Graphs and Networks

Supplement for MAT 3310

H. Hirst

This material is inspired by the CCICADA Module "Who is really in charge?"
http://ccicada.org/education/ccicada-education-modules/

Graphs

Introduction to Graphs and Networks

A **graph** is an ordered pair $G = (V, E)$, where V is a set of **vertices** (aka **nodes**) and E is a set of **edges** (aka **links**) between the vertices in V.

Consider the graph defined by $V = \{u, v, w, z\}$ and $E = \{(u, w), (v, w), (z, w), (v, z)\}$. The specific way in which nodes are arranged is not part of the graph, so both visual representations below are the same graph in the mathematical sense.

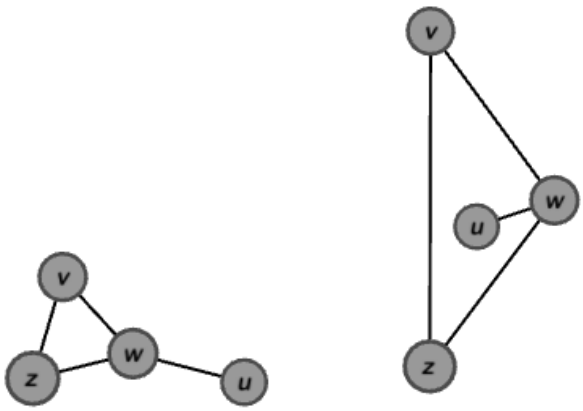


Figure 1. Visual representations of the graph $G = (V, E)$,
with $V = \{u, v, w, z\}$, $E = \{(u, w), (v, w), (z, w), (v, z)\}$

Key Terms

- **Directed / Undirected**: A graph is directed if there is a direction associated with an edge. The graph in Figure 1 is undirected. We will be working with undirected graphs.
- **Loop**: An edge that begins and ends at the same node. There are no loops in Figure 1.
- **Simple Graph**: A graph with no loops or repeated edges. Figure 1 is a simple graph. We will be working with simple graphs.
- **Weighted / Unweighted**: A graph is weighted if the edges have been assigned values that indicate the strength of the relationship between the incident nodes. In Figure 1, we are assuming that the edges are all equal, which we can think of as an unweighted graph. Later we will see that it is helpful to assign all the edges a weight equal to one.
- **Order**: The number of vertices. This value is usually designated by N. In Figure 1, $N = 4$.
- **Degree of a Vertex**: The number of incident (directly connected) edges, denoted "deg." In Figure 1, $\deg(w) = 3$.

- **Adjacent Vertex**: Two vertices are adjacent if they are directly connected by an edge. In Figure 1, u and w are adjacent; u and v are not.
- **Path Between Vertices**: Any sequence of edges that connect a sequence of adjacent vertices. The **length** of the path is the number of edges in the path. An example of a path in Figure 1 is (u, w, v) with length two. Note: Some graph theorists add the provision that a path cannot include vertices more than once.
- **Diameter**: The maximum of the set of shortest path lengths between two nodes. The diameter of Figure 1 is two, because the complete list of shortest paths in Figure 1 are:

(u,w):	1;	(u,w,v):	2
(u,w,z):	2;	(w, v):	1
(w,z):	1;	(v,z):	1

  Note that all possible start-end node pairs are included in this list. Using a counting argument, we can be sure all node pairs are listed by checking that we have listed $N(N-1)/2$ nodes, where N is the order of the graph.
- **Subgraph**: Take any subset of the nodes in a graph and form the set of edges by including all edges between nodes in the subset. $V = \{z, w, v\}$; $E = \{(z, w), (z, v), (w, v)\}$ is a subgraph of Figure 1.
- **Complete Graph**: The complete graph on N nodes has edges connecting every pair of nodes. The number of edges in such a graph (aka the **size** of the graph) can be calculated as $N(N-1)/2$. To extend Figure 1 to a complete graph, we would add (u, v) and (u, z) to the edge set.

