Tracking the Untrackable

A Social Network Analysis Approach to Investigate ISIS’s Recruitment Techniques Using Twitter Data

KATHRYN COULTER

A SENIOR RESEARCH PROJECT PROPOSAL PRESENTED TO THE DEPARTMENT OF POLITICAL SCIENCE AND THE DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE OF STETSON UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF ARTS AND BACHELOR OF SCIENCE STETSON UNIVERSITY
Acknowledgments

I would like to thank my advisor, Dr. Branton, for guiding my research and work ethic from start to finish. To Dr. Miles, for keeping me calm and supporting me in more ways than I could list. To Dr. Eckroth, for allowing me to sit in on his courses and for being a valuable resource throughout the process. To Dr. Plante, for teaching me, although somewhat against my will, how to use LaTeX. And, of course, to my mom, Dr. Coulter, for teaching me to love math.
Contents

1 Introduction 10

2 Mathematical Concepts 11

2.1 Introduction to Graph Theory ........................................ 11

2.1.1 Purpose of Graph Theoretical Concepts ....................... 11

2.1.2 Basic Concepts and Definitions .............................. 12

2.1.3 Types of graphs ...................................................... 14

2.1.4 Functions of graphs .................................................. 16

2.1.5 Graph operations ..................................................... 17

2.1.6 Matrices ................................................................. 18

2.1.7 Eigenvalues and applications .................................... 22

2.2 Introduction to Social Network Analysis ......................... 26

2.2.1 Purpose of SNA ...................................................... 26

2.2.2 Basic Concepts and Definitions ............................... 28

2.2.3 Centrality Measures ................................................ 30

2.3 Text Analysis ............................................................. 41
2.3.1 Basic Definitions ........................................ 41
2.3.2 Parsing Text .............................................. 42
2.3.3 Representing and Analyzing Documents ................. 44
2.4 Twitter Application ........................................... 44
2.4.1 R ........................................................... 44
3 Prior Literature and Work ......................................... 45
3.1 Framework for Application ..................................... 45
3.2 Previous Applications .......................................... 52
3.3 Twitter and ISIS ............................................... 57
  3.3.1 Why Use Twitter? ........................................... 57
  3.3.2 Qualitative Analysis of ISIS’s Use of Twitter ........... 58
  3.3.3 Quantitative Analysis of ISIS’s Use of Twitter .......... 60
  3.3.4 Attempts to Counter ........................................ 62
4 An Analysis of Recruitment Strategies and Their Success ...... 64
  4.1 Example of Methods and Techniques on “#greekweek” ........ 64
    4.1.1 Future Work for “greekweek” Data ......................... 74
4.2 Analysis of ISIS Tweets
List of Tables

1. Original meta-matrix conceptualization ........................................... 29
2. Geodesic path value for Agent$_B$ .................................................. 33
3. List of Geodesic Paths ................................................................. 35
4. Impact of Agent Removal from [26] .............................................. 47
5. Frequency table of meta-matrix thesaurus from [31] ....................... 50
6. Characteristics of the term document matrix for tweets containing “#greek-week” ................................................................. 66
7. Characteristics of sparse frequent term document matrix for tweets containing “#greekweek” ......................................................... 68
8. Characteristics of more frequent term document matrix for tweets containing “#greekweek” ......................................................... 70
9. Polarity scores for sentiment analysis of tweets from “greekweek” data . . 74
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A graph</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Different drawing of Figure 1</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>A digraph</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>A multigraph</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>A pseudograph</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>A weighted graph</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>A subgraph of the graph, G, presented in 1</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>Figure 1 after deleting vertex c</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>Graph for Connectivity by Eigenvalue example</td>
<td>24</td>
</tr>
<tr>
<td>10</td>
<td>Example graph for calculating the betweenness centrality</td>
<td>32</td>
</tr>
<tr>
<td>11</td>
<td>Example graph for calculating the closeness centrality</td>
<td>35</td>
</tr>
<tr>
<td>12</td>
<td>A weighted graph for Eigenvector Centrality example</td>
<td>37</td>
</tr>
<tr>
<td>13</td>
<td>Example map from [38]</td>
<td>49</td>
</tr>
<tr>
<td>14</td>
<td>Example visualization of meta-matrix thesaurus from [31]</td>
<td>51</td>
</tr>
<tr>
<td>15</td>
<td>Qualitative relational matrix of 9/11 hijackers</td>
<td>53</td>
</tr>
<tr>
<td>16</td>
<td>Trusted prior contacts and meeting tie weighted graph, from [40]</td>
<td>54</td>
</tr>
<tr>
<td>Page</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>An example of a hierarchical network structure, from [41]</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Jemaah Islamiyah Graph - Bali Operation, from [43]</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Image posted by Dawn on numerous Twitter accounts warning of Baghdad invasion, from [46]</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Location claimed in ISIS supporter profiles from [47]</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Tweets per day on average by account from [47], showing the success of the small hyperactive group</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Frequency chart displaying terms with frequencies greater than or equal to 10 for tweets containing “#greekweek”</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Word cloud displaying terms with frequencies greater than or equal to 3 for tweets containing “#greekweek”</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>“#greekweek” network with 89% sparsity drawn by the Fruchterman-Reingold algorithm</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>“#greekweek” network using entire data</td>
<td></td>
</tr>
</tbody>
</table>
Abstract

