimensional.

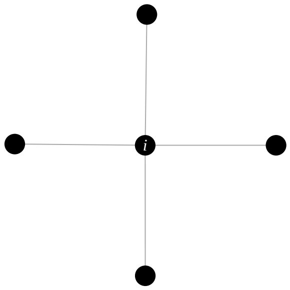
If you are handed an adjacency matrix: look at degree sequence, degree distribution, distances between all the nodes, look at the maximum of these: if this is not very small, again ask yourself why. The maximum distance will upper bound the average distance, you might want to look at the gap between those: can look at the histogram of distances.

More centrality scores: *closeness* and *betweenness*. When they were inventing these things in Harvard in the 50s, decided to call them these, and have stuck with them.

- *Closeness centrality* of node i is n divided by its average distance to other nodes. Formally, the canonical definition is as follows:

  $$C(i) = \frac{n}{\frac{1}{n-1}\sum_{j \neq i} d_{ij}}$$

  The denominator is average distance between node i and everything else in the graph. It is large if node i is close to everything else, and small if i is far from everything else. Normalize this by the maximum possible value, $n-1$ (which occurs in a chain, where the node at the other end has a maximum distance n) to guarantee a number between 0 and 1. Then, normalize this by the number of nodes, and invert so that nodes closer to others have a higher score (centralities usually used as relative ranking within a graph. But sometimes people calculate the maximum possible closeness for a graph of n nodes, and normalize by that). For instance, if you have a node connected to everybody else in a five-node graph,

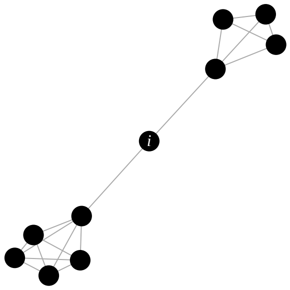
  it has a distance of 1 to everybody else. So, average distance is 1, and its centrality score is 4.

  What about multiple connected components? Two solutions. 1: Use only the component of node i. 2: Use the harmonic mean, which won't give you the exact same numbers:

  $$C_i' = \frac{1}{n-1}\sum_{j \neq i} \frac{1}{d_{ij}}$$

  So infinities become 0, and contribute 0 to the sum. Cosma thinks it makes a bit more sense (and he didn't even come up with it), but not typically used.

- *Betweenness*: how many shortest paths go through a node. Picture you should have in mind is one big cluster, and another big cluster, and one node sitting in the middle:



That one node does not have high degree, and not directly very close to lots and lots of nodes, but clearly if you take i away here, the graph becomes disconnected. It is some kind of choke point for whatever flow is going through the graph: betweenness tries to capture this notion. It does by counting these shortest paths.

Introducing some notation:

$g_{st} = \#$ of geodesic paths linking s and t. Eg., 4-star, clearly only one path between any pair of nodes. But in two clusters connected by one node, many alternative shortest paths.

$g_{sti} = \#$ of geodesic paths linking s and t through i. Then, the betweenness is:

$$b_i = \sum_{s,t} \frac{g_{sti}}{g_{st}}$$

High betweenness doesn't mean the node is 'influential': might be that it is constrained on two sides and is very 'frustrated', but betweenness is a model.

Maximum value is $\binom{n}{2}$, could divide by that to normalize.

(There are many centralities (is work that takes an axiomatic approach to centralities), but betweenness is frequently used in the literature and is implemented in software packages, so it is good to be familiar with it.)

- Circular definitions and eigenvectors: central nodes are ones with many links from other central nodes. (Famous definition of celebrity: somebody who is famous for being famous. Social sciences have anticipated this: one of ways of measuring this is a node is important if it gets links from other important, central nodes.) ("As one of my physics teachers used to say, one man's vicious circle is another man's successful approximation"). This is,

$$v_i(t+1) = \sum_{j=1}^{n} A_{ij} v_j(t)$$

with, say, $v_i(0) = 1$: start with each node having the same importance. Iterate to convergence, and ask what the equilibrium is.