Centrality

In analyzing networks – e.g., for spread of information or contagion – identifying key nodes can help determine how to enhance or interrupt the flow. Network analysts have defined several quantities associated with a node that can help determine important nodes in the network. We will examine four: **Degree Centrality**, **Betweenness Centrality**, **Closeness Centrality**, and the **Clustering Coefficient**. In each case we will define a normalized version, so that the measures are in the interval [0,1], with higher values indicating higher centrality or clustering.

- **Degree Centrality**: A measure of the direct connections for a node. A node's degree is the number of edges incident to it, which can also be thought of as the number of nodes adjacent to it. Obviously this number is one or larger so to normalize it, we divide by the number of other nodes in the network.

  In Figure 1, the degree centrality figure for w is calculated as $3/3 = 1$ because w has three incident edges and there are three other nodes in the network. In contrast, u has degree centrality $1/3$.

  What does degree centrality tell us? The node with the most connections has the most ability to influence other nodes directly and immediately, which is important to know if we are trying to stop someone with a communicable disease from infecting others around him for example.
- **Betweenness Centrality**: A measure of how often a node appears in the interior of shortest paths in the network. This is calculated by dividing the number of shortest paths that pass through a node by the number of all shortest paths that do not start or end on the node (remember, we are interested in *between*).

  Recall that the list of shortest paths in Figure 1 is:

(u,w):	1;	(u,w,v):	2
(u,w,z):	2;	(w, v):	1
(w,z):	1;	(v,z):	1

  Thus the betweenness centrality for node w in Figure 1 is $2/3$ because w is in the interior of two of the three shortest paths that start and end elsewhere (i.e., do not count (u, w), (w, v), or (w, z)). Note that all the other nodes have betweenness centrality zero.

Why would we want to know the betweenness centrality of nodes in a network? Betweenness measures how critical a node is for flow through an entire network, especially if our goal is interrupting flow of cash or information in a terrorist network.

Here is a more algorithmic approach to this calculation for a specific node, n:

1. Compute $C = \sum_{s \neq n \neq t} \frac{\sigma_{st}(n)}{\sigma st}$, where
   - s and t are nodes different from n
   - σ_{st} denotes the number of shortest paths from s to t
   - $\sigma_{st}(n)$ denotes that number of those shortest paths that include n
2. Divide C by the number of node pairs excluding n, i.e., $(N-1)(N-2)/2$

- **Closeness Centrality**: A measure of the average distance from a node to all other nodes in the network. To calculate closeness centrality, take the reciprocal of the average of lengths of the shortest paths between the node and all other nodes on the network.

  The closeness centrality of w in Figure 1 is calculated as the reciprocal of $(1+1+1)/3$ because the shortest path from w to each of the other nodes has length one, and there are three other nodes. Thus the closeness centrality for w is 1.0. For node v we calculate the reciprocal of $(2+1+1)/3 = 4/3$, so the closeness centrality for v is 0.75.

  Why would we want to know the closeness centrality of nodes? Closeness indicates that a node can get information out most efficiently to the entire network.

  Here is a more algorithmic version of this calculation for a specific node n:

  1. Calculate $L(n,m)$, the length of the shortest path from n to m, for all nodes m different from n.
  2. Find the reciprocal of $\frac{\sum_{m \neq n} L(n,m)}{(N-1)}$

- **Clustering Coefficient**: A measure of the likelihood that two nodes adjacent to a node are also adjacent to each other, (i.e., the likelihood that a node is part of a triangle). To calculate this coefficient, look at the subgraph formed by all nodes adjacent to the given node and divide the number of edges in that subgraph by the number of edges in a complete graph with the same number of nodes.

  For example, consider z in Figure 1. The nodes connected to z are v and w. There is one edge between v and w, and the complete graph on two nodes has one edge. Thus z has clustering coefficient $1/1 = 1.0$. On the other hand w has three adjacent nodes, u, v, and z; that subgraph has one edge (remember we are taking out w), and the complete graph on three nodes has three edges, so the clustering coefficient for w is $1/3 = 0.333$.

  Here is a more algorithmic version of this calculation for a specific node n:

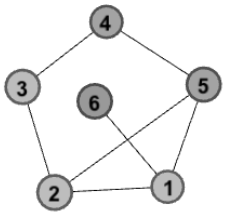
  1. Calculate $N = \deg(n)$.
  2. Determine the subgraph of nodes adjacent to n (not including n).
  3. Divide the number edges in the subgraph by the number of edges in the complete subgraph of the same order $(N(N-1)/2)$.