Tracking the Untrackable:

A Social Network Analysis Approach to Investigate

ISIS’s Recruitment Techniques Using Twitter Data

Kathryn Coulter

May, 2016

Advisor: Dr. Michael Branton

Department: Department of Mathematics and Computer Science, Department of Political Science

The rise of non-state actors and globalization in the international system presents clear challenges to the historical methodology on power, legitimacy, and decision making. Terrorist groups like ISIS have since found new means to establish legitimacy in their control and power. Among their most successful techniques is the use of social media, particularly Twitter, to recruit and spread propaganda. Social network analysis and text analysis can provide frameworks to uncover why ISIS uses Twitter and can identify key trends in how they reach prosperity so an attempt at countering their momentum can be made. In this research, tweets pertaining to ISIS are analyzed in the hope to investigate their recruitment strategies and techniques.
1 Introduction

The international system has faced numerous changes throughout history, but the most recent significant change came at the end of the Cold War with the dissolution of the Soviet Union. No longer was the world a bimodal power system, but instead the world moved towards a unipolar system with the United States becoming the sole hegemonic power. The shift in polarity was accompanied by increasing importance of intergovernmental organizations, a movement away from the state being the sole important political actor, and an enormous influx of technology occurring through all states. Due to all of these factors, security and conflicts at large were no longer necessarily state vs. state instances but rather a nation vs. a state or even non-governmental organizations without even a nation at times waging war against states. These extremist - or militant - led conflicts are not the typical warfare that powerful states like the U.S. have methods to counter, but instead require a new methodology of retaliation and prevention.

Given the changing nature of international and domestic security alike, a new approach is needed to address terrorist concerns. Social network analysis (SNA), which makes use of graph theory, statistics, and data mining, can be useful for providing a framework for these strategies. The underlying premise in SNA is that nodes (for most purposes, people) are connected to other nodes through complex relationships that form edges. These graphs thus form networks that display a group related by some overarching theme. These networks may also depend on an order, although this order may be complex or secretive (like it is in terrorist networks). Terrorist groups are well suited for this application, given they consist of numerous people that span a geographical region and attempt to accomplish a goal centered on an ideology [1].

Following the attacks of September 11th, 2001, a shift in focus occurred and media
outlets and scholars attempted to understand terrorist and covert networks in a new way [1]. However, even prior to the attacks there was a link between the importance of SNA and its application to fighting terrorism [2]. Terrorist networks, for instance, are not based on a typical hierarchical structure that previous network analysis explained. Networks no longer need to have a clear leader, and thus the use of SNA as an asymmetric approach to mapping these networks using specific relational criteria is needed to accurately understand them. SNA is among the clearest approaches to understanding and defining complex networks, given its changing nature and its multifaceted approaches and definitions. Through SNA, terrorist networks can be analyzed to a new degree and thus hopefully understood better. With this understanding, eventually preemptive steps could be taken to combat terrorist activities from ever occurring in the first place.

Many of the core concepts and assumptions of SNA are borrowed from graph theory, a branch of mathematics concerned with discrete relational structures [6]. As such, in order to understand SNA, we must first begin with its roots in Graph Theory. These definitions are primarily taken from [3], [7] and [8].

2 Mathematical Concepts

2.1 Introduction to Graph Theory

2.1.1 Purpose of Graph Theoretical Concepts

Beginning in Königsberg, Prussia, many attribute the field of graph theory to Leonhard Euler, who sought to create a walk that the city’s residents could enjoy that crossed each
of the seven bridges in the town exactly once. [3] While the usefulness of graph theory immediately following this time was only applied limitedly to puzzles and games, since the 1800s mathematicians have used the field to model many things. Since then, the field has completely blossomed and continues to grow with new applications existing each all the time.

