Catching Paul Revere, An Example Of The Analysis Of A Social Network.

Brian Borchers
Department of Mathematics
New Mexico Tech
Socorro, NM 87801
borchers@nmt.edu

Social Network Analysis

A social network is a mathematical model of relationships between individuals. The model is a graph in which each node corresponds to an individual each edge corresponds to a relationship between two individuals. The edges may also be weighted to reflect the frequency or intensity of interaction between individuals.

Although sociologists once had to painstakingly gather social network data by observing people, social data is now commonly gathered from records of interactions on the internet. For example, the Facebook social network has billions of nodes and edges.

Social Network Analysis

You might expect social networks to have random but largely uniform structure. However, it turns out in practice that most social networks show significant variation. Some individuals are highly connected and important links between groups while other individuals are not so important. In this talk we'll introduce various measures of the importance or centrality of an individual in a social network.

Incidence Matrices

Suppose that we have n individuals, each of which might be a member of any of m groups. We'll use an n by m incidence matrix, A, to represent the memembership of the individuals in the groups.

$$A_{i,j} = \begin{cases} 1 & \text{if individual } i \text{ is a member of group } j \\ 0 & \text{otherwise.} \end{cases}$$

Adjacency Matrices

We'll consider individuals i and j to be connected if they are both members of some group. If we multiply A times A^T, we get

$$(AA^T)_{i,j} = \sum_{k=1}^{m} A_{i,k} A_{j,k}$$

This gives us a count of the number of groups that individuals i and j are both members of. To produce a simple unweighted adjacency matrix, we let

$$G = \text{sign}(AA^T)$$

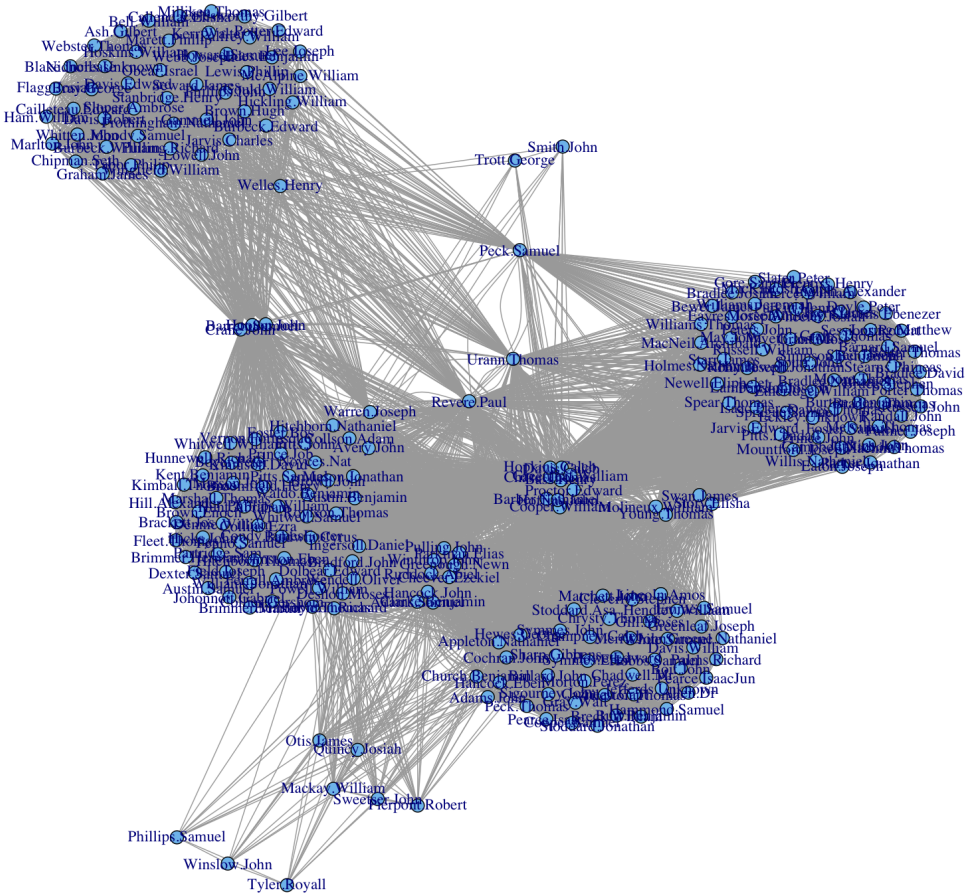
Some of the analysis that follows will be easier if we eliminate self loops by setting

$$G_{i,i} = 0 \;\; i = 1, 2, \ldots, n.$$

The Paul Revere Data Set

In his book, "Paul Revere's Ride", historian David Hackett Fischer gave membership lists for seven social clubs in Boston in 1775. These clubs were hot beds of revolutionary activity. Fischer's incidence matrix will be the basis for our example data set.

