Who Is Really In Charge?

An Application of Graph and Network Theory to Analyzing Social Networks

An Undergraduate Teaching Module

Prepared By:

Maurino Bautista, Rochester Institute of Technology, maurino.bautista@rit.edu
Donna Beers, Simmons College, donna.beers@simmons.edu
Mary Goodloe, Belmont University, mary.goodloe@belmont.edu
Note to teachers: Teacher notes appear in dark red in the module, allowing faculty to pull these notes off the teacher version to create a student version of the module.

Summary of the Module:
The goal of this module is to introduce undergraduates from across disciplines to the new and vital field of network science. Our focus will be to show students how mathematical and computational tools can be used to analyze social networks in order to advance health, prosperity, and security. We will also illustrate the significance of metadata.

Target Audience:
First- or second-year undergraduates

Prerequisites:
There are none.

Learning goals
- Students will acquire knowledge of the terminology, notation, and structures from introductory graph theory which form the foundation for network analysis.
- Students will understand the meanings of - and the differences between - these three measures of centrality for a social network: degree centrality, betweenness centrality, and closeness centrality.
- Students will be able to determine the appropriate measure of centrality that is needed to answer the question, Who is really in charge?, in a variety of real-world settings.
- Students will be able to use the network analysis software Gephi to (1) input and display a given social network, and (2) calculate the degree centrality, betweenness centrality, and closeness centrality metrics for the network.

Learning objectives
- Students will be able to state the definitions for basic concepts from graph theory and matrix theory including: graph, vertex, degree of a vertex, edge, order of a graph, size of a graph, path in a graph, the diameter of a graph, matrix, transpose of a matrix, adjacency matrix of a graph, and matrix multiplication.
- Students will be able to explain the meaning of each of these measures of centrality for a network: degree centrality, betweenness centrality, and closeness centrality.
- When presented with a given network, students will be able to carry out the procedures for calculating the degree centrality, the betweenness centrality, and the closeness centrality measures for the network.
- When presented a set of real-world data, students will be able to use Gephi to input the data, display the data (and revise the display, if necessary, to achieve greater clarity), and calculate the measures of degree centrality, between centrality, and closeness centrality for each node in the network.
Specify Anticipated Number of Class Periods:

Section 2 covers the basics of matrix arithmetic and graph theory. For students with no background in these areas, this section should take two class meetings. If students are already comfortable with matrix arithmetic, only one class meeting should be needed.

Section 3 introduces networks, network analysis, and centrality measures. This section should also take about two class meetings. With a class of mathematically strong students, it might be possible to cover sections 2 and 3 in a week.

Section 4 uses the results of a simple survey to illustrate the basics of the Gephi software. If the instructor preselects survey questions and does all the work in Gephi, it would be possible to conduct the survey at the end of one class period and demonstrate Gephi in the next. However, if the students themselves are working in Gephi, section 4 could take two to three days. How quickly students achieve some basic skill with Gephi will depend on how comfortable students already are with utilizing various software programs.

Section 5 uses both a simple example and a very interesting Revolutionary War example to further illustrate what can be done with network analysis and with Gephi. If the instructor wished to simply show students an interesting application, the highlights of the Kieran Healy article Using Metadata to Find Paul Revere could be covered in a day. If the students work through the various Excel and Gephi processes, this section would take two to three days.

Section 6 provides discussion questions for the article Uncloaking Terrorist Networks by Valdis Krebs. With the article as assigned reading, one class day could be spent in discussion.

Section 7 gives three possible student projects recommended for outside of class meetings. If the students have themselves worked in Gephi in sections 4 and 5, the first project should be straightforward. If the instructor simply demonstrated the activities in sections 4 and 5, then the first project would be an opportunity for the students to explore. The second project is very open-ended. Upper-level, independent students could teach themselves Netlytic and then design their own study. The third project, while again for upper-level students, is a more directed introduction to Netlytic.
# Table of Contents

Section 1: Introduction  
Section 2: Background Concepts from Mathematics  
Section 3: Background Concepts from Network Analysis  
Section 4: Gephi Basics and Activity  
Section 5: The American Revolution  
Section 6: Uncloaking Terrorist Networks  
Section 7: Student Projects  
Endnotes  
References
1. Introduction

Have you ever wondered who is really in charge? Who is at the center of whatever is trending? Who influences policy decisions? Who motivates groups that bring about change?

For example, in the social network below, where two individuals are connected if they regularly interact with each other (e.g., they see each other at the gym or exchange emails or otherwise spend time with each other), who do you think is most influential and why?

Leaders, whether in business, government, or education, are interested in who is at the heart of all that is happening, whether those central individuals are considered to be trendsetters, revolutionaries, subversive elements, or terrorists.

This module serves as an introduction to network analysis by asking you to discover who is in charge in various networks; and, for the one shown above, you may be surprised at the conclusions we will reach by the end of this module. Before directly addressing networks, we cover the necessary mathematics of matrices and graph theory. We proceed to the language of networks, graphs as models for networks, and several measures of centrality for networks. A short tutorial is provided for Gephi, a tool to draw and analyze the graphs that represent networks. Then we use all we have covered to find out who was really in charge among revolutionaries prior to the American Revolution and among the terrorists of 9/11. The final section provides some possible projects for students who wish to study networks further.
2. Background Concepts from Mathematics

In this section we will develop two mathematical structures, matrices and graphs, which are key to representing networks. Every graph has an associated adjacency matrix that tells how the vertices in the graph are connected. Once we have learned how to multiply matrices, we will see how calculations with the adjacency matrix may be used to determine and analyze pathways through the graph. This is an essential part of network analysis.

2.1 Matrices

Many kinds of numerical data such as payroll data and gradebook spreadsheets lend themselves to being stored in tables. A table with m rows and n columns of numbers is called a matrix of size \( m \times n \) (read as “m by n”).

Here are three ways to represent an \( m \times n \) matrix:

1. An upper case letter: \( A, B, C \), etc.
2. A representative element of the matrix enclosed in a bracket: \([a_{ij}], [b_{ij}], [c_{ij}], \) etc. The first subscript indicates the row where the number is located while the second number indicates the column where the number is located.

\[
\begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \cdots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
\]

3. A rectangular array of numbers: \( A = [a_{ij}] = 
\begin{bmatrix}
35 & 40 & 20 & 30 & 10 \\
35 & 25 & 25 & 35 & 30 \\
25 & 25 & 35 & 30 & 10
\end{bmatrix}
\]

Here are some examples of matrices and of matrix notation.

**Example 2.1.1** Let \( A = 
\begin{bmatrix}
35 & 40 & 20 & 30 & 10 \\
35 & 25 & 25 & 35 & 30 \\
25 & 25 & 35 & 30 & 10
\end{bmatrix}
\)

\( A \) is a 3 \times 5 matrix. Since the number 40 is located in row 1 and column 2 of \( A \), we write \( a_{1,2} = 40 \). As another example, \( a_{3,3} = 35 \) because the number that is located in row 3 and column 3 of the matrix is 30.
Activity 2.1 Let \( A = \begin{bmatrix} 35 & 40 & 20 & 30 & 10 \\ 35 & 25 & 25 & 35 & 30 \\ 25 & 25 & 35 & 30 & 10 \end{bmatrix} \).

Fill in the following blanks: \( a_{1,1} = \quad a_{2,4} = \quad a_{3,1} = \quad \)

Answer key for Activity 2.1

\( a_{1,1} = 35 \quad a_{2,4} = 35 \quad a_{3,1} = 25 \)

Example 2.1.2 Let \( B = \begin{bmatrix} 0 & 214 & 3465 & 5946 \\ 214 & 0 & 3631 & 6042 \\ 3469 & 3631 & 0 & 6749 \\ 5946 & 6042 & 6749 & 0 \end{bmatrix} \).

Then \( B \) is a 4 x 4 matrix. \( B \) exhibits a couple of interesting patterns. First, \( b_{i,i} = 0 \), for each \( i = 1, 2, 3, 4 \). Furthermore, matrix \( B \) exhibits symmetry. For example, \( a_{1,2} = a_{2,1} = 214 \) and \( a_{3,4} = a_{4,3} = 6749 \); more generally, \( a_{ij} = a_{ji} \) whenever \( i \neq j \). In fact, the numerical data in matrix \( B \) represent the air distances in miles between the pairs of cities indicated in the table below. The mileage was computed by using the online calculator at http://www.distancefromto.net/. The patterns we observe are due to these facts: the airline distance between any city and itself is 0 and the airline distance from city \( i \) to city \( j \) is the same as the distance from city \( j \) to city \( i \).

<table>
<thead>
<tr>
<th>From/To</th>
<th>London</th>
<th>Paris</th>
<th>NYC</th>
<th>Tokyo</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>0</td>
<td>214</td>
<td>3465</td>
<td>5946</td>
</tr>
<tr>
<td>Paris</td>
<td>214</td>
<td>0</td>
<td>3631</td>
<td>6042</td>
</tr>
<tr>
<td>NYC</td>
<td>3465</td>
<td>3631</td>
<td>0</td>
<td>6749</td>
</tr>
<tr>
<td>Tokyo</td>
<td>5946</td>
<td>6042</td>
<td>6749</td>
<td>0</td>
</tr>
</tbody>
</table>
2.2 The transpose of a matrix

To each \( m \times n \) matrix, \( A \), we can associate an \( n \times m \) matrix \( C = [c_{ij}] \) called the transpose of \( A \), where \( c_{ij} = a_{ji} \), for all \( i, 1 \leq i \leq n \), and for all \( j, 1 \leq j \leq m \). In other words, the transpose of \( A \) is the matrix obtained by switching the rows and columns of \( A \). The symbol for the transpose of \( A \) is \( A^T \).

To illustrate, for the \( 3 \times 5 \) matrix in Example 2.1.1, \( A = \begin{bmatrix} 2 & 2 & 3 & 4 & 5 \\ 0 & 5 & 6 & 7 & 8 \\ 4 & 1 & 2 & 3 & 4 \end{bmatrix}, \) then the transpose of \( A \) is \( A^T = \begin{bmatrix} 2 & 0 & 4 \\ 2 & 5 & 3 \\ 4 & 1 \\ 5 & 2 & 3 \end{bmatrix} \).

Activity 2.2 For each of the following matrices, find its transpose.

a. If \( A = \begin{bmatrix} 2 & -1 \\ 0 & \sqrt{5} \\ 4 & 1 \end{bmatrix} \), then \( A^T = \begin{bmatrix} 2 & 0 & 4 \\ 35 & 40 & 20 \\ 35 & 25 & 35 \end{bmatrix} \).

b. If \( B = \begin{bmatrix} 2 \\ -3 \\ 15 \\ 6 \end{bmatrix} \), then \( B^T = \begin{bmatrix} 2 & -3 & 15 & 6 \end{bmatrix} \).

Answer key for Activity 2.2

a. If \( A = \begin{bmatrix} 2 & -1 \\ 0 & \sqrt{5} \\ 4 & 1 \end{bmatrix} \), then \( A^T = \begin{bmatrix} 2 & 0 & 4 \\ 35 & 40 & 20 \\ 35 & 25 & 35 \end{bmatrix} \).

b. If \( B = \begin{bmatrix} 2 \\ -3 \\ 15 \\ 6 \end{bmatrix} \), then \( B^T = \begin{bmatrix} 2 & -3 & 15 & 6 \end{bmatrix} \).
2.3 Multiplication of matrices

Matrix multiplication has many real-world uses. We motivate the formal definition with an example. Suppose that a local lunch truck specializes in crepes and that it offers three kinds of crepes - smoked salmon, ham and cheese, and roasted vegetables - which are priced as follows:

<table>
<thead>
<tr>
<th>Crepe</th>
<th>Smoked salmon</th>
<th>Ham and cheese</th>
<th>Roasted vegetable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>$7.00</td>
<td>$5.00</td>
<td>$4.00</td>
</tr>
</tbody>
</table>

Also, suppose that for last week, customer demand for the crepes was as follows:

<table>
<thead>
<tr>
<th>Lunch crepes</th>
<th>Mon</th>
<th>Tues</th>
<th>Wed</th>
<th>Thurs</th>
<th>Fri</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoked salmon</td>
<td>35</td>
<td>40</td>
<td>20</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>Ham and cheese</td>
<td>35</td>
<td>25</td>
<td>25</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>Roasted vegetables</td>
<td>25</td>
<td>25</td>
<td>35</td>
<td>30</td>
<td>10</td>
</tr>
</tbody>
</table>

Then the value of the total sales (in dollars) for Monday is computed by summing up the Monday sales for each of the three kinds of crepes: 

\[(7 \times 35) + (5 \times 35) + (4 \times 25) = 520\] dollars. Note that we say that this product is found by "pouring" \[ \begin{bmatrix} 7 & 5 & 4 \end{bmatrix} \] down the Monday sales column.

Similarly, the value of the total sales (in dollars) for Tuesday is computed by summing up the Tuesday sales for each of the three kinds of crepes: 

\[(7 \times 40) + (5 \times 25) + (4 \times 25) = 505\] Again, this product is found by pouring \[ \begin{bmatrix} 7 & 5 & 4 \end{bmatrix} \] down the Tuesday sales column.