2.1.2 Basic Concepts and Definitions

**Definition 2.1.1.** A graph, G, consists of two finite sets, V(G) and E(G). Each element of V(G) is called a vertex and each element of E(G) is referred to as an edge. A graph depicts the relationship between elements of the vertex set, V(G), by connecting the two vertices by an edge from the edge set, E(G). [3] The edge set is represented as dyads from the vertex set. For example, for a vertex set of V(G) = \{a, b, c\}, an edge could be (a, c). Edges that are connected to a vertex are said to be incident to that vertex. Vertices that are connected by an edge are said to be adjacent. Graphs clearly have natural visual representations, and typically look like a diagram consisting of small circles (vertices) and curves (edges) that connect the circles. Figure 1, shown below, consists of 5 vertices \{a, b, c, d, e\} and 6 edges.

![Figure 1: A graph](image)

Note that graphs can be drawn in different ways while displaying the same information.
For example, the graph shown in Figure 1 can be redrawn as shown in Figure 2 and display the same vertices and the same edge sets, but offer a different visualization of this same information.

![Figure 2: Different drawing of Figure 1](image)

There are numerous algorithms for producing different drawings of graphs given a vertex and edge set, among them are **force-directed graph drawings**. The purpose of this type of drawing is to position the vertices in a way so that there are as few edges crossing as possible and all edges are close to being the same length. This works by assigning forces within the vertices based on their relative positions and then using these forces to minimize the energy needed their respective edges [4].

One example of this is the Fruchterman-Reingold Algorithm [5] that draws vertices that are connected by an edge near each other, but not too close. This algorithm is quite complex, and for our purposes we need to simply be able to recognize it as a force-directed drawing and that it tends to place vertices with the most connections in the center and other vertices that are connected with each other near each other.
2.1.3 Types of graphs

Definition 2.1.2. A simple graph is a graph having no loops, an edge that connects a vertex to itself, or multiple edges between any two vertices [8]. Figure 1 is also an example of a simple graph. Simple graphs, following Definition 1.1, have unordered pairs as edges.

Definition 2.1.3. A digraph, or directed graph, is a graph, G, that has an edge set $E(G)$, that contains ordered pairs of vertices. Each edge of a digraph thus has a specific orientation or direction. As shown in 3, there are still 5 vertices of the graph, but there are now 8 edges. 4 of the vertex pairings simply have one direction edges, but between vertices $(b,c)$ and $(c,e)$ there are edges in both directions, thus implying a relationship in both directions for these vertices.

![Figure 3: A digraph](image)

Definition 2.1.4. A multigraph can be directed or undirected, but allows multiple edges to form between two vertices, thus allowing elements to repeat in the edge set. This graph implicitly creates a weighting system to the graph, where any two vertices can be connected by no edge at all or numerous edges. A multigraph is shown below in Figure 4. As you can see, there are two edges connecting $(a,c)$ and $(b,e)$. 
Definition 2.1.5. A pseudograph is similar to a multigraph, but also allows for the possibility of loops to occur within a graph. An example of a pseudograph is presented in Figure 5, where you will note the loop that occurs at vertex $a$.

Definition 2.1.6. A weighted graph associates a label or value to every edge in the edge set $E(G)$. These labels or weights are usually real numbers, but can be confined to a subset of real numbers like positive integers. Weights can represent anything, from distance to perceived importance of a relationship. Figure 6 is an example of a weighted graph.
2.1.4 Functions of graphs

**Definition 2.1.7.** A walk is a sequence of vertices connected by edges where the vertices themselves may repeat. A path is similar to a walk, but a path is a walk with distinct vertices[3]. In Figure 5, an example of a walk would be \((a, c, b, d, e, c, a)\), whereas an example of a path would be \((a, c, b, d, e)\). A geodesic is the shortest path between any two vertices.

**Definition 2.1.8.** A subgraph, \(H\), of a graph, \(G\), is a graph such that the vertex set of \(H\) is a subset of the vertex set of \(G\) and the edge set of \(H\) is a subset of the edge set of \(G\). Consider Figure 1 from before. A subgraph of this graph is presented below.

![Figure 6: A weighted graph](image)

![Figure 7: A subgraph of the graph, G, presented in 1](image)
2.1.5 Graph operations

**Definition 2.1.9. Vertex deletion** produces the graph, G, without the vertex that was deleted as well as any edges incident to that vertex. Consider Figure 1. Deleting vertex c would thus produce the following graph.