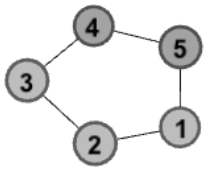
  What would the clustering coefficient tell us? A large coefficient (close to one) indicates that a node is very interconnected with its neighbors, so in a situation like transmission of disease the likelihood of spreading to all neighbors is magnified.

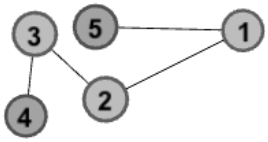
  We can also look at the average clustering coefficient for the entire network (i.e., average all the coefficients for all the nodes in the graph), and if this average is large and the average path length is small ($\leq O(\log(N))$), then graph theorists say that the network exhibits the **"small world" property**, implying that information will spread rapidly and there are fewer key nodes. This situation makes interrupting flow difficult.

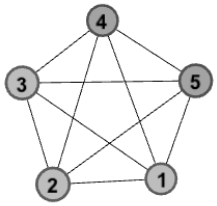
Exercises

For the graphs below, find the order, size, and diameter. Given an example of a pair of adjacent vertices and an example of a path. Find the complete set of shortest paths between pairs of nodes. Calculate the three measures of centrality and the clustering coefficient for each node in the graph.

1. 

2. 

3. 

4. 

For the graphs below, draw two visually different representations and calculate the three measures of centrality and the clustering coefficient for each node in the graph.

5. $G = (V, E)$, where $V = \{a, b, c, d, e, f\}$ and $E = \{(a, b), (a, c), (a, d), (a, e), (a, f)\}$
6. $G = (V, E)$, where $V = \{0, 1, 2, 3, 4, 5, 6, 7\}$ and
   $E = \{(0, 1), (0, 3), (0, 4), (1, 4), (1, 2), (2, 3), (2, 4), (2, 5), (3, 5), (5, 6), (6, 7)\}$
7. $G = (V, E)$, where $V = \{0, 1, 2, 3, 4, 5, 6, 7\}$ and
   $E = \{(0, 1), (1, 2), (2, 3), (3, 4), (4, 0), (4, 2), (1, 5), (3, 6), (5, 7)\}$
8. $G = (V, E)$, where $V =$ {george, steve, andy, scott, phil} and
   $E =$ {(george, steve), (george, andy), (george, scott), (george, phil), (steve, scott), (steve, phil), (andy, scott), (andy phil), (scott, phil)}

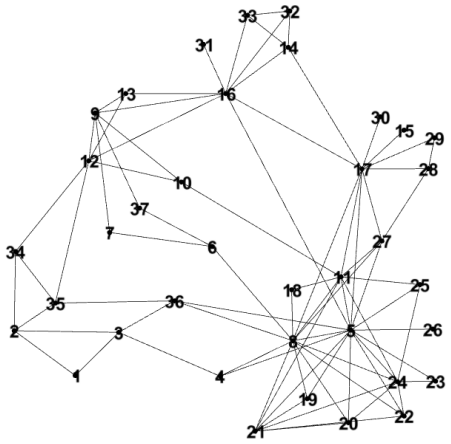
Let's think about centrality and clustering using a larger example...

Terrorist Network Exploration

In his article "Uncloaking Terrorist Networks" (http://journals.uic.edu/ojs/index.php/fm/article/view/941/863), Valdis Krebs analyzes the contact information for the 9/11 Terrorists . He was looking at the flow of money, information, and expertise among more than 70 people identified after the attack as being part of the network of the 19 hijackers.