The Paul Revere Data Set



Measures Of Centrality

Now we'll explore different ways of measuring the importance or centrality of an individual within a social network. The individuals that are most central to a network are important because much of the communication in the network passes through them. If we're trying to stop a terrorist conspiracy, and we can eliminate a high centrality conspirator, this will disrupt communication between the members of the conspiracy.

We'll use the notative $C(v)$ for some measure of the centrality of an individual, v. Ideally, the measures that we use should tell us something about how individual v fits into the overall structure of the network. $C(v)$ should also be efficient to compute, especially for very large social networks.

Degree

The simplest measure of centrality is the number of other individuals that someone is connected to. In graph theory, this is the degree of a node.

$$C_D(v) = \sum_{j=1}^{n} G_{v,j}$$

The degree is a very easy to compute measure of centrality, but the measure is also quite local. It doesn't take into account anything but the immediate connections of an individual.

In the Paul Revere data, it turns out that Paul Revere has degree 248, and three other individuals are tied for second with degree 200.

Distance Based Centrality Measures

Let $d_{i,j}$ be the length (in number of edges) of the shortest path between individuals i and j.

Let the farness of individual v be the sum of the distances from v to all other individuals. Let the closeness of individual v be the reciprocal of the farness. Then the closeness centrality of an individual v is

$$C_C(v) = \frac{1}{\sum_{j \neq v} d_{v,j}}.$$

In order to compute $C_C()$, we need to solve the all-pairs shortest path problem.

The Floyd-Warshall Algorithm

The Floyd-Warshall algorithm finds the shortest distances between all pairs of vertices in a graph.

We begin with the basic facts that $d_{i,i} = 0$ for $i = 1, 2, \ldots, n$, and that for every edge (i, j), $d_{i,j} = 1$.

The algorithm makes use of a clever recursive formula. Suppose that $d(i, j, k-1)$ gives the length of the shortest path from i to j using only intermediate vertices from the set $1, 2, \ldots, k-1$. Then

$$d(i, j, k) = \min(d(i, j, k-1), d(i, k, k-1) + d(k, j, k-1)).$$

By looping over $k = 1, 2, \ldots, n$, we can eventually compute $d(i, j, n)$ for each pair of vertices i and j. This gives us the desired lengths of the shortest paths from i to j.

The Floyd-Warshall Algorithm

```
function D=floydwarshall(G)
%
% Get the size of G.
%
n=size(G,1);
%
% Initialize D.
%
D=Inf*ones(n,n);
%
% We can reach node v from node v in 0 steps.
%
for v=1:n
  D(v,v)=0;
end
```

The Floyd-Warshall Algorithm

```
%
% For each edge (i,j) we can get from i to j in 1 step.
%
for i=1:n
  for j=1:n
    if (G(i,j)==1)
      D(i,j)=1;
    end
  end
end
```

The Floyd-Warshall Algorithm

```
%
% Now, work through nodes k=1, 2, ...,n, and for all
% node pairs (i,j) establish the length of the shortest
% path from i to j using only intermediate nodes
% 1, 2, ..., k.
%
for k=1:n
  for i=1:n
    for j=1:n
      if (D(i,j) > D(i,k) + D(k,j))
        D(i,j)=D(i,k)+D(k,j);
      end
    end
  end
end
```

The Floyd-Warshall Algorithm

The Floyd-Warshall algorithm can also be extended to produce actual (i, j) shortest paths and to count the number of shortest paths between each pair of nodes. The algorithm can also be extended to deal with graphs in which the edges have lengths other than 1.

Closeness Centrality In The Paul Revere Graph.

It turns out that as measured by C_C, Paul Revere is the most central individual in our social network with $C_C = 0.003876$. He is followed by Nathaniel Barber, William Cooper, and John Hoffins, who are all tied for second at 0.003268.

Betweenness Centrality

Let $\sigma_{i,j}$ be the number of shortest paths from individual i to individual j. This can be computed easily by a variant of the Floyd-Warshall algorithm. Let $\sigma_{i,j}(v)$ be the number of shortest paths from i to j that pass through individual v. We can find $\sigma_{i,j}(v)$ by running Floyd-Warshall on the graph with vertex v removed and subtracting this count of paths from $\sigma_{i,j}$.