![Figure 8: Figure 1 after deleting vertex c](image)

**Definition 2.1.10. Edge deletion** is similar to vertex deletion, but only deletes an edge while leaving the vertices incident to that particular edge.

**Definition 2.1.11.** A graph is said to be **connected** if every pair of vertices can be joined by a path, otherwise a graph is **disconnected**. All examples presented thus far have been connected with the exception of Figure 8.

**Definition 2.1.12.** The members of the set of maximally connected subgraphs of a graph are called its **connected components**. Thus all examples have had one component, with the exception, again, of Figure 8, which has two components.

**Definition 2.1.13.** The set of vertices adjacent to a vertex is said to be its **neighborhood**. Continuing with Figure 8, then, vertex a has no neighborhood while vertex d has vertices b and e in its neighborhood.

**Definition 2.1.14.** The **degree** of a vertex v is the number of edges incident with v [7]. Thus the degree of vertex a in Figure 8 is 1, while all other vertices have degree 2.
2.1.6 Matrices

As we have seen thus far, graphs are very visual objects. Thus far our graphs have been small enough to identify the degree of the vertex by simply counting on our fingers. However, larger graphs require computer assistance and thus a matrix representation is needed. A computer can understand a graph via a matrix representation.

**Definition 2.1.15.** An adjacency matrix $A$ of a graph $G$ is simply an $n \times n$ matrix where $n$ is the number of vertices in the graph. Adjacency matrices are typically symmetric, but a digraph will result in a directed adjacency matrix, which would not necessarily be symmetric. As defined by [9], an adjacency matrix is defined as the following for the entry for vertices $i$ and $j$:

$$a_{ij} = \begin{cases} 
0 & \text{if no edge exists between } v_i \text{ and } v_j \\
1 & \text{if an edge exists between } v_i \text{ and } v_j 
\end{cases}$$

Let us consider Figure 1, from Section 2.1.3. The adjacency matrix, $A$, for this undirected graph would be:

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

As compared to the adjacency matrix of Figure 3, from Section 2.1.3. The directed adjacency matrix, $A$, for this directed graph would be:
An adjacency matrix for a weighted graph would be modified so that

$$a_{ij} = \begin{cases} 
0 & \text{if no edge exists between } v_i \text{ and } v_j \\
 w_{ij} & \text{if an edge exists between } v_i \text{ and } v_j
\end{cases}$$

where $w_{ij}$ is the weight of the edge between vertices $i$ and $j$.

For example, consider Figure 6. The *weighted adjacency matrix* would be:

$$A = \begin{pmatrix} 
0 & 3 & 0 & 0 \\
3 & 0 & 5 & 1 \\
0 & 5 & 0 & 1 \\
0 & 1 & 1 & 0
\end{pmatrix}$$

Note, though, that weights can be normalized by dividing between the maximum weight value so they all fall between 0 and 1. Thus a normalized adjacency matrix for a weighted graph would be
\[ a_{ij} = \begin{cases} 
0 & \text{if no edge exists between } v_i \text{ and } v_j \\
\frac{w_{ij}}{\max(w_{ij})} & \text{if an edge exists between } v_i \text{ and } v_j 
\end{cases} \]

**Definition 2.1.16.** The degree matrix \( D \) of a graph \( G \) is an \( n \times n \) matrix where \( n \) is the number of vertices in the graph. As defined by [9], a degree matrix is defined as the following for the entry for vertices \( i \) and \( j \):

\[ d_{ij} = \begin{cases} 
\deg(v_i) & \text{if } i = j \\
0 & \text{otherwise}
\end{cases} \]

Thus for Figure 1, the degree matrix would be:

\[
D = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 3
\end{pmatrix}
\]

For Figure 6, the weighted degree matrix would be:

\[
D = \begin{pmatrix}
3 & 0 & 0 & 0 \\
0 & 9 & 0 & 0 \\
0 & 0 & 6 & 0 \\
0 & 0 & 0 & 2
\end{pmatrix}
\]

**Definition 2.1.17.** The distance matrix \( M \) of a graph \( G \) is an \( n \times n \) matrix where \( n \)
is the number of vertices in the graph and the entry is the length of a geodesic that connects the two nodes. Thus, considering Figure 1 again, the distance matrix would be:

$$
D = \begin{pmatrix}
0 & 2 & 1 & 3 & 2 \\
2 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 2 & 1 \\
3 & 1 & 2 & 0 & 1 \\
2 & 1 & 1 & 1 & 0
\end{pmatrix}
$$

**Definition 2.1.18.** Given a matrix $A$, the **transpose** of $A$, denoted as $A^T$, is the matrix where the rows of $A$ are the columns of $A^T$ and the columns of $A$ are the rows of $A^T$.