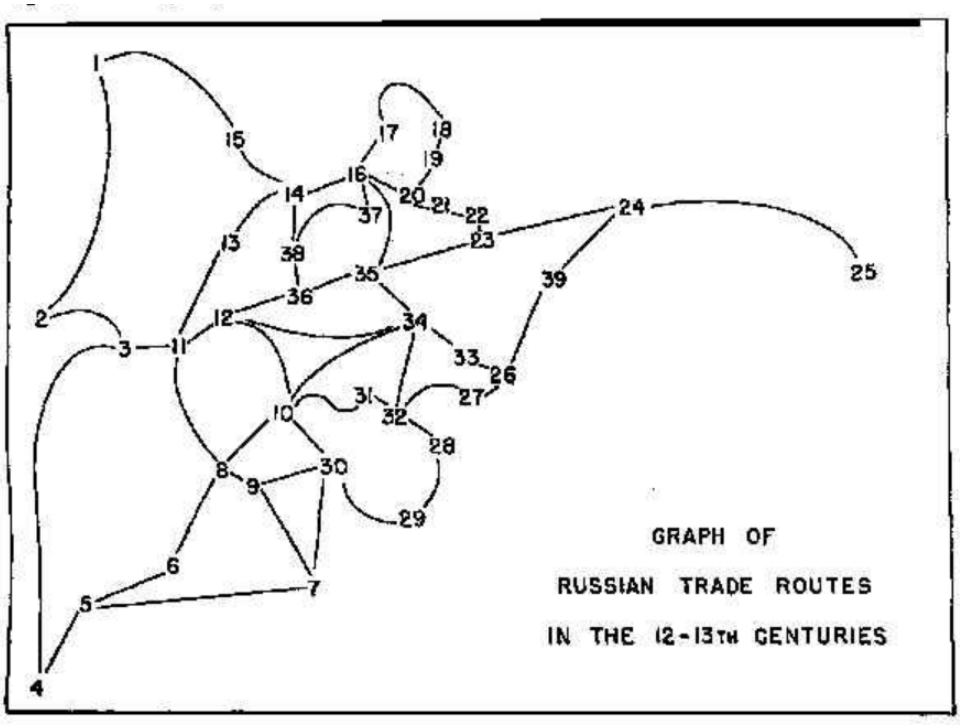
Using a subset of the data that includes the 19 hijackers and their closest associates, the following graph depicts an abbreviated version of the terrorist network. There are 36 nodes and 80 edges, so calculating measures of center by hand would be challenging.

Exploration: Watch the video on using the software package GEPHI. Using the terrorist.csv file and GEPHI, calculate the measures of centrality and the clustering coefficient, and explain which terrorist you think is Mohammed Atta, the purported mastermind of the network. If you were given this terrorist network information and asked to name the top three contacts to target in a raid, which would you name? Explain.



Russian Trade Route Exploration

Many historians have debated why Moscow grew into a center for trade and migration, dominating other towns in Russia in the 13th century. In 1965, Forrest Pitts produced a graph of the major medieval urban locations along Russian river transportation routes (people and goods traveled by boat in summer and by sleigh on the frozen rivers in winter) in an effort to see if a graphical analysis could shed light on the situation. He used a self-designed analysis of shortest paths to show that Kolomna (34 on the network) was the in fact the best connected place with Moscow (35) ranking second. Below is the graph of the river network. Do our measures of centrality support the idea that Moscow was the most important place in the trade network? Which measures do you think are most appropriate? Do the clustering coefficients tell you anything? Explain your results.



From http://www.analytictech.com/networks/pitts.htm

Ego Network Exploration

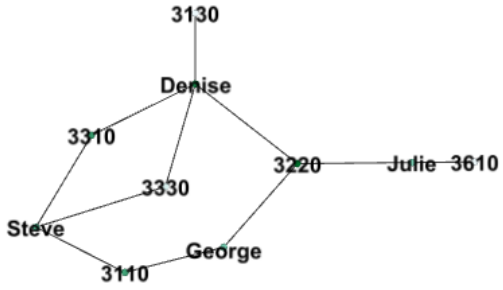
Track your electronic social contacts (texting, email, facebook, etc.) in one 24 hour period, and then ask each one to confirm who on your list they contacted directly. Calculate the diameter of the graph and the degree, betweenness, and closeness centralities for each node, and produce a nice visual of the graph for your social network. Are you the "center" in your social network? Is there someone in your network that is the most "distant" contact? Explain by describing what the diameter, the clustering coefficients, and each of the centrality measures indicate.