Continuing in like manner, the value of the total sales for Monday-Friday can be calculated by computing the product of the 1 x 3 crepe price matrix, \(P = \begin{bmatrix} 7 & 5 & 4 \end{bmatrix}\), and the 3 x 5 Monday-Friday demand matrix, \(D = \begin{bmatrix} 35 & 40 & 20 & 30 & 10 \\ 35 & 25 & 25 & 35 & 30 \\ 25 & 25 & 35 & 30 & 10 \end{bmatrix}\) as follows:

\[PD = \begin{bmatrix} 7 & 5 & 4 \\ 35 & 40 & 20 & 30 & 10 \\ 35 & 25 & 35 & 30 & 10 \end{bmatrix} = \begin{bmatrix} 520 & 505 & 405 & 505 & 260 \end{bmatrix}\]
Whenever the dimensions of two matrices, A and B, “match” appropriately, we can multiply them. Formally, if $A = [a_{ij}]$ is an $m \times n$ matrix and $B = [b_{ij}]$ is an $n \times p$ matrix, where $m$, $n$, $p$ are positive integers, then we can define the product of A and B, written as $AB$, to be the $m \times p$ matrix $C$, where $C = [c_{ij}]$ and $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$. We say that $c_{ij}$ is obtained by “pouring row i of matrix A down column j of B” or, equivalently, by taking the dot product of row i of matrix A and column j of matrix B. Here is a visualization for $c_{ij}$:

$$c_{ij} = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix} \begin{bmatrix}
b_{11} & b_{12} & b_{ij} & b_{1p} \\
b_{21} & b_{22} & b_{2j} & b_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
b_{n1} & b_{n2} & \cdots & b_{nj} & \cdots & b_{np}
\end{bmatrix}$$

**Activity 2.3** Find $AB$, if it is defined, for each of the following pairs of matrices, A and B:

a. If $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$, then $AB = \begin{bmatrix} \_ & \_ \end{bmatrix}$.

b. If $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$, then $AB = \begin{bmatrix} \_ \end{bmatrix}$.

c. If $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 6 & 7 \\ 8 & 9 & 0 \end{bmatrix}$, then $AB = \begin{bmatrix} \_ & \_ & \_ \end{bmatrix}$.

d. If $A = \begin{bmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$, then $AB = \begin{bmatrix} \_ & \_ & \_ \end{bmatrix}$.

**Answer key to Activity 2.3**

a. If $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$, then $AB = \begin{bmatrix} 5 & 11 \\ 11 & 25 \end{bmatrix}$.

b. If $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$, then $AB = \begin{bmatrix} 17 \\ 39 \end{bmatrix}$.

c. If $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 6 & 7 \\ 8 & 9 & 0 \end{bmatrix}$, then $AB = \begin{bmatrix} 21 & 24 & 7 \\ 47 & 54 & 21 \end{bmatrix}$.

10
d. If \(\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}\) and \(\mathbf{B} = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}\), then \(\mathbf{A}\mathbf{B} = [-1]\).

2.4 Graphs, vertices, and edges

A **graph** is an ordered pair \((V, E)\), where \(V\) is a set of **vertices** and \(E\) is a multiset of **edges** between the vertices in \(V\). Alternate names for vertices and edges are **nodes** and **links**, respectively.

**Example 2.4.1** In Figure 1 below, we see a graph with vertex set \(V = \{u, v, w, z\}\) and edge multiset \(E = \{\{u, u\}, \{u, w\}, \{v, w\}, \{v, w\}, \{w, z\}\}\). Note that there are two edges between \(v\) and \(w\) and that they are both listed in the multiset. Also, an edge that begins and ends at itself is called a **loop**.

![Figure 1. A graph with a repeated edge and a loop](https://en.wikipedia.org/wiki/Incidence_matrix#/media/File:Labeled_undirected_graph.svg)

**Example 2.4.2** The graph in Figure 2 is called a **simple graph** because it has no loops and no repeated edges. Its vertex set is \(V = \{1, 2, 3, 4\}\) and its edge set is \(E = \{e_1, e_2, e_3, e_4\}\).

![Figure 2. A simple graph](https://en.wikipedia.org/wiki/Incidence_matrix#/media/File:Labeled_undirected_graph.svg)
**Activity 2.4** The graph below is called the Petersen graph\(^i\). Find its vertex and edge sets.

![Petersen graph](http://www.rawbw.com/~davidm/zome/solvingPetersen.html)

Figure 3. The Petersen graph

Answer key for Activity 2.4

The vertex set for the Petersen graph is \( V = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \).

The edge set is \( E = \{\{0, 1\}, \{1, 2\}, \{2, 9\}, \{9, 7\}, \{7, 0\}, \{4, 3\}, \{3, 6\}, \{6, 5\}, \{5, 8\}, \{8, 4\}, \{0, 4\}, \{2, 3\}, \{8, 9\}, \{6, 7\}, \{4, 0\}\} \).

**2.5 The order and size of a graph**

By the **order of a graph** we mean the number of its vertices. The **size of a graph** is the number of its edges. To illustrate, the graph in Example 2.4.1 has order 4 and size 5. (A loop is counted as one edge.); also, the graph in Example 2.4.2 has order 4 and size 4.

**Activity 2.5** The graph in Figure 4 below is famous because it models the Königsberg Bridge problem\(^ii\) that sparked the beginning of the field of graph theory. Find the order and size of this famous graph.

![Königsberg Bridge Problem](https://mvngu.files.wordpress.com/2011/03/konigsberg.png)

Figure 4. The graph for the Königsberg Bridge Problem
Answer key to Activity 2.5

The graph has vertex set $V = \{a, b, c, d\}$. It has edge multiset $E = \{\{a, b\}, \{a, b\}, \{b, c\}, \{b, c\}, \{c, d\}, \{b, d\}, \{a, d\}\}$. Therefore, the order of the graph is 4. The size of the graph is 7.

2.6 The degree of a vertex

By the degree of a vertex we mean the number of edges that are incident with it. Also, by the degree distribution of a graph we mean a table that lists all the vertices of the graph along with their degrees. To illustrate, let's look again at the graph in Figure 2 (Example 2.4.2):

![Graph](http://blog.quecheeclub.com/Portals/290247/images/Cape%20Air%20to%20Quechee%20Vt.jpg)

There are three edges that are incident with vertex 1: $e_1$, $e_2$, and $e_3$. So, vertex 1 has degree 3. Also, there is one edge incident with vertex 2: $e_1$. Continuing this analysis, we can summarize our findings in Table 1, which gives the degree distribution table for the above graph:

<table>
<thead>
<tr>
<th>vertex</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>degree</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1. Degree distribution table for Figure 2

Activity 2.6 One of today's smallest airlines is Cape Air. The New England airports served by Cape Air are shown in Figure 5.

![Cape Air route map](http://blog.quecheeclub.com/Portals/290247/images/Cape%20Air%20to%20Quechee%20Vt.jpg)
If we restrict our attention to the Massachusetts airports served by Cape Air, represent each of these airports by a vertex, and draw an edge between any two airports connected by a Cape Air direct flight, then we get the graph in Figure 6 below.

Give the degree distribution for the Cape Air graph in Figure 6.

![Figure 6. Cape Air graph for Massachusetts](image)

**Answer key for Activity 2.6**

The degree distribution table for the Cape Air graph in Figure 6 is as follows:

<table>
<thead>
<tr>
<th>vertex</th>
<th>P</th>
<th>B</th>
<th>N</th>
<th>NB</th>
<th>MV</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>degree</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

**2.7 The adjacency matrix of a simple, undirected graph**

For the remainder of Section 2 we focus on graphs whose edges are bi-directional. These graphs are called **undirected graphs**. We will show how every undirected graph has an associated matrix called its **adjacency matrix** that tells how the nodes in the graph are connected.

To illustrate what we mean by an adjacency matrix, let's return to the graph in Figure 2 (Example 2.4.2), repeated below:
Since this graph has four vertices, its adjacency matrix, $A$, will be a 4 x 4 matrix. The rows and columns of this matrix will correspond to the four vertices 1, 2, 3, 4, respectively, as follows. First, because vertex 1 is connected to vertices 2, 3, and 4 but not to itself, then the first row of the adjacency matrix, will be: 0111. Next, because vertex 2 is connected only to vertex 1, then the second row of the adjacency matrix will be: 1000. Continuing in like manner, we complete the adjacency matrix and get:

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

Of particular note, the adjacency matrix is symmetric, because an edge from vertex $i$ to vertex $j$ is also an edge from vertex $j$ to vertex $i$, i.e., all of the edges are bidirectional. As well, the row totals for each of the rows 1, 2, 3, 4 give the degrees for vertices 1, 2, 3, 4, respectively.

To generalize, whenever $G = (V, E)$ is a simple, undirected graph, we associate with $G$ a matrix that is based upon the vertices in $V$ and the edges in $E$ called the adjacency matrix for $G$. Specifically, if the vertex set, $V$, for $G$ has $n$ vertices 1, 2, ..., $n$, then the adjacency matrix $A$ for $G$ is the $n \times n$ matrix,

$$A = [a_{ij}] = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{nn} \\
\end{bmatrix}$$

where $a_{ij} = 1$ if vertices $i$ and $j$ are joined by an edge, and $a_{ij} = 0$ otherwise.

**Activity 2.7** Give the adjacency matrix for the Cape Air graph (Figure 6) that is reproduced below:
2.8 Paths in a graph and how to use the adjacency matrix to count them

For any graph, a path from vertex $v_i$ to vertex $v_n$ is any sequence of vertices $v_1, v_2, ..., v_n$ where $v_i$ and $v_{i+1}$ are the endpoints of an edge in $G$. To illustrate, we return to the graph in Figure 2 (Example 2.4.2):

Because there are four edges in graph $G$, we say that $G$ has four paths that are each 1 edge long, or have length one. They are: 1,2; 1,3; 1,4; 3,4. In general, the length of a path between two vertices is the number of edges in that path.

Now, all the paths of length 1 can be "read off" the adjacency matrix, $A$, for any graph $G$. For example, the above graph $G$ (Figure 2), has adjacency matrix $A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$. From row one of matrix $A$, we see that vertex 1 is connected by an edge to vertices 2, 3, and 4; these edges represent three of the 4 paths of length 1 in $G$. From row 3, we see that vertex 3 is connected to vertex 4 by an edge, and this edge is the fourth path of length 1. So, the adjacency matrix, $A$, gives us a full accounting of all four paths of length 1 in graph $G$. (Recall that $A$ is symmetric, so we only need to use the top half of the matrix.)
Moving from paths of length one to paths of length 2, we observe that there are 13 paths of length 2 in the above graph G (Figure 2). They are: 1,2,1; 1,2,1; 1,3,1; 2,1,2; 3,1,3; 3,4,3; 4,1,4; 4,3,4; (these 8 paths are called cycles); 1,4,3; 1,3,4; 2,1,3; 2,1,4; 3,1,4.

Now, paths of length 2 in a graph G can be "read off" the matrix $A^2$ (i.e., the square of the adjacency matrix, $A$, or the matrix which is the product of $A$ with itself). Why? Let $B = A^2$. From the definition of the multiplication of matrices, the entries $b_{ij}$ of $B$ are obtained as follows: $b_{ij} = \sum_{k=1}^{n} a_{ik}a_{kj}$. So, $b_{ij} = 1$ if and only if there is a vertex $k$ with $a_{ik} = 1$ and $a_{kj} = 1$. But this is equivalent to saying that there is a path of length 2 from vertex $i$ to vertex $j$. So we can conclude that $b_{ij}$ is the number of paths of length 2 from $i$ to $j$. (If $i = j$, then we just get the degree of the vertex).

Double-check: If we return, again, to the graph G in Figure 2 (Example 2.4.2, shown above) whose adjacency matrix is $A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$, can we "read off" all the paths of length two from $A^2 = \begin{bmatrix} 3 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$? Yes. Let $B = A^2 = \begin{bmatrix} 3 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$. For example, $b_{1,1} = 3$ is the number of paths of length 2 from vertex 1 to vertex 1; those paths are: 1, 2, 1; 1, 3, 1; 1, 4, 1. As another example, $b_{1,3} = 1$ is the number of paths of length two from vertex 1 to vertex 3; the one path is 1,4,3. In this manner, we can check that when we add up the entries in the upper half of the matrix $A^2$, we can account for the 13 paths of length two which we found by inspection. ($A^2$, like $A$, is symmetric because graph G is undirected.)

We note that it may be shown using mathematical induction that, for any simple graph G with $n$ vertices $v_1, v_2, ..., v_n$ and adjacency matrix $A$, the number of paths of length $k$ in G from vertex $v_i$ to vertex $v_j$ is equal to $b_{ij}$, where $[b_{ij}] = B = A^k$, for any positive integer $k$.

**Activity 2.8** Let $G$ be the following graph:
a. Find all the paths of length 2 in G.
b. Find the adjacency matrix, $A$, associated with G.
c. Compute $A^2$ and explain which entry of $A^2$ corresponds to the paths you found in part a.