Consider, again, the directed adjacency matrix $A$ of Figure 3, from Section 2.1.3.

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

The **transpose** of this matrix is thus

$$
A^T = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0
\end{pmatrix}
$$

21
Definition 2.1.19. A **sparse matrix** is a matrix whose of elements are primarily zero, compared to a **dense matrix** whose elements are primarily non-zero. To calculate a matrix’s sparsity, you take the number of zero elements and divide by the number of total elements. For example, in the previous transpose matrix, there are 17 zero elements, 8 nonzero elements and 25 total elements. Thus this matrix’s sparsity is $\frac{17}{25}$, or the matrix has 68% sparsity. Thus this matrix is sparse, although not overwhelmingly so.

2.1.7 Eigenvalues and applications

Definition 2.1.20. Let $A$ be an $n \times n$ matrix. An **eigenvector**, $x$ of $A$ is a vector such that $Ax = \lambda x$ for some real or complex number $\lambda$, the **eigenvalue** of $A$. Note $Ax = \lambda x$ can be rewritten as

$$(A - \lambda I)x = 0$$

where $I$ is the $n \times n$ identity matrix. In order for $x$ to be non-zero, $A - \lambda I$ must be singular, or not invertible or, equivalently, $det(A - \lambda I) = 0$, where $det$ is the determinant of $A$.

Definition 2.1.21. For an eigenvalue $\lambda$, its **algebraic multiplicity** is defined as the multiplicity of $\lambda$ as a root of the characteristic polynomial. Its **geometric multiplicity** is the maximal number of linearly independent eigenvectors corresponding to it [10].

Definition 2.1.22. As defined in [11], the **Laplacian matrix** of a graph is the $n \times n$ matrix $L_G = L_{ij}$ where

$$L_{ij} = \begin{cases} \deg(v_i) & \text{if } i = j \\ -A_{ij} & \text{if } i \neq j \end{cases}$$
where $\text{deg}(v_i)$ is the degree of vertex $i$. This definition is equivalent to the statement $L_G = D_G - A_G$ where $D_G$ is the degree matrix and $A_G$ is the adjacency matrix of a graph $G$.

The Laplacian matrix is a discrete analog of the Laplacian operator, which behaves roughly as an averaging operator or difference. It behaves as an average rate of change. The traditional Laplace operator is continuous, but the discrete Laplace operator is an analog of the continuous version. The discrete Laplace operator thus has meaning on a graph. The Laplacian matrix is this same discrete analog of the Laplacian operator and serves a similar purpose by measuring the extent a graph differs at one vertex from its values at nearby vertices [12].

Consider Figure 6. The Laplacian matrix of this graph would be:

\[
L = \begin{pmatrix}
3 & -3 & 0 & 0 \\
-3 & 9 & -5 & -1 \\
0 & -5 & 6 & -1 \\
0 & -1 & -1 & 2
\end{pmatrix}
\]

**Theorem 1.** The **Perron-Frobenius** theorem states that if an $n \times n$ matrix has nonnegative entries, then there exists a unique, real eigenvalue $\lambda$, called the spectral radius, whose absolute value has a maximum value among all eigenvalues. This eigenvalue has multiplicity of 1, so is thus a **simple** root, and its associated eigenvectors are positive.

Note that the Perron-Frobenius Theorem implies that if $G$ is connected, then the largest eigenvalue of its adjacency matrix has multiplicity 1. For the Laplacian matrix, if $G$ is connected, then its eigenvalue 0 has multiplicity 1. Note also that for a general graph,
the multiplicity of the 0 eigenvalue of the Laplacian is equal to the number of connected components [11]. So, it follows

**Theorem 2.** A graph $G$ has $k$ connected components if and only if the algebraic multiplicity of eigenvalue 0 for the graph’s Laplacian matrix is $k$ [13].

One example should serve to demonstrate the definitions presented in this section.

Consider the following graph.