Bipartite Graphs – Group Membership

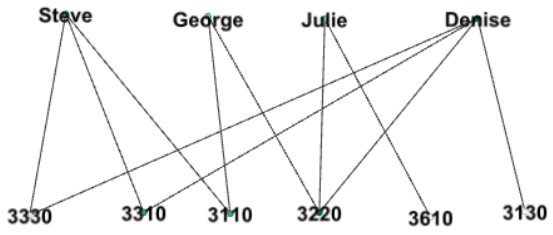
Sometimes we have lists of individuals and the groups they are associated with to analyze, and so the graphs generated really have two different kinds of nodes. For example, consider the following list of students and classes they are taking.

Denise:	3130, 3220, 3310, 3330
Paul:	3310, 3330
George:	3110, 3220
Steve:	3110, 3310, 3330
Julie:	3220, 3610
Kate:	3110

Drawing a graph showing students connected with edges to the classes they are taking could yield:



Since there are really two kinds of nodes, another more standard way of representing this graph is to organize the nodes into two rows (or columns).



This graph is called **bipartite** because it can be organized into two subsets of nodes, with no edge connections within either subset.

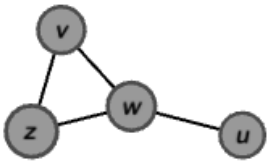
Typical questions we might ask of situations like this are: "Which group (or class in this case) is more influential?" "Which individuals are more strongly connected?"

To answer these questions, we can use linear algebra techniques.

Linear Algebra Connections to Graph Theory

A Brief Detour

For any undirected graph, we can create a matrix representation, referred to as an **adjacency matrix**. Order the node list and create a matrix, A, where the entry $a_{ij} = 1$ if there is an edge between v_i and v_j and 0 if not. This is analogous to assuming all edges have a weight of one.



Recall our graph from Figure 1 (reproduced above). If we let u be node 1, v be 2, w be 3, and z be 4, then the adjacency matrix for the graph is

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Examining the upper triangular portion of this matrix provides us with a complete listing of all paths of length one (i.e., direct edge connections) in the graph. This is not that useful since we can obviously see them in the graph. But! Consider what happens when we calculate A^2 using matrix multiplication.

$$A^2 = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 3 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$$

Consider entry 2, 1 in A^2: The calculation that yielded this value was the dot product of row 2 of A with column 1of A:

$$a_{21} = \sum_{k=1}^{4} a_{2k} \times a_{k1}$$

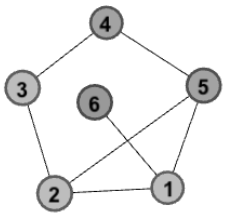
Look at one factor in this sum for a fixed value of k: If there is an edge between nodes 2 and k then $a_{2k} = 1$ and if there is an edge between nodes 1 and k then $a_{k1} = 1$; the product would be one indicating one path of length two between nodes 1

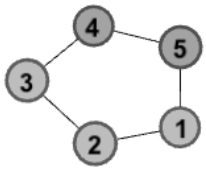
and 2, namely $(1, k, 2)$. If one of these edges does not exist then the product is zero indicating no path of length 2 between nodes 1 and 2. The sum will accrue values for each path of length two between nodes. Cool!

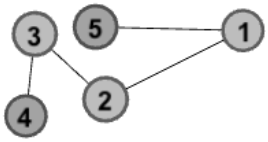
We can use a similar argument to interpret A^3 – the number of paths of length three between two nodes – and higher powers. Note that this interpretation of path allows nodes to be repeated on a path.

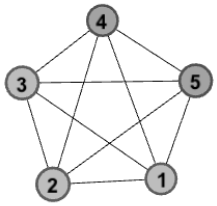
Exercises

For the graphs below, find the adjacency matrix. Use it to find the number of paths of length 3 between each node in the graph.

9. 

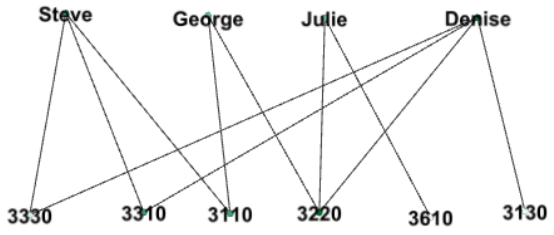
10. 