Answer key for Activity 2.8

a. The paths of length 2 are as follows (the highlights below are explained in part c):

1, 2, 1; 1, 5, 1; 2, 3, 2; 2, 4, 2; 2, 5, 2; 2, 1, 2; 2, 4, 3; 2, 3, 4; 2, 5, 4; 1, 5, 2; 1, 4, 5;
3, 2, 3; 3, 4, 3; 3, 2, 5; 3, 4, 5;
4, 2, 1; 4, 5, 1; 4, 3, 2; 4, 5, 2; 4, 2, 3; 4, 5, 3; 2, 4, 3; 2, 5, 3; 4, 5, 3;
5, 2, 1; 5, 4, 2; 5, 1, 2; 5, 2, 3; 5, 4, 3; 5, 2, 4; 5, 3, 1; 5, 2, 5; 5, 4, 5;

b. Find the adjacency matrix, $A$, associated with G.

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

c. Compute $A^2$ and explain which entry of $A^2$ corresponds to the path you found in part a.

$$A^2 = \begin{bmatrix}
2 & 1 & 1 & 2 & 1 \\
1 & 4 & 1 & 2 & 2 \\
1 & 1 & 2 & 1 & 2 \\
2 & 2 & 1 & 3 & 1 \\
1 & 2 & 2 & 1 & 3
\end{bmatrix}$$

Note: If we let $B = A^2$, then $b_{ij}$ is the number of paths of length 2 in graph $G$ from vertex $i$ to vertex $j$. In part a, we list all the paths that begin at vertex 1, then those that begin at vertex 2, etc. Also, for all the paths that begin with vertex 1, we first list all
the paths from vertex 1 to vertex 1, then the paths from vertex 1 to vertex 2 etc., using highlighting to indicate the beginning of the next group of paths.

Here is how the entries of $B = A^2$ correspond to the paths we listed in part a:

- $b_{1,1} = 2$ corresponds to 2 paths of length 2 from vertex 1 to vertex 1: 1, 2, 1; 1, 5, 1
- $b_{1,2} = 1$ corresponds to 1 path of length 2 from vertex 1 to vertex 2: 1, 5, 2
- $b_{1,3} = 1$ corresponds to 1 path of length 2 from vertex 1 to vertex 3: 1, 2, 3
- $b_{1,4} = 2$ corresponds to 2 paths of length 2 from vertex 1 to vertex 4: 1, 2, 4; 1, 5, 4
- $b_{1,5} = 1$ corresponds to 1 path of length 2 from vertex 1 to vertex 5: 1, 2, 5

### 2.9 Diameter of a graph

The **diameter of a graph** is the maximum length of all shortest paths between two nodes (i.e., the longest graph geodesic). To illustrate, consider the graph in Figure 7 (repeated below):

![Graph](image)

We see that:

- the shortest path from vertex 1 to vertex 2 is: 1, 2
- the shortest path from vertex 1 to vertex 3 is: 1, 2, 3
- the shortest path from vertex 1 to vertex 4 is: 1, 2, 4 or 1, 5, 4
- the shortest path from vertex 1 to vertex 5 is: 1, 5
- the shortest path from vertex 2 to vertex 3 is: 2, 3
- the shortest path from vertex 2 to vertex 4 is: 2, 4
- the shortest path from vertex 2 to vertex 5 is: 2, 5
- the shortest path from vertex 3 to vertex 4 is: 3, 4
- the shortest path from vertex 3 to vertex 5 is: 3, 4, 5 or 3, 2, 5
- the shortest path from vertex 4 to vertex 5 is: 4, 5

Thus, the maximum length of all of the shortest paths is 2; so, the diameter of the graph is 2.

**Activity 2.9** Find the diameter for the following Cape Air graph, repeated from Figure 6:
Answer key for Activity 2.

We begin by listing the shortest paths between each pair of distinct vertices in the Cape Air graph:

- the shortest path from P to B is: P, B (length of the path: 1)
- the shortest path from P to N is: P, B, N (length of the path: 2)
- the shortest path from P to NB is: P, B, NB (length of the path: 2)
- the shortest path from P to MV is: P, B, MV (length of the path: 2)
- the shortest path from P to H is: P, B, H (length of the path: 2)
- the shortest path from B to N is: B, N (length of the path: 1)
- the shortest path from B to NB is: B, NB (length of the path: 1)
- the shortest path from B to MV is: B, MV (length of the path: 1)
- the shortest path from B to H is: B, H (length of the path: 1)
- the shortest path from N to NB is: N, NB (length of the path: 1)
- the shortest path from N to MV is: N, MV (length of the path: 1)
- the shortest path from N to H is: N, H (length of the path: 1)
- the shortest path from NB to MV is: NB, MV (length of the path: 1)
- the shortest path from NB to H is: NB, MV, H (length of the path: 2)
- the shortest path from MV to H is: MV, H (length of the path: 1)

Since the maximum length of the above shortest paths is 2, then the diameter of the Cape Air graph is 2.
3. Background Concepts from Network Analysis

Social network usage is huge and growing. As may be seen in the Statista bar graph below, the number of social network users across the globe was .97 billion in 2010, and it is projected to grow to 2.44 billion in 2018.

![Growth of social network users](http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/)

Figure 8. Growth of social network users

Among American online users, a Pew Research report that was published in January 2015 noted these features:

- 74% of online adults use social networking sites
- 71% of online adults use Facebook
- Multi-platform use is growing: 52% of online adults now use two or more social media sites, up from 42% in 2013
- A first: About half of internet-using young adults ages 18-29 (53%) use Instagram
- The portion of internet users with college educations using LinkedIn hit 50%

![Social Media Site Usage in 2014](http://www.pewinternet.org/2015/01/09/social-media-update-2014/pi_15-01-09_socialmediaupdate_260x260/)

Figure 9. Social media usage
One social media trend of particular note is that the portion of americans getting their news from Twitter or Facebook is rising:

![Facebook and Twitter News Use is on the Rise](http://www.journalism.org/2015/07/14/the-evolving-role-of-news-on-twitter-and-facebook/)

As these data show, social media platforms have greatly penetrated our lives. In the next section, we will show how analyzing social networks can enhance the quality of our lives.

### 3.1 What is a network?

A **network** is any interconnected system of things (people or objects). A **graph** is a mathematical model for representing the network. In the graph of a network, the vertices represent the actors in the network and the edges drawn between vertices signify that there is a connection between the actors. In network analysis, the terms **nodes** and **links** are used to correspond to vertices and edges, respectively.

**Example 3.1.1 The Internet**

The most vast among communications networks is the Internet: It is the network of all networks. Its nodes are computers or other devices and its edges are the physical or wireless connections between them.
Example 3.1.2 Cape Air Routes in New England and upstate New York

Transportation systems offer rich examples of networks. As we saw in Section 2, Cape Air is a small, U.S. airline, that serves New England and upstate New York. In the route system below, nodes represent the cities served by Cape Air and edges denote airline routes between the cities.

Beyond its applications to communications and transportation networks, network analysis is presently being used by biologists to study human disease and genetic disorders. Figure 13 shows the Disease Gene Network (DGN) published in 2008 by the National Academy of Sciences.

"In the DGN, each node is a gene, with two genes being connected if they are implicated in the same disorder. The size of each node is proportional to the number of disorders in which the gene is implicated. Nodes representing genes with links to multiple classes are
colored dark grey, whereas unclassified genes are colored light grey. Genes associated with more than five disorders, and those mentioned in the text, are indicated with the gene symbol. Only nodes with at least one link are shown.iii

Figure 13. The Disease Gene Network
http://www.nature.com/scitable/resource?action=showFullImageForTopic&imgSrc=/scitable/content/6437/10[1].1073_pnas0701361104-fig2b_full.jpg

3.2 What is social network analysis?

Valdis Krebs, a leader and expert in social and organizational network analysis and "founder and chief scientist of OrgNet, LLCiv, defines **Social Network Analysis (SNA)** as

> ... the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships.v

Now, as never before, networks are being used to represent human relationships in order to model human behavior. Citing the work of Valdis Krebs among others, a recent article in Forbes asserts that, "Whereas before, human behavior was a domain dominated by liberal arts majors, its study is now being revolutionized by mathematicians."vi

Example 3.2.1 Zachary's Karate Clubvii
A classic example of a human network is Zachary's Karate Club which comes from a sociology study. In Figure 14 below, the left-hand graph is the network of friendships between 34 members of a karate club run by Wayne Zachary at an American university in the 1970s. The interactions between members were observed over a two-year period. When a strong disagreement between the club's instructor and administrator arose, the club split into two clubs, as shown in the red and blue coloring in the right-hand graph in Figure 14. The use of social network analysis and visualization techniques helps us to understand, after the fact, why the members split the way they did. In the right-hand figure below, we can see that the members of the club formed two clusters: friends of person 1 (the instructor) and the friends of person 34 (the club's administrator). The members of these clusters had friendships almost exclusively with each other and led to the formation of two new clubs after the break-up.

![Networks](https://networkdata.ics.uci.edu/data.php?id=105)

**Figure 14. Zachary's Karate Club**

**Example 3.2.2 Tracing contacts to detect the spread of a contagion**

Early in 2014, U.S. news outlets began reporting on the outbreak in West Africa of a mysterious and lethal disease, Ebola. The map below, in Figure 15, shows the countries in West Africa that were affected and the total number of cases, as of August 12, 2015, in each of the color-coded regions.
Responding to the Ebola crisis, epidemiologists from organizations including the World Health Organization, the U.S. Center for Disease Control and Prevention, and Doctors Without Borders were dispatched to West Africa to interview the family members and friends who were caring for the victims. Their goal was to understand how the disease was being transmitted in order to slow, and eventually stop, its progress.

Figure 15. Total Ebola cases by August 2015

Figure 16 below displays a partial contact map for the Ebola outbreak. The vertices (i.e., human figures) represent patients who had succumbed to Ebola and edges indicate that there has been physical contact between individuals.

Figure 16. Partial contact map for Ebola outbreak
According to the CDC, as of August 16, 2015, the total cases (Suspected, Probable, and Confirmed) stood at 27,988; of these, 15,220 were laboratory-confirmed. The total number of deaths as of this date was 11,299.

For a timeline of the Ebola outbreak and of the international response to this crisis, read the New York Times article, "How Ebola Roared Back".

Activity 3.1 Ego networks
An ego network is a network whose vertices consist of an individual (the ego), all of the individuals to whom they are tied (the alters), and the ties among these people. The diagram below illustrates what an ego network looks like. This diagram is due to Steve Borgatti, who is Professor and Endowed Chair of the Department of Management at the Gatton College of Business and Economics of the University of Kentucky and a prominent researcher and scholar in the area of social network analysis.

Of course, a single human being has different kinds of ties with other human beings. For example, there are similarity ties such as 'loves the same sport' or 'likes the same music', there are social relationship ties such as kinship and friendship, and there are interaction ties such as 'spends time with' or 'exchanges text message'. Therefore, when designing an ego network, it is necessary to clearly specify what the tie in the ego network represents.

Your assignment is to develop your ego network based on data from the last 24 hours. Follow these steps:

a. Write down the first names of all the people with whom you interacted within the last 24 hours by sending or receiving an Instagram, text message, or email; these individuals are your "alters". For the purposes of this activity we will only track these 3 forms of interaction. (In case you have two friends named Mary, modify their names by appending the first letter of their last name to their first name; e.g., Mary Smith becomes MaryS.)

b. Beside each name, write down the total number of interactions you had with each of your alters (e.g., if you exchanged 2 Instagrams, 1 text message, and 0 emails with Mary, then the total number of your interactions with Mary is 3).
c. Contact each of your alters, and share with them the list of your alters. Ask them to send you the total number of interactions (Instagrams/text messages/emails) they had with each of your other alters during the same 24-hour period you are using.
d. Draw a network where the vertices represent you (the ego) and your alters, and where you draw an edge between any pair of your alters who interacted by exchanging Instagrams/text messages/emails.
e. Label each of the edges in your network with the total number of interactions which took place between the individuals in your network
f. According to American sociologist and Stanford University professor Mark Granovetter, the stronger the tie between EGO and two of her alters, the greater the likelihood that the alters enjoy at least a weak tie. Does your ego network support Granovetter's hypothesis? Explain your answer.

Answer key to Activity 3.1

Answers to this activity will vary.

3.3 Who is really in charge? Centrality as power in a network

In real estate there is an adage that the worth of a property is more than just the physical property itself; it's a matter of location, location, location.

Power is an essential property of social structures; and, finding the key players in a network, i.e., the actors who wield the most power, is at the heart of social network analysis. Sociologists Robert Hanneman and Mark Riddle note that the power of the actors in a network correlates with how they are situated in the network, i.e., correlates with "how close they are to the "center" of the action in a network. ... Network analysts are more likely to describe their approaches as descriptions of centrality than of power."

We now examine three measures of centrality: degree, betweenness, and closeness. We look at these measures in relation to three simple network topologies, Star, Circle, and Line, which are covered by sociologists Robert A. Hanneman and Mark Riddle in their textbook Introduction to Social Network Methods. We follow their treatment.

Figure 18. Various network topologies
3.4 Degree centrality and network topology

In Section 2 we defined the degree of a node (in an undirected graph) as the number of edges that are incident with it. The most basic measure of centrality of a node is degree centrality.