![Figure 9: Graph for Connectivity by Eigenvalue example](image)

The graph’s adjacency matrix, degree matrix, and Laplacian matrix are presented below.

$$
A = \begin{pmatrix}
0 & 0 & 3 & 2 \\
0 & 0 & 0 & 0 \\
3 & 0 & 0 & 5 \\
2 & 0 & 5 & 0
\end{pmatrix}
$$

$$
D = \begin{pmatrix}
5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 8 & 0 \\
0 & 0 & 0 & 7
\end{pmatrix}
$$
To find the characteristic polynomial, you take

\[
\det(A - \lambda I_2) = \begin{vmatrix}
5 - \lambda & 0 & -3 & -2 \\
0 & -\lambda & 0 & 0 \\
-3 & 0 & 8 - \lambda & -5 \\
-2 & 0 & -5 & 7 - \lambda
\end{vmatrix} = \lambda^4 - 20\lambda^3 + 93\lambda^2 = \lambda^2(\lambda^2 - 20\lambda + 93)
\]

Thus giving three eigenvalues for \(L\), \(\lambda = 10 + \sqrt{7}, 10 - \sqrt{7}, 0\). Note that the only eigenvalue with a greater algebraic multiplicity than 1 is \(\lambda = 0\), which has multiplicity of 2.

The associated eigenvectors of \(L\) for \(\lambda = 0\) are \(v_1 = (1, 0, 1, 1)\) and \(v_2 = (0, 1, 0, 0)\). The two eigenvectors are linearly independent, thus giving \(\lambda = 0\) a geometric multiplicity of 2.

By \(\lambda = 0\) having an algebraic multiplicity of 2, there are 2 connected components to the graph \(G\), as confirmed by Figure 9.

Note that there are numerous algorithms for computing eigenvalues of large matrices, many of which are the basis for computer programs and packages that compute eigenvalues. The Arnoldi [14] and Lanczos algorithms [15] are both iterative algorithms to find such eigenvalues based on the Power iteration [17], which also is an algorithm,
albeit incredibly more simple, that produces the eigenvalue with the greatest absolute value. The Arnoldi iteration finds the eigenvalues of general matrices, and because we are only using symmetric matrices with real entries this becomes the Lanczos algorithm. The Lanczos algorithm lacks stability because it uses floating point arithmetic, and it is different than other iterative algorithms in that it makes much better use of the information obtained by remembering all the previous computations and always operates on a vector orthogonal. In its original form, Lanczos generates an iterated sequence of linearly independent vectors, each being a linear combination of the previous vector, and the iteration automatically stops when the proper degree of the polynomial has been reached. The coefficients of the final linear combination provide the coefficients of the characteristic polynomial. This method favors large eigenvalues, as they become relatively larger than small eigenvalues by using this process, and so Lanczos modified his method so that the generated vectors were orthogonal to each other [16]. Immediately following its success, the Lanczos Algorithm was largely ignored for two decades until it, again, captured the interest of scientists. Today the Lanczos method is one of the most frequently used numerical methods and is the common product of an international community of applied mathematicians [16]. The Lanczos algorithm is computationally efficient and is used widely in a variety of fields and has numerous variations [18].

2.2 Introduction to Social Network Analysis

2.2.1 Purpose of SNA

The analysis of networks is not a recent trend [9]. Instead, the field has gained popularity recently as it has been applied to numerous subjects. As Linton Freeman describes in [19]
“The social network approach is grounded in the intuitive notion that the patterning of social ties in which actors are embedded has important consequences for those actors. Network analysts, then, seek to uncover various kinds of patterns. And they try to determine the conditions under which those patterns arise and to discover their consequences.”

Thus under the assumption of SNA, people do not act independently from the context of their social interactions, environment, and overall network. Furthermore, there are patterns of behavior that people tend to, given the understanding of this said network. Network analysis is attractive for many fields, including biology, the social sciences, and physics. As utilized here, SNA contains five key features that Freeman presents:

1. Social network analysis is motivated by a structural intuition based on ties linking social actors.

2. It is grounded heavily in systematic empirical data.

3. It draws heavily on graphic imagery.

4. It relies on the use of mathematical and/or computational models.

5. There is a wide range of empirical phenomena that can be explained in terms of their structural patterning.

The true significance of SNA, however, lies in what it is attempting to explain. While traditional analysis focuses on individuals and their characteristics or attributes, network analysis, and in particular SNA, focuses on relationships between these individuals. Where traditionalists assume independence, SNA looks at dependent observations and the shaping of individual choices and decisions based on the context of the network. Seeking to identify the organizational structure of a group, SNA uses graph theory as well as numerous other mathematical fields to study and attempt to understand social
dynamics. The ultimate goal, thus, is to be able to predict how and why actors act in these networks.