Equilibrium is going to have to be, if I take the old scores and plug them in, I have to get back the same scores,

$$v_i = \sum_j A_{ij} v_j \implies \mathbf{v} = \mathbf{A}\mathbf{v}$$

Which implies the solution will be the eigenvectors of $\mathbf{A}$, and $\mathbf{v}$ must be an eigenvector with eigenvalue 1.

But, maybe no such eigenvector with eigenvalue 1: add on a fudge factor, α,

$$v_i(t+1) = \alpha \sum_{j=1}^{n} A_{ij} v_j(t)$$

Then $v_i = \alpha \sum_j A_{ij} v_j$, and $\mathbf{v} = \alpha\mathbf{A}\mathbf{v}$. So $\mathbf{v}$ must be an eigenvector of $\alpha\mathbf{A}$ with eigenvalue 1.

With multiple connected components, say q, get eigenvectors $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, ..., \mathbf{v}^{(q)}$ with $\mathbf{v}^{(k)}$ non-zero only on component component $\#k$, $k = 1, , q$. (Can be used to find connected components, but doing the eigenvector decomposition of a big graph is considerably slower than other ways of identifying connected components).

Fun linear algebra fact: $\mathbf{A}$ is not any matrix, it is a special matrix. Adjacency matrix, so entries are all 0 or 1, so, all nonnegative. Those matrices have special properties if you look at their eigenvalues. Eigenvalues are all finite, and no negative eigenvalues bigger than positive eigenvalues. $\mathbf{A}$ has non-negative entries so:

1. Biggest eigenvalue is positive, say κ (traditional to call it this).
2. Corresponding eigenvector has all non-negative entries.
3. One such eigenvector per connected component.

Called the Perron-Frobenius theorem (or, Frobenius-Perron). Fundamental fact of matrices with non-negative entries. Comes up with eigenvector problems, and transition matrices for Markov problems, with are also matrices with non-negative entries.

Why bring this up? Set $\alpha = \frac{1}{\kappa}$, then $\mathbf{v}$ is just leading eigenvector of $\frac{1}{\kappa}\mathbf{A}$. Want the eigenvector of $\alpha\mathbf{A}$.

Would it ever not converge? Start with vector $\mathbf{u}$

$$\mathbf{u}(t+1) = \alpha\mathbf{A}\mathbf{u} = (\alpha\mathbf{A})^{t+1}\mathbf{u}(0) = (\alpha\mathbf{A})^{t+1} \sum_{\mathbf{v}_i \in \text{eigenvec}(\alpha\mathbf{A})} \mathbf{v}_i(\mathbf{v}_i \cdot \mathbf{u}(\mathbf{0}))$$

Eigenvectors are a perfectly good basis for space. Take $\mathbf{u}(\mathbf{0})$ and rewrite in this basis.

$$\begin{aligned}
&= \sum_{i=1}^{n} (\alpha\mathbf{A})^{\mathbf{t+1}} \mathbf{v_i}(\mathbf{v_i} \cdot \mathbf{u}(\mathbf{0})) \\
&= \sum_{i=1}^{n} (\mathbf{v_i} \cdot \mathbf{u}(\mathbf{0}))(\alpha\mathbf{A})^{\mathbf{t+1}} \mathbf{v_i} \\
&= \sum_{i=1}^{n} (\mathbf{v_i} \cdot \mathbf{u}(\mathbf{0}))(\alpha\mathbf{A})^{\mathbf{t}} \lambda_{\mathbf{i}} \mathbf{v_i} \\
&= \sum_{i=1}^{n} (\mathbf{v_i} \cdot \mathbf{u}(\mathbf{0})) {\lambda_{\mathbf{i}}}^{\mathbf{t+1}} \mathbf{v_i}
\end{aligned}$$

We know $|\lambda_i| \leq 1$, because of α; the largest eigenvalue is 1, and the others are smaller and go away. So, will converge to 1 if there is only one connected component. If there are multiple, will converge to some mixture.

If you modify this very slightly, come up with PageRank (how much time does a random walk spend on your page, with every so often sending you to a random page).

Diffusion and random walks: closely related operator called the Graph Laplacian, that encodes a lot of information about the graph.

When plotting, can take the eigenvector decomposition, and locate each node according to those eigenvectors. Another way is to turn it into a physics problem.