A node has degree centrality in a network if it is the node of maximum degree in the network. To illustrate degree centrality, we look at the position of actor A in the "Star", "Circle", and "Line" networks. By inspection, we can see from its location in "Star" that A has the most ties, so its position is most favored. In “Circle”, all the actors have the same number of ties, so all the nodes are equally advantaged (or disadvantaged). In "Line", A and G, which are both located at the end of the line, have fewer ties than the other nodes (which all have the same degree). In a network where an edge signifies information exchange or sharing, they are less favored than the other nodes.

The normalized degree centrality measure of a node $u$ (undirected network) is defined as the degree of $u$ divided by $N-1$, where $N$ is the number of nodes in the network.

Example 3.4.1 Consider the small, kite-shaped graph in Figure 19 which has 5 nodes. Since node a is connected to only 1 node, its normalized degree centrality measure is 1/4. Node b, on the other hand, is connected to 3 nodes; so its normalized degree centrality measure is 3/4. Further, nodes c, d, and e each have degree 2 and therefore have normalized degree centrality measure is 2/4. Therefore, node b has degree centrality for this graph.

![Figure 19. A kite-shaped graph](http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/)

Activity 3.2 The graph on the cover page, repeated below in Figure 19, is the well-known Krackhardt kite graph which was developed by David Krackhardt, a noted researcher in social networks.
Find the normalized degree centrality measure for each of the nodes in the Krackhardt kite graph. Determine which node has normalized degree centrality.

Answer key to Activity 3.2

Here are a couple of hints:
- The number of nodes is $N = 10$.
- Since Andre is connected to 4 nodes (Carol, Fernando, Diane, Beverly), his degree is 4; so the normalized degree centrality measure for Andre is $4/9 = 0.44$.

Here is a table showing the normalized degree centrality measures for all ten nodes.

<table>
<thead>
<tr>
<th>Node</th>
<th>Normalized degree centrality measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diane</td>
<td>0.67</td>
</tr>
<tr>
<td>Fernando</td>
<td>0.56</td>
</tr>
<tr>
<td>Garth</td>
<td>0.56</td>
</tr>
<tr>
<td>Andre</td>
<td>0.44</td>
</tr>
<tr>
<td>Beverly</td>
<td>0.44</td>
</tr>
<tr>
<td>Carol</td>
<td>0.33</td>
</tr>
<tr>
<td>Ed</td>
<td>0.33</td>
</tr>
<tr>
<td>Heather</td>
<td>0.33</td>
</tr>
<tr>
<td>Ike</td>
<td>0.22</td>
</tr>
<tr>
<td>Jane</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Diane has degree centrality.
3.5 Betweenness centrality and network topology

The next metric we examine for measuring node centrality is a metric called betweenness. The **betweenness of a node** is the number of shortest paths that pass through the node divided by the number of all shortest paths in the network.

Again, we look at the position of actor A in the "Star", "Circle", and "Line" networks to see how favorably it is situated in these networks relative to its betweenness.

In “Star”, A lies between each pair of actors. Every shortest path in the network must pass through A. Therefore, A is in the most favored position. In “Circle”, all actors are equally favored. In “Line”, the end actors A and G do not lie between any pairs; actors closer to the middle of the line lie on many paths among pairs and so are in a more advantaged position than A or G.

For most other networks, we use the following formula for calculating betweenness: The **normalized betweenness centrality of a node n**, denoted $C_b(n)$, is computed as follows:

- First compute $\sum_{s \neq t \neq n} (\sigma_{st}(n) / \sigma_{st})$, where
  - $s$ and $t$ are nodes in the network different from $n$
  - $\sigma_{st}$ denotes the number of shortest paths from $s$ to $t$
  - $\sigma_{st}(n)$ is the number of shortest paths from $s$ to $t$ that $n$ lies on
- To find $C_b(n)$, divide $\sum_{s \neq t \neq n} (\sigma_{st}(n) / \sigma_{st})$ by the number of node pairs excluding $n$
  (observe that the number of node pairs excluding $n$ is $(N-1)(N-2)/2$, where $N$ is the number of nodes in the network).

**Example 3.5.1** Let’s find the betweenness centrality of node b in the network below (Figure 19); this network, and the calculation of the betweenness centrality of node b are from [http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/](http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/).
We have:

\[ C_b(b) = \]
\[ (\sigma_{ac}(b) / \sigma_{ac}) + (\sigma_{ad}(b) / \sigma_{ad}) + (\sigma_{ae}(b) / \sigma_{ae}) + (\sigma_{cd}(b) / \sigma_{cd}) + (\sigma_{ce}(b) / \sigma_{ce}) + \]
\[ (\sigma_{de}(b) / \sigma_{de}) / 6 \]
\[ = ((1 / 1) + (1 / 1) + (2 / 2) + (1 / 2) + (0 / 1) + (0 / 1)) / 6 = 3.5 / 6 = 0.583^{\text{xiii}} \]

**Activity 3.3** For the network in Figure 19, calculate the betweenness centrality for nodes a, c, d, and e. Determine which of the five nodes in the network has the highest betweenness centrality.

**Answer key to Activity 3.3**

<table>
<thead>
<tr>
<th>Node</th>
<th>Normalized betweenness centrality measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>( C_b(a) = \sum_{x \neq a \neq t} (\sigma_{st}(a) / \sigma_{st}) / 6 )</td>
</tr>
<tr>
<td></td>
<td>= ((\sigma_{ba}(a) / \sigma_{ba}) + (\sigma_{bd}(a) / \sigma_{bd}) + (\sigma_{be}(a) / \sigma_{be}) + (\sigma_{cd}(a) / \sigma_{cd}) + (\sigma_{ce}(a) / \sigma_{ce}) + (\sigma_{de}(a) / \sigma_{de})) / 6 )</td>
</tr>
<tr>
<td></td>
<td>= 0/1 + 0/1 + 0/2 + 0/2 + 0/1 + 0/1 = 0</td>
</tr>
<tr>
<td>b</td>
<td>( C_b(b) = \sum_{x \neq b \neq t} (\sigma_{st}(b) / \sigma_{st}) / 6 )</td>
</tr>
<tr>
<td></td>
<td>( \approx 0.583 ) (details above)</td>
</tr>
<tr>
<td>c</td>
<td>( C_b(c) = \sum_{x \neq c \neq t} (\sigma_{st}(c) / \sigma_{st}) / 6 )</td>
</tr>
<tr>
<td></td>
<td>= ((\sigma_{ab}(c) / \sigma_{ab}) + (\sigma_{ac}(c) / \sigma_{ac}) + (\sigma_{ad}(c) / \sigma_{ad}) + (\sigma_{bd}(c) / \sigma_{bd}) + (\sigma_{bc}(c) / \sigma_{bc}) + (\sigma_{de}(c) / \sigma_{de})) / 6 )</td>
</tr>
<tr>
<td></td>
<td>= ((0 / 1) + (0 / 1) + (1 / 2) + (0 / 1) + (1 / 2) + (0/1)) / 6 = 1 / 6 \approx 0.17</td>
</tr>
<tr>
<td>d</td>
<td>( C_b(d) = \sum_{x \neq d \neq t} (\sigma_{st}(d) / \sigma_{st}) / 6 )</td>
</tr>
<tr>
<td></td>
<td>= ((\sigma_{ab}(d) / \sigma_{ab}) + (\sigma_{ac}(d) / \sigma_{ac}) + (\sigma_{ad}(d) / \sigma_{ad}) + (\sigma_{bd}(d) / \sigma_{bd}) + (\sigma_{bc}(d) / \sigma_{bc}) + (\sigma_{ce}(d) / \sigma_{ce})) / 6 )</td>
</tr>
<tr>
<td></td>
<td>= ((0 / 1) + (0 / 1) + (1 / 2) + (0 / 1) + (1 / 2) + (0/1)) / 6 = 1 / 6 \approx 0.17</td>
</tr>
<tr>
<td>e</td>
<td>( C_b(e) = \sum_{x \neq e \neq t} (\sigma_{st}(e) / \sigma_{st}) / 6 )</td>
</tr>
<tr>
<td></td>
<td>= ((\sigma_{ab}(e) / \sigma_{ab}) + (\sigma_{ac}(e) / \sigma_{ac}) + (\sigma_{ad}(e) / \sigma_{ad}) + (\sigma_{bd}(e) / \sigma_{bd}) + (\sigma_{bc}(e) / \sigma_{bc}) + (\sigma_{de}(e) / \sigma_{de})) / 6 )</td>
</tr>
<tr>
<td></td>
<td>= ((0 / 1) + (0 / 1) + (0 / 1) + (0 / 1) + (0/1) + (1/2)) / 6 = 1/12 \approx 0.08</td>
</tr>
</tbody>
</table>

We conclude that node b has betweenness centrality.
3.6 Closeness centrality and network topology

One commonly used definition of the closeness of a node is that it is the average number of “hops” from that node to all the other nodes in the network. So, the node which has closeness centrality will have the smallest average number of hops from that node to all other nodes.

Again, we look at the position of actor A in the "Star", "Circle", and "Line" networks to see how favorably it is situated in these networks relative to its closeness. In “Star”, A is “closer” to more actors than any other actor: its geodesic distance (length of shortest path) is 1 from all other actors; the shortest distance between other pairs of actors is 2. So, the location of A gives it an advantage over the other nodes. In “Circle”, all actors have the same distribution of geodesic distances from the others. In “Line”, D is closer to all the other actors than are the set {C,E}, {B,F}, and the set {A, G}. A and G, being at the end of the line, are at a disadvantage when compared to the other nodes.

Once again, the simplicity of "Star", "Circle", and "Line" network topologies enabled us, by eye, to draw our conclusions about the closeness of node A.

We note that when we apply the commonly used definition of closeness, then the node which has closeness centrality will have the smallest average number of hops from that node to all other nodes. However, when we computed degree centrality and betweenness centrality, our goal was to find the node(s) having the highest centrality measure. For consistency with our previous focus on finding nodes having highest centrality measures, we modify our formula for calculating the closeness of a node by taking a reciprocal. Specifically, we define the closeness centrality of a node, denoted $C_c(n)$, as the reciprocal of the average shortest path length. It is computed as follows:

$$C_c(n) = \frac{1}{L(n,m)}$$

where $L(n,m)$ is the length of the shortest path between two nodes $n$ and $m$.

The closeness centrality of a node is a number between 0 and 1.

To illustrate, let’s compute the closeness centrality of node b in Figure 19 (Activity 3.2), shown below; this network and the calculation of the closeness centrality of node b are from http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/.
To compute the average shortest path length for node b, we begin with the path lengths between b and all other nodes:

\[ L(b,a) + L(b,c) + L(b,d) + L(b,e) = 1 + 1 + 1 + 2 = 5. \]

Then we calculate the average of these by dividing by 4:

\[ [L(b,a) + L(b,c) + L(b,d) + L(b,e)]/4 = 5/4 = 1.25 \]

Finally, to calculate the closeness centrality of node b, we take the reciprocal:

\[ C_c(b) = \frac{1}{[(L(b,a) + L(b,c) + L(b,d) + L(b,e))/4]} = 1/1.25 = 0.8^{xiv} \]

**Activity 3.4**

For the network in Figure 19, calculate the closeness centrality for nodes a, c, d, and e. Determine which of the five nodes in the network has the highest closeness centrality.

**Answer key to Activity 3.4**

<table>
<thead>
<tr>
<th>Node</th>
<th>Closeness centrality measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>[ C_c(a) = 1/[(L(a,b) + L(a,c) + L(a,d) + L(a,e))/4] = 4/(1+2+2+3) = 0.5 ]</td>
</tr>
<tr>
<td>b</td>
<td>[ C_c(b) = 0.8 ] (details are shown above)</td>
</tr>
<tr>
<td>c</td>
<td>[ C_c(c) = 1/[(L(c,a) + L(c,b) + L(c,d) + L(c,e))/4] = 4/(2+1+2+1) = 0.67 ]</td>
</tr>
<tr>
<td>d</td>
<td>[ C_c(d) = 1/[(L(d,a) + L(d,b) + L(d,c) + L(de))/4] = 4/(2+1+2+1) = 0.67 ]</td>
</tr>
<tr>
<td>e</td>
<td>[ C_c(e) = 1/[(L(e,a) + L(e,b) + L(e,c) + L(e,d))/4] = 4/(3+2+1+1) = 0.57 ]</td>
</tr>
</tbody>
</table>

We conclude that node b has closeness centrality.

**Quiz:** For the Krackhardt kite graph shown on the cover of this module, you found the normalized degree centrality measures for each of its ten nodes in Activity 3.2. Now determine the betweenness and closeness centrality measures for each of the ten nodes. Which node has betweenness centrality? Which node has closeness centrality?