While a relationship is present between two nodes in the form of a link or edge, it is important to note that the entire network being analyzed must have something in common as well. A network of people thus usually relate to a structured situation or a group of interest. This can be anything from a hierarchical structure of a government to all Facebook users that “like” a certain musician’s page.

In order to delve deeper into any analysis or application, a few definitions, measurements and key concepts must be addressed. These definitions are borrowed primarily from [9], [6] and [20].

2.2.2 Basic Concepts and Definitions

**Definition 2.2.1.** A network is a collection of points called agents linked through some type of association. The points can represent anything, and the links can represent any relationships between the points. A network can be represented by a graph, as Figure 1 could represent a network.

**Definition 2.2.2.** A meta-matrix is an approach to network design borrowing from operations research that represents multi-dimensional data of social networks. This method displays these various relationships by integrating multiple and related network matrices into a single interrelated unit [30]. The meta-matrix enables the representation in terms of entity classes, ontologically distinct categories of agents, and relation classes, or the link between concepts within different entity classes [31].

A simple grouping of entity classes that is used in [31] [26] [34] and [37] would be the following:
1. People

2. Knowledge or resources

3. Tasks or events

4. Groups or organizations

These entity classes can be changed, but these suffice for our purposes. The usage of these classes facilitates in the systematic thinking of the overall network structure and provides a hierarchy for structuring the network of concepts [31]. The following table from [31] provides common labels for the network created by linking the entity class on the row and column.

<table>
<thead>
<tr>
<th>Meta-Matrix entities</th>
<th>People</th>
<th>Knowledge or resources</th>
<th>Tasks or events</th>
<th>Groups or organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>Social network</td>
<td>Knowledge network/resource network</td>
<td>Attendance network/assignment network</td>
<td>Membership network</td>
</tr>
<tr>
<td>Knowledge or resources</td>
<td>Information network/substitution network</td>
<td>Needs network</td>
<td>Organizational capability</td>
<td></td>
</tr>
<tr>
<td>Tasks or events</td>
<td></td>
<td>Temporal ordering/task flow/precedence</td>
<td></td>
<td>Institutional attack/support</td>
</tr>
<tr>
<td>Groups or organizations</td>
<td></td>
<td></td>
<td>Inter-organizational network</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Original meta-matrix conceptualization

Numerous relations can be formed between these entity classes. For example, in the traditional approach with people as both entity classes of the network, the relation class could be communication, friendship, etc. As previously noted, the entities themselves can be altered. So, a network could be made where a twitter account is one of the entity
classes and a hashtag is the other entity class and frequency is perhaps the weighted edge. The meta-matrix approach can be applied to other forms of analysis, like text analysis.

2.2.3 Centrality Measures

One of the main purposes of SNA is to classify and analyze patterns of ties among actors and to analyze the impact (either beneficial or constraining) of the social structure on the actors and networks [20]. Conversely, one can classify relationships among actors in a network and then analyze the impact that removing an agent within the graph would have on network as a whole. To do this, certain measures are needed. Centrality measures are used to answer the question of which nodes in the graph are the most important. For example, the degree centrality would measure how many links each agent has, but this may be limiting in graphs that are uniquely shaped. For example, as [9] does, imagine a large network like a business. We tend to think of the chief executive officer (CEO) as important, but these individuals would have a low degree centrality because they are not connected to all members of the branch, but instead just secretaries and managers. So these secretaries and individuals would have a high betweenness centrality given they lie along the communication lines of numerous people in the office. These individuals would also have a high closeness centrality given they require the fewest number of steps to others. To see who is connected to important agents or agents with many links, like the CEO may be, we need to use an eigenvector centrality. These measures are defined below.

**Definition 2.2.3.** The degree centrality of an agent $i$ within a network is defined as

$$C_{Di} = \sum_{j=1}^{n} a_{ij}$$
Thus the degree centrality is simply the sum of all the links connected to that particular agent. From a given adjacency matrix, this equates to summing across the rows [9]. To standardize this measurement and thus compare it to other networks of varying sizes of $n$, we can divide by $n - 1$ as follows

$$C'_{D_i} = \frac{1}{n - 1} \sum_{j=1}^{n} a_{ij}$$

Thus given Figure 1 and its adjacency matrix on page 24, the degree centrality of agent $c$ is 3 and its standardized degree centrality is $\frac{3}{4}$.