The betweenness centrality of individual v is

$$C_B(v) = \sum_{i \neq v, j \neq v} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}}.$$

It shouldn't be surprising that Paul Revere comes out on top under this measure, with a betweenness centrality of 7,328.5.

Eigenvector Centrality

A desirable property of a centrality measure is that it should reflect the centrality of nearby individuals. Let $\delta(v)$ be the set of individuals that are connected by edges to individual v. We would like to have

$$C(v) = \frac{1}{\lambda} \sum_{i \in \delta(v)} C(i)$$

We can write this in matrix form as

$$Gx = \lambda x$$

where G is the adjacency matrix, and x is the vector of centralities. This is just an eigenvector problem. Since G is a symmetric matrix, it will have real eigenvalues and eigenvectors. We'll also require that our eigenvector x be nonnegative.

The Perron-Frobenius Theorem

An adjacency matrix G is irreducible if there is some power k such that every element of G^k is positive. In practice this is nearly always true of social networks.

A version of the Perron-Frobenius theorem states that if the adjacency matrix G is irreducible, then the the largest eigenvalue of G will be simple and the only eigenvector of G with all positive entries will be associated with the largest eigenvalue of G.

The eigenvector centrality $C_{EV}(v)$ is simply the x_v entry of this normalized eigenvector.

Once again, Paul Revere has the highest centrality at 0.1448, with Barber, Cooper, and Hoffins all ties at 0.1386.

Matrix Exponential Measures Of Closeness

The number of paths of length 2 from individual i to j is easy to compute by looking at all possible intermediate nodes k.

$$\gamma_{i,j}^{(2)} = \sum_{k=1}^{n} G_{i,k} G_{k,j}$$

This is simply a formula for matrix multiplication. That is, the number of paths of length two from i to j is $(G^2)_{i,j}$. By induction, the number of paths of length k is

$$\gamma_{i,j}^{(k)} = (G^k)_{i,j}.$$

If we weight the number of paths by the factorial of the length, we end up with the sum

$$\gamma_{i,j} = \sum_{k=0}^{\infty} \frac{(G^k)_{i,j}}{k!} = \left(e^G\right)_{i,j}.$$

Exponential Measures Of Centrality

We can define an exponential centrality as

$$C_{Exp}(v) = \left(e^G\right)_{v,v}.$$

We can also define an exponential betweenness centrality by measuring the average reduction in connvectivity that occurs by removing an individual v from the network. Let $H(v)$ be the adjacency matrix of the graph with individual v disconnected from its neighbors. $H(v)$ can be obtained by zeroing out row v and column v of G.

$$C_{ExpBetween}(v) = \frac{1}{(n-1)^2 - (n-1)} \sum_{i \neq v, j \neq v} \frac{(e^G)_{i,j} - (e^{H(v)})_{i,j}}{(e^G)_{i,j}}$$

Centrality Measures For Very Large Graphs

Centrality measures that require the solution of all-pairs shortest path problems work fine for small social networks, but the computational effort becomes prohibitive for very large social networks.

Because the largest eigenvalue of G is typically much larger than other eigenvalues, iterative methods can be used to quickly estimate this eigenvalue and the corresponding eigenvector. Similarly, there are methods for approximating the matrix exponential of large and sparse matrices.

Further Reading

My interest in the Paul Revere data set began with a blog posting by Kieran Healy, Using Metadata to find Paul Revere http://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/

The data from our example comes from the book "Paul Revere's Ride" by David Hackett Fischer. http://www.amazon.com/Paul-Reveres-David-Hackett-Fischer/dp/0195098315

The material on exponential measures comes from

E. Estrada and D. J. Higham. Network Properties Revealed Through Matrix Functions. SIAM Review 52:696-714, 2010. http://dx.doi.org/10.1137/090761070

Further Reading

Some books on the analysis of social (and other networks) include

M. E. J. Newman. Networks: An Introduction. Oxford University Press, 2010. http://www.amazon.com/Networks-An-Introduction-Mark-Newman/dp/0199206651

D. J. Watts. Small Words: The Dynamics of Networks Between Order and Randomness. Princeton University Press, 1999. http://www.amazon.com/Small-Worlds-Randomness-Princeton-Complexity/dp/0691117047