**Answer key for Quiz**

A table with the specified network centrality measures may be found at [http://www.orgnet.com/sna.html](http://www.orgnet.com/sna.html). The node which has betweenness centrality is Heather. The nodes with closeness centrality are Fernando and Garth.
4. Gephi Basics and Activity

In this section we will introduce the open software Gephi using a small network representing shared sustainability interests. Possible student survey questions are provided. We will demonstrate the steps to take in Gephi and the results that can be achieved using a short example survey with five participants. Please note that this module does not provide a complete, self-contained tutorial on Gephi nor on the installation of Gephi. For teachers and students who wish to learn more about Gephi, we refer you to these references: the Github website [http://gephi.github.io/users/download/](http://gephi.github.io/users/download/) for information on how to download Gephi, the Gephi forum at [https://forum.gephi.org/viewtopic.php?f=3&t=3580](https://forum.gephi.org/viewtopic.php?f=3&t=3580) for details on installing Gephi, and the Github tutorial at [http://gephi.github.io/users/quick-start/](http://gephi.github.io/users/quick-start/) for how to use Gephi.

4.1 Sustainability survey

Possible sustainability questions are included in Table 2. We recommend you choose eight to twelve of them for your survey. In order to keep the data manageable, do not use too many questions. We suggest that for a larger group fewer questions should be used.

The survey should include some questions that many participants will answer in the negative. If all participants say yes to all questions, the graph representing your network will be a complete graph, meaning every node is directly connected to every other node. The network measures we will consider will be rather boring for a complete graph!

Questions number two, twenty-five, and twenty-six do not ask about specific sustainability activities, but they do indicate a willingness to share activities with others. If someone wished to start a new sustainability project of organization on campus, targeting individuals who are likely to spread the work could be useful. Someone marketing “green” products would also be interested in such individuals.

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Do you recycle paper, plastic, and aluminum products?</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Do you regularly do volunteer work with your community?</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Is your main vehicle an inherently low emissions vehicle?</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Do you turn off the lights when you leave a room?</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Are your showers less than five minutes long?</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Do you turn off the water when you brush your teeth?</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Are all of the light bulbs in your house low energy bulbs?</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Do you turn off your television/computer/sound system at the wall?</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Do you keep your thermostat at 68 or below in the winter and at 78 degrees or above in the summer?</td>
<td></td>
</tr>
</tbody>
</table>
10. Do you frequently walk or bike instead of drive?
11. Do you use a bus, train, or carpool whenever possible?
12. Do you reuse things instead of buying new?
13. Do you regularly use reusable shopping bags?
14. Do you compost food waste?
15. Do you usually carry a reusable beverage container?
16. Do you grow your own food or buy locally grown food?
17. Do you have a wildlife friendly garden?
18. Do you reuse the plastic bags stores give you?
19. Are all of your appliances rated as energy efficient?
20. Do you use a drying rack for clothes instead of a dryer?
21. Do you use plant-based cleaning supplies?
22. Do you recycle ink cartridges, CDs, DVDs, and CFLs?
23. Do you use refillable pens?
24. Do you use 100% post-consumer recycled content paper and notebooks?
25. Are you a member of the student government association?
26. Are you a member of the campus hiking or rock-climbing club?

<table>
<thead>
<tr>
<th>Table 2. Possible survey questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. Do you frequently walk or bike instead of drive?</td>
</tr>
<tr>
<td>11. Do you use a bus, train, or carpool whenever possible?</td>
</tr>
<tr>
<td>12. Do you reuse things instead of buying new?</td>
</tr>
<tr>
<td>13. Do you regularly use reusable shopping bags?</td>
</tr>
<tr>
<td>14. Do you compost food waste?</td>
</tr>
<tr>
<td>15. Do you usually carry a reusable beverage container?</td>
</tr>
<tr>
<td>16. Do you grow your own food or buy locally grown food?</td>
</tr>
<tr>
<td>17. Do you have a wildlife friendly garden?</td>
</tr>
<tr>
<td>18. Do you reuse the plastic bags stores give you?</td>
</tr>
<tr>
<td>19. Are all of your appliances rated as energy efficient?</td>
</tr>
<tr>
<td>20. Do you use a drying rack for clothes instead of a dryer?</td>
</tr>
<tr>
<td>21. Do you use plant-based cleaning supplies?</td>
</tr>
<tr>
<td>22. Do you recycle ink cartridges, CDs, DVDs, and CFLs?</td>
</tr>
<tr>
<td>23. Do you use refillable pens?</td>
</tr>
<tr>
<td>24. Do you use 100% post-consumer recycled content paper and notebooks?</td>
</tr>
<tr>
<td>25. Are you a member of the student government association?</td>
</tr>
<tr>
<td>26. Are you a member of the campus hiking or rock-climbing club?</td>
</tr>
</tbody>
</table>

Several of the possible survey questions were inspired by Brown University’s Green Room Packing Supplies list. If you wish to see the full list in order to be a bit more green yourself go to http://www.brown.edu/initiatives/brown-is-green/sites/brown.edu.initiatives.brown-is-green/files/uploads/Brown%20Green%20Room%20Supplies%202012.pdf.

We chose twelve of these questions for a sample sustainability survey of five theoretical individuals. The survey is shown in Table 3.

Note that we have included a line for a survey identification code. Using a code will allow labeling nodes of the graph that will eventually be produced without identifying specific individuals by name.

**Sustainability Survey**

Survey identification code: __________

Name: ___________________________________________
<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Do you recycle paper, plastic, and aluminum products?</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Do you regularly do volunteer work with your community?</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Do you turn off the lights when you leave a room?</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Are your showers less than five minutes long?</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Do you turn off the water when you brush your teeth?</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Do you turn off your television/computer/sound system at the wall?</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Do you keep your thermostat at 68 or below in the winter and at 78 degrees or above in the summer?</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Do you regularly use reusable shopping bags?</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Do you usually carry a reusable beverage container?</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Do you grow your own food or buy locally grown food?</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Do you use a drying rack for clothes instead of a dryer?</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Do you use 100% post-consumer recycled content paper and notebooks?</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Sample survey

4.2 Recording survey data

In order to generate the graph to illustrate the network, we need Excel or another spreadsheet program, Notepad or another text editor, and of course the network visualization software Gephi. Gephi can be downloaded at [http://gephi.github.io/](http://gephi.github.io/). For our example we have the data in Table 4.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack recycles, turns electronics off at the wall, keeps his home cool in winter and warm in summer, and uses 100% post-consumer recycled content paper.</td>
<td></td>
</tr>
<tr>
<td>Jan recycles, volunteers, turns off lights, water, and electronics, carries reusable bags, and eats locally grown food.</td>
<td></td>
</tr>
<tr>
<td>Jessie turns off lights, takes short showers, carries reusable beverage containers, and uses recycled paper.</td>
<td></td>
</tr>
<tr>
<td>Jill recycles and carries reusable bags.</td>
<td></td>
</tr>
<tr>
<td>Jim volunteers, turns off the water, adjusts the thermostat, and carries both reusable bags and beverage containers.</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Raw data

Clearly we need to organize the data. To prepare for this, create some one-word abbreviations for each sustainability practice. We used recycles, volunteers, lights, showers, water, electronics,
thermostat, bags, containers, food, rack, and paper. Each participant’s name or survey identification code should also be recorded as a single string, since Gephi reads two words separated by a space as two separate nodes. Put the information into a spreadsheet in one of the following two ways.

1. The first possibility is to work down the first column using one cell for each individual and his or her activities and using semi-colons for separation. For example, the first cell of the first row would be: Jack;recycles;electronics;thermostat;paper. The information for the next individual would go into the first cell of the second row, and we continue through the fifth row producing Table 5.

<table>
<thead>
<tr>
<th>Jack</th>
<th>recycles</th>
<th>electronics</th>
<th>thermostat</th>
<th>paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>recycles</td>
<td>volunteers</td>
<td>lights;water;electronics;bags;food</td>
<td></td>
</tr>
<tr>
<td>Jessie</td>
<td>lights;showers</td>
<td>containers;paper</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jill</td>
<td>recycles</td>
<td>bags</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jim</td>
<td>volunteers;water;thermostat</td>
<td>bags;containers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Data in a single column

2. Alternatively, put Jack in the first cell of the first row and put each of his activities in subsequent cells. Put Jan and her activities in the second row and continue until producing Table 6.

<table>
<thead>
<tr>
<th>Jack</th>
<th>recycles</th>
<th>electronics</th>
<th>thermostat</th>
<th>paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>recycles</td>
<td>volunteers</td>
<td>lights;water;electronics</td>
<td>bags</td>
</tr>
<tr>
<td>Jessie</td>
<td>lights;showers</td>
<td>containers</td>
<td>paper</td>
<td></td>
</tr>
<tr>
<td>Jill</td>
<td>recycles</td>
<td>bags</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jim</td>
<td>volunteers</td>
<td>water</td>
<td>thermostat</td>
<td>bags;containers</td>
</tr>
</tbody>
</table>

Table 6. Data in multiple columns

### 4.3 Creating a graph in Gephi

After entering all the data in one of the above ways, use procedure 4.3.1 to create a graph.

**Procedure 4.3.1: How to create a graph with both participants and activities as nodes**

1. Save the spreadsheet created as a comma delimited (CSV) file format.
2. Open Gephi.
3. Click file, new project, and file open.
4. Browse your computer files and choose your CSV file.
5. Click open, and choose “undirected” as the graph type in the pop-up box.
6. Click OK, and a graph should appear.
7. To label the nodes, click the button with the darkened T to the left and below the graph.
8. To make the graph appear larger, click on the “layout” box on the far left, choose “expansion” from the drop down menu, and then click “run” a few times.
9. To make the nodes of the graph appear more uniformly spaced, click on “layout,” choose “FruchtermanReingold,” click “run,” and then click “stop.” You may need to run “expansion” again afterwards.

You should have a graph similar to Figure21.

Figure21. A first graph

We now have a graph connecting individuals to their sustainable activities. We may look at an individual and see all the activities he or she does, or you may look at an activity and see all the individuals who do it. This graph treats activities and individuals both as the same type of node, and is a bipartite graph.

We may use features in Gephi to modify the appearance of the graph. Click on “overview” and then “graph.” One possibility is the drag feature. With the mouse, hover over a node, when the edges not connected to that particular node fade away, click on the node and drag it. Dragging all
the activities to the left and all the individuals to the right will clearly show the bipartite nature of the graph.

Additional modifications may be made using the preview settings. Clicking on “preview” brings up a column of preview settings on the left. Under “nodes” you may reduce the opacity of the nodes so that the labels are more visible. Under “node labels” you may enlarge the font of the node labels, and under “edges” you may choose for the edges of the graph to be curved. To see the results of any changes, click on the “refresh” button at the bottom of the preview settings. A version of our graph incorporating several of these changes is shown as Figure 22.

![Figure 22](image_url)

**Figure 22.** A visually modified version of the first graph

### 4.4 Analysis of the graph

We can analyze the graph representing our network. In Gephi, click the “overview” tab in the top left, then look to the right of the screen. We find properties of the network including the number of nodes, number of edges, and the fact we have chosen an undirected graph to represent the network. Our sample graph has 16 nodes representing 5 individuals and 11 of the twelve possible activities on the survey. It also has 22 edges.

Running the various items under “network overview” gives properties of the entire graph, but also calculates properties for individual nodes. The average degree is just that, an average of the number of edges connected to each node. When you run average degree, you will see that for our example the average degree is 2.75. You will also see a degree report which indicates the number
of nodes of each degree. In our example there are 2 nodes of degree 1, 8 nodes of degree 2, 2 nodes of degree 3, 2 nodes of degree 4, 1 node of degree 5, and 1 node of degree 7. Notice that \((2 \times 1) + (8 \times 2) + (2 \times 3) + (2 \times 4) + (1 \times 5) + (1 \times 7) = 44\). But the number of edges is only 22! So, what happened? Each edge was counted with both of the nodes it connects.

Since we have not assigned any edge weights, the average weighted degree is currently the same as the average degree.

The network diameter gives the length of the longest, shortest path between any pair of nodes. Here the network diameter is 5, meaning that there is at least one pair of nodes for which it would take 5 hops to move from one to the other. Within the graph distance report the average path length between any two nodes is given as 2.39.

Running network diameter also gives measures of betweenness centrality, closeness centrality, and eccentricity distribution. Click on “data table” at this point and you will see that Gephi has added columns showing the applicable measures for each node. Notice that Jan has the highest betweenness measure and the lowest closeness measure.

### 4.5 Applying analysis results to the graph

Now that we have run these analyses, we can apply some of the results to the graph.

Click the ranking tab on the far left of the screen and be sure you are looking at node measures. Choose a rank parameter such as betweenness, and click apply. If you look at the graph in “overview,” the nodes will now be colored according to their betweenness centrality. To see the colored nodes under preview, click on refresh. Notice that the node for Jan, who had the highest betweenness measure, is now red. You may also size the nodes based on rank parameter values. Click the red diamond above the rank measure. Choose a min size of 5 and a max size of perhaps 20 and apply. You will now see larger dots for the nodes with a higher betweenness measure. Your graph may now look something like Figure23.
Figure 23. Graph highlighting betweenness

Enjoy changing the appearance of your graph to emphasize different measures or aspects of the network.

If you would like to export a picture of your graph you may do so at the bottom of the preview settings window. If you are having trouble with the exported graph cutting off parts of the labels there is not a clean solution but there is a way to work around the problem. In the graph window under the overview tab click on the node pencil. Then click on the graph to add a node at the far top, bottom, left, and right of the graph. Gephi will include these nodes when the graph is exported, leaving the full label to show on the other nodes. Just be sure to delete the nodes if you decide to run additional network measures!