11. 

12. 

Matrices for Bipartite Graphs

Since there are two types of nodes, we will examine a slightly different version of the adjacency matrix: the **reduced adjacency matrix**. Consider again the bipartite graph of students and classes.



To build the reduced adjacency matrix, we will let one type of nodes be the rows and the other the columns of the matrix. Here is a matrix set up with students as rows and classes as columns.

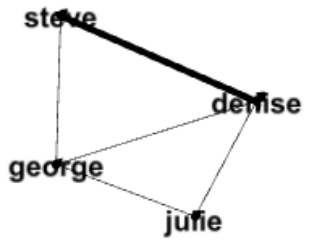
	3330	3310	3110	3220	3610	3130
Steve	1	1	1	0	0	0
George	0	0	1	1	0	0
Julie	0	0	0	1	1	0
Denise	1	1	0	1	0	1

This matrix is not square, so we cannot raise it to a power; however, we can perform the following two multiplications, both of which yield interesting information.

$$AA^\top = \begin{bmatrix} 3 & 1 & 0 & 2 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 2 & 1 & 1 & 4 \end{bmatrix} \qquad A^\top A = \begin{bmatrix} 2 & 2 & 1 & 1 & 0 & 1 \\ 2 & 2 & 1 & 1 & 0 & 1 \\ 1 & 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Person-Person Affiliation Network ($AA^\top$): Each row and column represents a person in the original row order. Matrix multiplication is at work here again. Upper triangular entries give the number of classes two students have in common. For example, Steve (person 1) and Denise (person 4) have two classes in common – the strongest connection between two students.

We can use the matrix to create a graph showing these interconnections, but now the edges will be weighted according to the corresponding entry in the matrix. For example, the edge from Steve to Denise should have a weight of 3 compared to the edge from Steve to George, which should have a weight of 1. One way to visualize the weights is to draw more heavily weighted edges thicker:



Group-Group ($A^\top A$): Each row and column represents a class in the original column order. Upper triangular entries give the number of students two classes have in common, and we can produce a weighted graph for these data as well. In this graph the edge from 3340 to 3310 is weighted most heavily, since there are two students in common, but all other pairs have either none or one student in common.

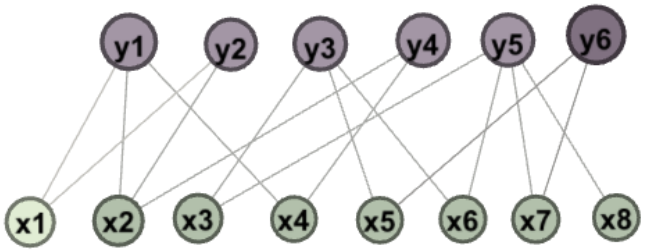
In both of these cases we could tell this information from the original student-class data, but in situations with many individuals associating in many groups, it helps to have the $AA^\top$ and $A^\top A$ matrices and the graph of associations to analyze the situation. Graph theorists refer to these graphs as **projections of the bipartite graph**.

Exercises

For the problems below, construct the bipartite graph and find the reduced adjacency matrix (RAM). Use the RAM to find the person-person and group-group matrices and projection graphs; comment on where the strong connections are between people and between groups.

13. $G = (V, E)$, where $V = \{a, b, c, d, e, f, g\}$ and $E = \{(a, d), (a, e), (a, f), (a, g), (b, e), (b, f), (c, d), (c, g)\}$

14. Abbie, Ben, Coleman, Dwayne, Ellen, Flynn, and Genie all ski. Abbie also likes hiking and kayaking. Coleman and Genie like to skateboard. Ben and Coleman like kayaking. Dwayne likes to skateboard and hike. Flynn likes to hike.

15. 