4.6 Assignment

Create a sustainability survey of your own. Choose a subset of the possible survey questions or write some new ones. Survey a group, record the data, create a graph, analyze the graph, and use the graph to illustrate various network measures.
Section 5: The American Revolution: Who was in Charge?

We begin this section with a small and simple example that can be done with pencil and paper but, more importantly, should be done in Gephi so that by the time we deal with the much larger American Revolution data set that cannot be done with pencil and paper, we will be familiar with the concepts and the Gephi software. It is unlikely that we will appreciate the power of network analysis in such a small example and may even be disappointed. We shall initially sacrifice the power of the method for the clarity of the smaller example.

5.1 A simple committee membership list

Example 5.1.1 Consider a simple list of three committees A, B, and C with their respective memberships given below:

Committee A: Bob, Dick, Sally, and Mary
Committee B: John, Dick, Sally, and Mary
Committee C: Bob, Dick and Sally

We can use a graph to exhibit this relationship of “being a member of some committee” by assigning a node or vertex to each committee and person and drawing a directed line or edge from vertex X to vertex Y if person X is a member of committee Y. We can let Gephi do the work by creating an edge table and reading it into Gephi as follows.

Procedure 5.1: How to enter an edge table into Gephi

1. Use Excel or any spreadsheet to create a table with 2 columns.
2. On the very first row (header row) of the table, type ‘source’ on the first column and ‘target’ on the second column. Do not type the single quote marks. This is the required first row of the table.
3. For each succeeding row, enter a person-committee pair if the person is a member of that committee. Make sure there are no spaces after the entries. The file should look like Table 7.

<table>
<thead>
<tr>
<th>source</th>
<th>target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>A</td>
</tr>
<tr>
<td>Bob</td>
<td>C</td>
</tr>
<tr>
<td>John</td>
<td>B</td>
</tr>
<tr>
<td>Dick</td>
<td>A</td>
</tr>
<tr>
<td>Dick</td>
<td>B</td>
</tr>
<tr>
<td>Dick</td>
<td>C</td>
</tr>
<tr>
<td>Sally</td>
<td>A</td>
</tr>
<tr>
<td>Sally</td>
<td>B</td>
</tr>
<tr>
<td>Sally</td>
<td>C</td>
</tr>
<tr>
<td>Mary</td>
<td>A</td>
</tr>
<tr>
<td>Mary</td>
<td>B</td>
</tr>
</tbody>
</table>

Table 7. Source-Target table
4. Use Save As to save the file using the CSV format and giving it a meaningful filename.
5. Open Gephi (if it is not already open) and select File -> New Project.
6. Select Data Laboratory -> Data Table -> Import Spreadsheet

![Figure 24. Import dialog box in Gephi](image)

7. Click in the box indicated by the arrow in Figure 24 and navigate to the file saved in step 4.
8. Make sure that the dropdown under the entry ‘As table:’ is showing Edges table and follow the prompts until Finish.
9. You should get a graph similar to Figure 25. I have customized the graph to change the node sizes, show the node labels, and rearrange the node positions to get a better display.

![Figure 25. Gephi graph from Table 7](image)

**Example 5.1.2** Consider the following publication information for a set of three books:

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Year</th>
<th>Pages</th>
<th>Publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>Calculus</td>
<td>1972</td>
<td>857</td>
<td>Oxforge</td>
</tr>
<tr>
<td>David</td>
<td>Algebra</td>
<td>1968</td>
<td>369</td>
<td>Cambridge</td>
</tr>
<tr>
<td>Susan</td>
<td>Poetry</td>
<td>1981</td>
<td>112</td>
<td>Oxforge</td>
</tr>
</tbody>
</table>

Table 8. Excel table of publication information for Example 5.1.2
Suppose we want a graph connecting Author to Publisher. We shall use a Gephi plugin (Excel/csv converter to network) to input this data into Gephi in order to create the graph. The plugin allows you to choose which columns form the relations and hence the network. This plugin is not part of the original Gephi install but it can be downloaded from the Gephi marketplace website: https://marketplace.gephi.org/plugin_categories/plugin-imports/. The procedure below assumes you have installed this plugin.

Procedure 5.2: How to enter an Excel table using the Gephi plugin Excel/csv converter to network to create a graph connecting any 2 columns of the table

1. Install the plugin in Gephi (File -> Tools -> Plugins -> Available).
2. Restart Gephi and select File -> Import Spigot…
3. In Category, select “Data importer” and follow the steps of the wizard.
4. In “Choose the field delimiter” dialog screen, choose “comma”.
5. In Select agents, choose “Author” in the first dialog box and “Publisher” in the second dialog box as shown in Figure 26.

Figure 26. Select agents dialog box in Gephi

6. Select Next in the succeeding dialog boxes until you get to Finish.
7. Finally, we selected directed edges and customized the graph to obtain the graph in Figure 27.

Figure 27. Gephi graph for Author and Publisher data in Table 8

Example 5.1.3 Consider a different view of the graph in Figure 25.
This is clearly a bipartite graph (like the one in Figure 22), whose vertices can be partitioned into two sets S and T so that the only edges go from a vertex in one set (S, for example) to a vertex in the other set (T). There are no edges between any two vertices in the same set. As we saw in Section 2.7, the adjacency matrix for this graph would be an 8 x 8 array because the graph has 8 vertices. However, we can create a “reduced adjacency matrix” (i.e., a matrix with smaller dimensions) for this graph, since the rows and columns of the matrix do not have to include all the vertices. If we put the names in the rows and the committees in the columns, then we obtain the following “reduced adjacency matrix” of people-by-groups

\[
M = \begin{pmatrix}
1 & 0 & 1 \\
0 & 1 & 0 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 0 \\
\end{pmatrix}
\]

But if we take the transpose, we then obtain from the same graph, a “reduced adjacency matrix” of groups-by-people

\[
M^T = \begin{pmatrix}
1 & 0 & 1 & 1 & 1 \\
0 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 0 \\
\end{pmatrix}
\]

If we now multiply the two matrices together, following the rules of matrix multiplication given in Section 2.3, we obtain two matrices depending on the order of multiplication:

\[
A = MM^T = \begin{pmatrix}
1 & 0 & 1 & 2 & 0 & 2 & 2 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & & & & & \\
\end{pmatrix} = \begin{pmatrix}
2 & 1 & 3 & 3 & 2 \\
2 & 1 & 3 & 3 & 2 \\
1 & 1 & 2 & 2 & 2 \\
\end{pmatrix}
\]

and
\[ B = M^T M = \begin{bmatrix}
1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
\end{bmatrix} \begin{bmatrix}
1 & 0 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
\end{bmatrix} = \begin{bmatrix}
4 & 3 & 3 \\
3 & 4 & 2 \\
3 & 2 & 3 \\
3 & 2 & 3 \\
3 & 2 & 3 \\
\end{bmatrix} \]

The 5 x 5 product matrix \( A = MM^T \) is not an adjacency matrix but is a “strength-of-connection” matrix of a graph whose vertices consists only of persons. The entry in the \( i,j \) position tells us the number of committees in which the pair of persons, \( i \) and \( j \), are co-members. Here is the same data but with the row and column labels:

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>John</th>
<th>Dick</th>
<th>Sally</th>
<th>Mary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>John</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dick</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Sally</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Mary</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 9. Person-by-person strength-of-connection matrix for Example 5.1.1

For example, \( a_{3,4} = 3 \), and since row 3 corresponds to Dick while column 4 corresponds to Sally, this states that Dick and Sally are connected by being co-members of the same 3 committees, although we do not know which 3 committees. On the other hand, John is not connected to Bob through committee membership since \( a_{2,1} = 0 \), but this does not imply that John and Bob do not have any connection at all! In fact, they could be good friends but that is not deducible from the data. Furthermore, since \( a_{2,2} = 1 \), this states that John is a member of only 1 committee, although again, we do not know which committee. Let us now use Gephi to draw the graph associated with this strength-of-connection matrix.

**Procedure 5.3: How to enter a square strength-of-connection matrix into Gephi**

1. Use Excel to create the table using the first row and first column to put meaningful labels to the rows and columns.
2. Save the file in CSV format giving it a meaningful filename.
3. Open the file in a text editor.
4. Replace all occurrences of comma with a semi-colon. It is best to use the Find and Replace All command. Be careful, especially if the file is large since the rows could wrap and continue to the next line. You do not want to put an extra line return anywhere! The file should look like Table 10.
5. Although it is not required and could be cumbersome, you may also want to replace all the diagonal entries with 0 since these edges represent loops (vertex connected to itself). Gephi seems to ignore these entries but they could affect the weight of the edges.

6. Restart Gephi and select File -> Open.

7. Navigate to the appropriate file and select it. Select Open in the dialog box.

8. In the Import Report dialog box, select Undirected under Graph Type and select OK.

You should get a graph similar to the Figure 29, although we have made some customizations on the graph.

![Figure 29. Person-by-person strength-of-connection graph](image)

You can see that Sally and Dick have the strongest connection since they share membership in the most number of committees. This can actually be inferred by looking at the committee membership list. However, such a conclusion is not as easy if there were more committees and more members.

On the other hand, the product matrix \( \mathbf{B} = \mathbf{M}^\top \mathbf{M} \) is a 3 x 3 committee-by-committee strength-of-connection matrix of a graph whose vertices are the committees. Here is the same data but with the row and column labels.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 11. Co-membership table
Using Gephi, we get the graph in Figure 30.

![Figure 30. Co-membership graph from Table 11](image)

What we get is not very interesting but nevertheless, it tells us that committee A is the most influential! Perhaps, this can be inferred from the table above but certainly not as easy to infer if the table is much larger.

We can now see that people are linked by their shared memberships in groups and groups are linked through the people that belong to them both.

**Activity 5.1.1** Create your own edges table in Excel illustrating a relationship (like friend of, has read, visited, etc) between two lists (of names, places, titles, etc.) and use Gephi to show a graph of the relationship.

**Activity 5.1.2** Create your own table similar to Example 2 in Excel containing data that is interesting to you and use the plugin to graph a relationship between 2 entities in your data.

### 5.2 What is metadata?

“The USA Freedom Act is a U.S. law enacted on June 2, 2015 that restored in modified form several provisions of the Patriot Act, which had expired the day before. The act imposes some new limits on the bulk collection of telecommunication metadata on U.S. citizens by American intelligence agencies, including the National Security Agency…. The bill was originally introduced in both houses of the U.S. Congress on October 29, 2013, following publication of classified NSA memos describing bulk data collection programs leaked by Edward Snowden that June.” [https://en.wikipedia.org/wiki/USA_Freedom_Act](https://en.wikipedia.org/wiki/USA_Freedom_Act)

Metadata is data about other data. It provides information about a certain item's content. For example, an image may include metadata that describes how large the picture is, the color depth, the image resolution, when the image was created, and other data. A text document's metadata may contain information about how long the document is, who the author is, when the document was written, and a short summary of the document. An ordinary letter, sent by the U.S. postal service, displays several kinds of metadata: the names and addresses of both the addressee and the sender, the time when the letter was posted, the location of the post office from which the
letter was sent, and the size of the information being mailed (as measured by the cost of the postage). In the figures below, you will see several examples of metadata.

Figure 31. Metadata from an ordinary letter

Internet routing is based on policy (i.e. economics) so traceroutes do not give shortest-paths

Figure 32. Metadata from an email

Figure 33. Twitter data
Actually, any publicly available data is metadata. It points to other data that may or may not be publicly available. It is amazing what one can intelligently infer from publicly available data using mathematical operations, but it is equally easy to fall into misinformation and discover suggestive, but ultimately incorrect or misleading, conclusions if the analysis is not done properly. There is a need to understand both the power and the limitation of the method.

**Activity 5.2.1** Find other examples of metadata.

**Answer key for Activity 5.2.1**

Exif data for camera photos, Library of Congress Cataloging-in-Publication Data, Bar Codes, Car Plate Number, etc.