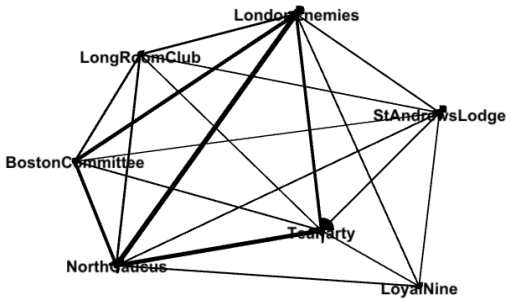
Revolutionary War Connections Exploration

Many of the Revolutionary War patriots were members of social clubs involved in planning (what were then terrorist) acts against the British in Boston. A listing of the memberships of seven of these social clubs was tabulated by historian David Hackett Fisher (http://www.amazon.com/Paul-Reveres-David-Hackett-Fischer/dp/0195098315) and converted to a spreadsheet by sociologist Kieran Healy (https://github.com/kjhealy/revere/blob/master/data/PaulRevereAppD.csv). For example, John Adams was a member of the North Caucus and the Long Room Club, and Paul Revere – perhaps the most famous revolutionary Bostonian – was a member of five of the seven clubs.

The file revolutionary-war-clubs-fulldata.csv contains the Healy data for our use. Taking this membership list and producing the group-group matrix (for these data that is $A^\top A$) gives the following 7×7 matrix.

$$A^\top A = \begin{bmatrix} 53 & 2 & 3 & 2 & 3 & 1 & 3 \\ 2 & 10 & 3 & 0 & 2 & 0 & 3 \\ 3 & 3 & 59 & 5 & 13 & 9 & 16 \\ 2 & 0 & 5 & 17 & 2 & 5 & 5 \\ 3 & 2 & 13 & 2 & 97 & 3 & 8 \\ 1 & 0 & 9 & 5 & 3 & 21 & 11 \\ 3 & 3 & 16 & 5 & 8 & 11 & 62 \end{bmatrix}$$

Examining the upper triangular entries indicates that, at 16 common members, the relationship between social clubs 3 (North Caucus) and 7 (London Enemies) was strongest with club 5 (Tea Party) also playing a possible strong role in collaboration. On the other hand, club 2 (Loyal Nine) has the weakest connection with the other clubs. We can see this also in the weighted graph:



What do the measures of center tell you in this case, you might ask? Along with the edge weights (in this case number of common members), they can confirm strong and weak ties between groups. For example, groups 1, 3, 5, and 7 all have closeness centrality 1, while group 2 (Loyal Nine) has the lowest centrality at .75.

What about looking at relationships between individual patriots? Examining the key roles the individual patriots played in collaborating requires calculating $AA^\top$, which is a 254×254 matrix.

Exploration: Watch the second video on using the software package GEPHI. Use the $AA^\top$ matrix in the file people-people-RAM.txt to explore the relationships among the patriots. Historians indicate that Paul Revere played a key role in activities in Boston. Examine the graph and the measures of center to see if you can confirm this conclusion. Who else appeared to play a key role based on these data?

Darknet Network Exploration

In their article "Bipartite Network Model for Inferring Hidden Ties in Crime Data" (https://arxiv.org/abs/1510.02343), Isah, Neagu, and Trundle were able access a list of Darknet vendors and their wares in an effort to look at criminal networks on the internet. The table below contains an excerpt from that list.

Create a bipartite graph and explore the vendor-vendor and product-product projections. What do the graphs and measures of centrality tell you about the strength of possible associations? Isah, Neagu, and Trundle hypothesize that obtaining illicit substances and data requires a high degree of organization and specialization. Do you see evidence for underlying collusion between the vendors?

Vendor Abbreviation	Products
MrH	Cocaine, Cannabis, Stimulants, Hash
PS24	Financial Account Numbers, Pirated Software, IDs and Passports, SIM Cards
SFex	Benzos, Cannabis, Cocaine, Stimulants, Fake Prescriptions
OzV	Pirated Software, Erotica, E-Books, Meth
OzDDi	Cannabis Seeds, Weed, Meth, MDMA
TT	Financial Account Numbers, IDs and Passports, Credit Card and CVV Numbers
PEAK	Mescaline, Stimulants, Meth, Psychedelics
PAMFET	MDMA, Speed, Stimulants, Ecstasy
PFish	Weight Loss Pills, Stimulants, Fake Prescriptions, Ecstasy