### 5.3 The American Revolution data set

In his 1994 book, *Paul Revere's Ride*, historian David Hackett Fischer provided membership lists for seven social clubs in Boston in 1775. These clubs were hot beds of revolutionary activity. Here is a short portion of the data in Fischer’s book from original sources:
Here is a partial view of the club membership data set created by Fischer from the original sources:

<table>
<thead>
<tr>
<th>Name</th>
<th>St. Andrews Lodge</th>
<th>Loyal Nine North</th>
<th>North Caucus</th>
<th>Long Room Club</th>
<th>Boston Tea Party</th>
<th>Boston Committee of Correspondence</th>
<th>London Enemies List</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams, John</td>
<td>1762</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adams, Samuel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allen, Dr.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appleton, Nathaniel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ash, Capt. Gilbert</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austin, Benjamin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austin, Samuel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avery, John</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baldwin, Cyrus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ballard, John</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barber, Nathaniel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barnard, Samuel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barrett, Samuel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bass, Henry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bell, Capt. William</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blake, Increase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boit, John</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bolter, Thomas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boyer, Peter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boynton, Richard</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brackett, Jos.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradford, John</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradlee, David</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradlee, Josiah</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12. Partial club membership list from David Hackett Fischer
http://www.amazon.com/Paul-Reveres-David-Hackett-Fischer/dp/0195098315
Sociologist Kieran Healy created an Excel spreadsheet of this same data and replaced the X with a 1 indicating membership in a club and filling the rest of the cells with 0 indicating non-membership in a club. Here is a portion of that data set:

<table>
<thead>
<tr>
<th>Name</th>
<th>StAndrewsLodge</th>
<th>LoyalNine</th>
<th>NorthCaucus</th>
<th>LongRoomCl</th>
<th>TeaParty</th>
<th>BostonCommittee</th>
<th>LondonEnemies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams.John</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adams.Samuel</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Allen.Dr</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Appleton.Nathaniel</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ash.Gilbert</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Austin.Benjamin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Austin.Samuel</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Avery.John</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Baldwin.Cyrus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ballard.John</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Barber.Nathaniel</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Barnard.Samuel</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Barrett.Samuel</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Bass.Henry</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Bell.William</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Blake.Increase</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Boit.John</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bolter.Thomas</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Boyer.Peter</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Boynton.Richard</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 13. Reduced adjacency matrix for Fischer's club membership list

For the complete list in CSV format, you can download it from:

[https://github.com/kjhealy/revere/blob/master/data/PaulRevereAppD.csv](https://github.com/kjhealy/revere/blob/master/data/PaulRevereAppD.csv)

or from:[https://raw.githubusercontent.com/kjhealy/revere/master/data/PaulRevereAppD.csv](https://raw.githubusercontent.com/kjhealy/revere/master/data/PaulRevereAppD.csv)

We are now going to create a 254 x 7 matrix whose rows corresponds to the names of persons and whose columns correspond to the names of the clubs. This is a reduced adjacency matrix for a bipartite graph. It may be of interest to create the graph of this relationship in Gephi. The general procedure for doing this was shown in Section 5.1. Some amount of information can actually be obtained from this bipartite graph but the graph is quite messy and the information is not as meaningful!

We want to bring out information that is in this data set but perhaps not that easy to see. If we take this “reduced” adjacency matrix, $A$, of people and groups (clubs) and take its transpose $A^T$, we then obtain a “reduced” adjacency matrix of groups (clubs) and people. We will now perform the same process we did in Example 5.1.3 of Section 5.1 and look at the graph corresponding to the 254 x 254 “person-by-person” strength-of-connection matrix $AA^T$ and the 7 x 7 organization-by-organization strength-of-connection matrix $A^TA$.

At this point, we have a computationally challenging task: that of multiplying two “relatively” large matrices and is certainly not one that could be done casually on a piece of paper! One
matrix has 254 rows (and a few columns) and the other has 254 columns (and a few rows)! You will need a very large piece of paper! Luckily, these operations can be done in Excel! In order to really profit from this discussion, we recommend that the reader actually perform these operations on a computer.

Procedure 5.4: How to multiply a matrix and its transpose in Excel:
1. We are going to assume that the m x n matrix A has been entered in an Excel workbook.
2. At the bottom left portion of the workbook next to the Sheet1 tab, click on the + sign to insert a new sheet and repeat the process 2 more times. This will insert 3 new sheets into the workbook, one to hold the transpose of the matrix, the second to hold the product \( AA^T \) and the third to hold the other product \( A^TA \). You can change the names of the sheet by double clicking on the tab and typing a meaningful name for that sheet.
3. Click on the sheet that will hold the transpose in order to bring it to the top.
4. Highlight or select an n x m array of cells (n rows and m columns) that will hold the transpose and while it is still highlighted, type the following formula:
   \[ =\text{TRANSPOSE}(\text{Sheet1}!\text{xx};\text{yy}) \]
   where xx is the cell address of the upper left corner cell of the original matrix and yy is the cell address of the lower right corner cell of the original matrix.
5. Click on the sheet that holds the original matrix and highlight or select the entire matrix and only the entire matrix.
6. Type the closing right parentheses.
7. Press the Shift key and the Ctrl key simultaneously and then press the Return key. This will change the formula you just typed into:
   \[ =\text{TRANSPOSE}(\text{Sheet1}!\text{xx};\text{yy}) \]
8. The result will be displayed on the highlighted cells of the sheet referred to in Step 3.
9. Click on the sheet that will hold the product \( AA^T \).
10. Since the product will be an m x m matrix, highlight or select an m x m array of cells (m rows and m columns) that will hold the product \( AA^T \) and while it is still highlighted, type the following formula: \( =\text{MMULT}(\text{Sheet1}!\text{xx};\text{yy},\text{Sheet2}!\text{ss};\text{tt}) \)
11. Click on the sheet that holds the original matrix and highlight or select the entire matrix and only the entire matrix.
12. Type a comma.
13. Click on the sheet that holds the transpose of the original matrix and highlight or select the entire matrix and only the entire matrix.
14. Type the closing right parentheses.
15. Press the Shift key and the Ctrl key simultaneously and then press the Return key. This will change the formula you just typed into:
   \[ =\text{MMULT}(\text{Sheet1}!\text{xx};\text{yy},\text{Sheet2}!\text{ss};\text{tt}) \]
   where xx and yy are the same as in step 7 while ss is the cell address of the upper left corner cell of the transpose of the original matrix and tt is the cell address of the lower right corner cell of the transpose of the original matrix.
16. The result will be displayed on the highlighted cells of the sheet referred to in Step 9.
17. Repeat steps 9-16 except that in step 10, the product \( A^TA \) is an n x n matrix hence steps 11 and 13 are interchanged.
We could display a portion of the 254 x 254 “person-by-person” strength-of-connection matrix $AA^T$ but we doubt that any information can be inferred from it. Instead, we use Gephi to graph the matrix where the vertices are the persons and an edge in the graph means co-membership in a group. People are linked by the “things” they share or have in common with, like membership in the same group. We used the layout Force Atlas 2 in Gephi and it has arranged everyone neatly, showing clusters of individuals, peripheral persons and persons that seem to “bridge” various clusters.

![Gephi graph of person-by-person strength-of-connection matrix for Fischer data set](http://kieranhealy.org/files/misc/revere-network-reduced.png)

If you did this in Gephi, you can actually zoom in anywhere on the graph and see who the individuals are. You should now be able to find out that person in the center of the network who seems to bridge the various groups. Gephi offers several ways of measuring the importance or centrality of a vertex in a network. Ideally, the measure one chooses should tell something about how the vertex fits into the overall structure of the network and be efficiently computable as well.
5.4 The central figure for the American Revolution data set

In Chapter 3, you encountered three centrality measures for determining the node (or nodes) of key importance in a given network: degree centrality, betweenness centrality, and closeness centrality. We briefly review the definitions of these measures and interpret them informally.

First, degree centrality measures which node has the greatest degree among all the nodes in the network and is the simplest measure of centrality. In a friendship network, degree centrality measures who is most popular while in epidemiology, degree centrality for an infectious disease contact network is seen as "a measure of immediate influence—the ability to infect others directly or in one time period."

Betweenness centrality measures how often a node appears on shortest paths between nodes in the network, i.e., it measures "... how critical a node is in connecting other nodes as a broker or gateway." A hacker who is trying to disrupt a network aims to knock out nodes having betweenness centrality.

Closeness centrality is the average distance from a given node to all other nodes in the network or, as we defined it in Section 3, it is the reciprocal of this average distance. (With the latter definition, the node of maximum closeness centrality is closest to the rest of the nodes). "Think of closeness in terms of a node’s ability to spread energy to all other nodes in a graph." If you have a new product you want to publicize, you want to share your news with nodes having closeness centrality.

Now, the Gephi visualization software not only provides powerful displays of data; it also calculates the above centrality measures and other important network metrics, including eigenvector centrality, HITS, and network modularity.

Eigenvector centrality is a measure of node importance in a network based on a node's connections. "In degree centrality, we consider nodes with more connections to be more important. However, in real-world scenarios, having more friends does not by itself guarantee that someone is important: having more important friends provides a stronger signal."

HITS actually computes two different scores: hubs and authority. The authority score indicates the value of the node itself, and hubs estimates the value of the links outgoing from the node.

Modularity measures how well a network decomposes into “communities.”

Here is a portion of the report from Gephi for the American Revolution data:
Table 14. Gephi statistics for Fischer data set

Activity 5.4.1 Using the measures of centrality from Section 3 and the data from Table 14, who do you identify as the most central, most important person in this group?

Answer key for Activity 5.4.1

A look at the entire report provides a convincing argument that Paul Revere is the most central, most important person in the group using any measure of centrality.

Similarly, we can do the same procedure for the 7 x 7 “organization-by-organization” strength-of-connection matrix \( A^T A \). In fact, here is the entire matrix:

Table 15. Organization-by-Organization matrix for Fischer data set

The corresponding graph from Gephi is shown in Figure 37.
It is not as revealing as the previous graph (although visually, the graph here is more easily interpreted) but it says that the connection between the North Caucus group and the London Enemies was the strongest and Loyal Nine was on the periphery. Groups are linked through the people they share.

**Homework Exercises for Section 5**

1. Select a small set of something (names, words, titles, places, blogs, or combination of these) for which a relationship between any two element in the set may or may not exist and create a graph of the relationship using Gephi or pencil and paper.
2. Do any of these phrases have resonance for you?
   a. The problem with this city is that it’s run by X’s cronies.
   b. Contracts in this town are all about kickbacks.
   c. Follow the money!
   d. It’s not what you know, it’s who you know.
   e. Board of directors of corporations across certain industries
   f. Lobbyists in Washington
   g. Political campaign contributors

   If so, and you can gather relevant data, network graphs can provide a means to display the connections involved in a way that your audience (and your editors) can readily understand.

**Answer key for homework exercises for Section 5**

Answers to questions 1 and 2 will vary.
6. Uncloaking Terrorist Networks

We recommend reading *Uncloaking Terrorist Networks*, by Valdis E. Krebs. In the article on mapping terrorists networks, Krebs reveals which hijacker was in charge of the 9/11 terrorist network according to several measures of network centrality. The abstract for the article is copied below.

*This paper looks at mapping covert networks using data available from news sources on the World Wide Web. Specifically, we examine the network surrounding the tragic events of September 11, 2001. Through public data we are able to map a portion of the network centered on the 19 dead hijackers. This map gives us some insight into the terrorist organization, yet it is incomplete. Suggestions for further work and research are offered.*

6.1 Possible discussion questions

1. What was Krebs’ purpose in constructing a map of the terrorist network?
2. Why was Krebs particularly qualified to undertake this task?
3. How did Krebs collect information?
4. How did Krebs construct the terrorism network?
5. What sorts of ties were particularly important to holding the terror network together?
6. When constructing the edges of the network, what did Krebs consider beyond just the presence or absence of a tie between two individuals?
7. How did Krebs rank the strength of ties between individuals?
8. The abstract mentions that the graph of the organization is incomplete. In what ways is it incomplete and why?
9. In graph theory a complete graph is one in which every node is directly connected to every other node. Thus, the map of the terrorist network is certainly incomplete in the graph theoretic sense. In fact, Krebs makes a particular note of how sparse the network is. What purposes would this serve for the terrorist organization? In what ways would sparseness hinder the organization?
10. What is the diameter of the network in Figure 2 and what is the average path length?
11. How does the addition of the shortcuts shown in Figure 3 change the diameter and the average path length?
12. Why would the terrorists not wish to make the shortcuts permanent?
13. Who was the key player behind the September 11, 2001 attacks?
14. If the key player were removed what other individuals could potentially fill that role?
15. The map of the network was clearly built without full knowledge of the network. What could happen to the various network measures if new conspirators or new connections among existing conspirators were found?
16. What should be done about individuals linked to known terrorists?
17. How do covert networks differ from normal social networks?
18. What steps can members of covert networks take to reduce the visibility of the network or to hide the strength of the ties within the network?
19. What types of ties between conspirators could an intelligence agency attempt to identify and what data sources might be useful?
20. What are the potential dangers of using incomplete data to label individuals as terrorists?
21. What is “snowball sampling” and why might it be useful?
22. What should multiple intelligence agencies do with their information and why?

6.2 Additional reading

The article *Connecting the Dots: Tracking Two Identified Terrorists*, also by Krebs would be an interesting addition to the article discussed above.

6.3 Responses to discussion questions

While we hope the discussion questions will lead to extensive conversations and even debates, some short, possible responses are given below to each question.

1. What was Krebs’ purpose in constructing a map of the terrorist network?

Krebs was looking for patterns, and mathematical analysis is ideal for finding patterns. He hoped to discover the forms of organization and communication preferred by the terrorist network Al Queda, and he felt that this knowledge could be useful in finding other such networks.

2. Why was Krebs particularly qualified to undertake this task?

Krebs had been researching organizational networks for many years. He had, and still has, his own consulting company, orgnet.com, and he had mapped numerous project teams over the years. His firm provides social network analysis software. Even with his qualifications, when Krebs realized mapping a covert network would be very different from mapping an overt one, he turned to the work of social network theorists who had studied secret networks.

3. How did Krebs collect information?

Krebs used publicly available information that was released through several major newspapers. He knew that there would be information that was either not known or not released by government investigators. He also discovered that some data provided to the media was either simply incorrect or intentional misinformation.
4. How did Krebs construct the terrorism network?

Creating the map of the network was a step-by-step process using network mapping and measuring software. Krebs would add nodes for each hijacker and links between hijackers as new information was published. He knew that some data was missing and some was inaccurate, so he did not add nodes or links hastily.

5. What sorts of ties were particularly important to holding the terror network together?

Prior contacts that had created strong, trusted ties were crucial to the network. Many of the hijackers were relatives or had gone to school together as children. Others had lived together for many years, and they apparently all came together at a terrorist training camp in Afghanistan. By only linking with individuals known and trusted for many years, the network was able to greatly reduce the opportunity for discovery by outsiders.

6. When constructing the edges of the network, what did Krebs consider beyond just the presence or absence of a tie between two individuals?

Krebs looked at the strength of the tie and also at the increase or decrease in that strength over time. For example, Figure 3 of the article shows weak temporary ties from a one-time meeting of some of the hijackers in Las Vegas.

7. How did Krebs rank the strength of ties between individuals?

Krebs chose to use three strengths based on the time two individuals had spent together. The hijackers who were from the same families, who had gone to school together, or who had lived together were considered to have the strongest ties. If two individuals did not have the relationships that built strong ties but had traveled or attended multiple meetings together, they were considered to have ties of medium strength. Finally, the ties between individuals with only infrequent contact were classified as weak ties.

8. The abstract mentions that the graph of the organization is incomplete. In what ways is it incomplete and why?

The data Krebs used was from news sources. It did not include information investigators chose not to reveal and could not include information investigators never discovered. As previously noted, some of the information was false. Therefore there were both nodes and links missing from the graph that Krebs created. In addition, the network was covert so even when individuals and ties were discovered, the strength of those ties could be difficult to determine.
9. In graph theory a complete graph is one in which every node is directly connected to every other node. Thus, the map of the terrorist network is certainly incomplete in the graph theoretic sense. In fact, Krebs makes a particular note of how sparse the network is. What purposes would this serve for the terrorist organization? In what ways would sparseness hinder the organization?

Limiting contacts between individuals helps to keep the network hidden by reducing visibility and the likelihood of leaks. Keeping very few links also protects the network in the event of the capture of one member of the organization. If that individual has no knowledge of parts of the network, then he cannot reveal anything about those parts. Sparseness could negatively affect the organization by reducing the ability to communicate easily, which in turn could limit the organization’s ease in accomplishing tasks. In short, the network trades increased secrecy for reduced efficiency.

10. What is the diameter of the network in Figure 2 and what is the average path length?

Recall that the diameter of a graph is the length of the longest, shortest path between any two nodes in the graph. For the network mapped in Figure 2, the diameter is 9 due to the length of the shortest path from MajedMoqed to either Wail Alshehri or SatamSuqami. As indicated in Table 1 of the article, the average path length is 4.75.

11. How does the addition of the shortcuts shown in Figure 3 change the diameter and the average path length?

With the shortcuts added the distance from MajedMoqed to either Wail Alshehri or SatamSuqami is reduced to 5, but the diameter only decreases to 7, the length of the shortest path from Ahmed Alghamdi to either Wail Alshehri or SatamSuqami. The average path length shrinks to 2.79, a reduction of over 40%.

12. Why would the terrorists not wish to make the shortcuts permanent?

While the shortcuts improved the ability of the group to communicate and allowed the pilots to work together more efficiently, the increased contact could raise the visibility of the network and increase the possibility of discovery.

13. Who was the key player behind the September 11, 2001 attacks?

Mohammed Atta was considered to be the leader of the network both by investigators and by bin Laden. Atta had the highest centrality measures for degree, closeness, and betweenness. As Krebs wrote:
“The network metric Degrees reveals Atta’s activity in the network. Closeness measures his ability to access others in the network and monitor what is happening. Betweenness shows his control over the flow in the network - he plays the role of a broker in the network. These metrics support his leader status.”

14. If the key player were removed what other individuals could potentially fill that role?

Other individuals with high network centrality measures theoretically could have become leaders if Mohammed Atta had been removed. These include Marwan Al-Shehhi, Hani Hanjour, Essid Sami Ben Khemais, and Nawaf Alhazmi. Notice that Essid Sami Ben Khemais was not one of the nineteen hijackers.

15. The map of the network was clearly built without full knowledge of the network. What could happen to the various network measures if new conspirators or new connections among existing conspirators were found?

Adding new conspirators or new connections to the network would certainly change the network measures. If new conspirators had only a few connections and if those new connections were to individuals who were not well-connected in the existing network, then the additions would be likely to cause the network diameter and the average path length to increase. On the other the hand, adding new conspirators with numerous connections or adding new connections between known conspirators would reduce the network diameter and average path length. In either case, the centrality measures for individuals would change and therefore the individual identified as being in charge could very well change.

16. What should be done about individuals linked to known terrorists?

This question could generate a variety of opinions.

Some may argue that those individuals linked to known terrorists should automatically be assumed to be terrorists. Therefore, if they are not U.S. citizens they should be deported (if in the U.S.) or perhaps even killed. If they are U.S. citizens, perhaps they should be put into internment camps as many U.S. citizens of Japanese descent were during WWII.

Others may argue that individuals with links to terrorists are a valuable resource. These persons of interest should be closely monitored, with no regard to privacy issues, in order to learn as much as possible.
Finally, some may argue that we must assume that every person is innocent until proven guilty. Thus any actions against these individuals or monitoring of their lives would be unfair profiling.

17. How do covert networks differ from normal social networks?

The members of covert networks take steps to hide the ties between them. In a normal social network, no effort is made to hide strong ties, which are usually obvious due to the frequency and length of contact. In fact, strong ties are often celebrated by parties, school reunions, family gatherings, etc. Members of secret groups may also avoid ties outside of the group in order to reduce the chance of leaks. Normal social networks, depending on their inclusivity or exclusivity, may or may not cultivate ties outside of the group.

18. What steps can members of covert networks take to reduce the visibility of the network or to hide the strength of the ties within the network?

Conspirators can keep communication to a minimum. Even members with strong, long-standing ties, can avoid communicating with each other in order for those ties to remain undiscovered by outsiders.

19. What types of ties between conspirators could an intelligence agency attempt to identify and what data sources might be useful?

In Table 3 of his article, Krebs identifies four types of relationships: trust, task, money and resources, and strategy and goals. He lists a variety of data sources including prior contacts, public records, records of various types of communications, travel records, financial records, court records, web sites, videos, and human observations.

20. What are the potential dangers of using incomplete data to label individuals as terrorists?

Clearly, there is a risk of falsely identifying an innocent individual as a terrorist. This innocent person could be subjected, at a minimum, to close scrutiny. Beyond this invasion of privacy, the individual could be questioned, have his or her travel restricted, or be denied certain jobs. Family and friends of someone incorrectly identified as a terrorist could also come under scrutiny.

An additional danger of a false positive identification of someone as a terrorist is the waste of resources spent pursuing information about that individual. The time and money spent focusing on a non-terrorist is time and money that cannot be used examining other individuals. Serious terrorist risks might be overlooked.
21. What is “snowball sampling” and why might it be useful?

Snowball sampling is a research technique in which individuals currently being studied are used to find additional subjects for study. In some studies, a researcher could simply ask his or her subjects to recruit their friends and families. Those new participants would in turn be asked to recruit their own friends and families. In this way, the number of subjects in the study could not only grow but grow more and more quickly. When trying to uncover a covert network, the researcher cannot ask those being studied for the names of additional people. The researcher can only add individuals as they are revealed to be in contact with someone already in the network.

Even with its limitations, snowball sampling does provide a logical place to look for possible terrorists, namely associates of known terrorists. A map of a network can slowly be pieced together, and analysis of this network might reveal who is at its center.

22. What should multiple intelligence agencies do with their information and why?

As became apparent after 9/11, agencies should share their information. Sharing would contribute to building a more comprehensive map of terrorist organizations. In turn, better knowledge would lead to a greater ability to combat terror.
7. Student Projects

In this section we propose projects for students interested in further exploration of networks.

7.1 Further analysis of the sustainability survey data of section 4 using methods of section 5

For this project you will use the methods illustrated in section 5 to discover that Paul Revere was “in charge” according to network centrality measures to find who is in charge of your personal sustainability network.

If you did exercise 4.6, then you already have the results of conducting a sustainability survey. If not, choose a few questions from the list provided in Table 2 and conduct a survey.

Once you have the results of a survey, summarize the data in a reduced adjacency matrix of people by groups as in Example 5.1.3. That is, create a matrix $M$ with names of individuals for the row headers and names of sustainability activities for the column headers. As noted in section 4, to prevent Gephi from representing a single person or activity with multiple nodes it is best to use a single string for each row or column header. In the matrix, enter a 1 if that row’s person participates in that column’s activity, otherwise enter a zero.

Continue to prepare for building a Gephi graph by transposing $M$ to find the reduced adjacency graph of groups by people $M^T$. Use Procedure 5.4 to multiply $MM^T$ in Excel to obtain a person-by-person adjacency. Use Procedure 5.3 to enter this adjacency matrix into Gephi and produce a graph.

Use the tools in Gephi to find each measure of centrality. After running network diameter, look at the data table. Clicking on a column header in the data table will sort the data based on the individual values for that measure. Doing this, you can see who is in charge of your sustainability network!

Once you have identified the people in charge, discuss what different powers might do with the knowledge that someone is a sustainability leader. An environmental organization, a marketer of green products, an energy company, and our government are each likely to view such an individual differently.

Clearly the methods used here can be applied to a variety of situations. You could conduct surveys in other areas of interest and analyze the results.
7.2 An introduction to Netlytic

Netlytic(\url{http://netlytic.org}) is “a cloud-based text and social networks analyzer used to collect, analyze, and visualize publicly available online conversations from social media websites such as Facebook, Twitter, and Instagram.”xviii As social creatures, our online lives just like our offline lives are intertwined with others within a wide variety of social networks. Each retweet on Twitter, comment on a blog or link to a YouTube video explicitly or implicitly connects one online participant to another and contributes to the formation of various information and social networks. Once discovered, these networks can provide researchers with an effective mechanism for identifying and studying collaborative processes within any online community.xix

We encourage you to explore Netlytic using data from various online social networks, starting with Twitter. Go to \url{http://netlytic.org} and create a Netlytic account. Create a Twitter account for Twitter data collection. Login to Netlytic and link the newly created Twitter account to your Netlytic account. (See how in this video tutorial: \url{https://www.youtube.com/watch?v=U4mLzxfAiTE}.) Complete the "Ebola" online tutorial using a live dataset \url{https://netlytic.org/home/?p=453}. Once you are comfortable with the basics of Netlytic, collect and analyze a dataset related to a topic of your choice.

7.3 A social media analytics assignment by AnatoliyGruzd

Objectives: The goal of this assignment is to apply techniques from social media analytics, text mining and social network analysis to analyze online discourse and network data from social media about a particular event, company, product or service in order to (1) identify main topical themes (e.g., what customers are saying about a particular product), (2) identify key influencers (both individual and organizational accounts), and (3) determine how to use information from (1) and (2) to improve products/services under examination as well as to develop a communication strategy to influence online discourse on this topic.

Software: Netlytic (\url{https://netlytic.org})
Software Documentation: \url{https://netlytic.org/home/?page_id=36}
Software video tutorials: \url{https://www.youtube.com/channel/UCNBfatUFOZJzeoDsV9oOg3w}

Data source: Publically available social media data from Twitter, Facebook, Instagram or YouTube. Datasets for this assignment will be collected by students using Netlytic.
Main Steps:
1. Review two case studies of applying social media analytics using both text and network analysis available in Netlytic:
   - Text analysis with Netlytic: Sochi 2014 - [https://netlytic.org/home/?p=168](https://netlytic.org/home/?p=168)
   - Network analysis with Netlytic: Oscars 2014 - [https://netlytic.org/home/?p=171](https://netlytic.org/home/?p=171)
2. Select a topic relevant to your professional interests (e.g., event, company, product, service).
3. Identify relevant social media platforms and public online groups/hashtags used to discuss the topic in question.
4. Using Netlytic, collect publicly available social media data (e.g., Twitter messages, Facebook group posts, Youtube comments, or Instagram posts) and analyze it with the help of text and network analysis. The outcome of this step will be a set of interactive visualizations.
5. Finally, use the resulting visualizations to complete Objectives 1-3 and prepare a final report (~10 pages). The written project report should be submitted electronically via the course website. The report template will be provided by the instructor.
6. Present the project results in class in a form of “lightning talk” ([http://en.wikipedia.org/wiki/Lightning_talk](http://en.wikipedia.org/wiki/Lightning_talk)). The presentation should be based on the results to be discussed in your final report. One delegate from each group will have 3 minutes to summary the findings. The presentation should be accompanied by up to 3 Power Point slides saved as PDF and submitted electronically via the course website 24 hours before the oral presentation.
Endnotes

i https://en.wikipedia.org/wiki/Petersen_graph

ii https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg


ix http://www.nytimes.com/2014/12/30/health/how-ebola-roared-back.html?_r=0

x Ibid

xi http://faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html


xiii http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/

xiv Ibid


xvi Ibid


xiv Ibid
References