**Definition 2.2.4.** To calculate the betweenness centrality of an agent $i$ within a network, we first identify all geodesics in the graph. To do this, first we calculate the total possible amount of paths present, calculated by $nP_r$ for a directed graph or $nC_r$ for undirected graphs where $nP_r$ and $nC_r$ are defined by

$$nP_r = \frac{n!}{(n - r)!}$$

$$nC_r = \frac{n!}{r!(n - r)!}$$

where $n$ is the total number of vertices in the network and $r$ is the number of vertices to be considered, which in our case is always 2, thus these can be re-written as

$$nP_2 = n(n - 1)$$

$$nC_2 = \frac{n(n - 1)}{2}$$

Consider Figure 10.
The total number of possible paths to consider would thus be

\[ 4C_2 = \frac{5(5 - 1)}{2} = 6 \]

We then identify the geodesic associated with each pair of vertices. Note there may be more than one geodesic path for a pairing. After, you consider which of these geodesic paths contain agent \( i \). The **betweenness score** of an agent \( i \) within a network is defined as

\[ C_{B_i} = \frac{\sum_{i<j} g_{jk}(Agent_i)}{g_{jk}} \]

where \( g_{jk} \) is the number of geodesic paths between agents \( j \) and \( k \) and \( g_{jk}(Agent_i) \) is the number of geodesic paths between agents \( j \) and \( k \) that contain agent \( i \). Note that if agent \( i \) is the initial agent or the final agent in the geodesic this value is 0. If \( n \) geodesic paths are present between the two agents, then each geodesic is assigned a value of \( \frac{1}{n} \). Thus \( g_{jk}(Agent_i) \) is the values of these geodesics where agent \( i \) is present.

Table 2 shows the geodesic path calculation for the numerator of Agent \( B \).
<table>
<thead>
<tr>
<th>Starting vertex</th>
<th>Ending vertex</th>
<th>Geodesic path(s)</th>
<th>Value of Agent $B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>(A,B)</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>(A,B,C)</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>(A,B,D)</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>(B,A)</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>(B,C)</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>D</td>
<td>(B,D)</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>(C,B,A)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>B</td>
<td>(C,B)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>(C,D)</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>A</td>
<td>(D,B,A)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>B</td>
<td>(D,B)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>(D,C)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Geodesic path value for Agent $B$

Note if the path has already been analyzed, even if the vertices are in a different order, the count is not considered for calculating the betweenness centrality. Thus the numerator for the betweenness centrality of Agent B is 2. For the denominator, we calculate the number of pairs of agents that do not include $n$, and so we are either permuting or combining $(n-1)$ two at a time. Thus the standardized measure for betweenness centrality becomes

$$C'_{Bi} = \frac{C_{Bi}}{(n-1)C_2}$$

which for Agent B from Figure 10 is $\frac{2}{3C_2}$ or $\frac{2}{3}$.

**Definition 2.2.5.** The **closeness centrality** of a agent, as defined in [9] is defined as

$$C_{Ci} = \left[ \sum_{j=1}^{n} d(n_i, n_j) \right]^{-1}$$

where $d(n_i, n_j)$ represents the geodesic distance between agents $i$ and $j$. This measure
is used to define the \textit{closeness} of agent $i$ relative to all other agents in the network. Noting that $\lim_{x \to \infty} \frac{1}{x} = 0$, $C_C$, will reach a maximum at $(n - 1)^{-1}$, noting that when one agent is adjacent to all other actors in the network and will reach a minimum when one agent in the network is not \textbf{reachable}, or if there is no path that exists to join the two agents. When this path does not exist, $d(n_i, n_j) = \infty$. This makes closeness only meaningful when applied to connected networks or to the connected components of a network.

Closeness can be standardized by the following equation

$$C'_{C_i} = \frac{(n - 1)}{\sum_{j=1}^{n} d(n_i, n_j)}$$

This standardization results in an index that ranges from 0 to 1. The inverse of this measurement is referred to as the \textbf{average path length}. To calculate the closeness centrality, we thus must calculate the average path length for every pair of agents.

Consider Figure 11.

We must first, again, calculate the list of geodesic paths present in the graph, as shown in Table 3.

We now must calculate the average path length to every other agent. So, for example, Agent D has geodesics: $(D,E,A)$, $(D,E,B)$, $(D,C)$, $(D,E)$ with lengths 2, 2, 1, 1, respectively. Therefore, its average path length is

$$\frac{2 + 2 + 1 + 1}{(n - 1)} = \frac{6}{4} = \frac{3}{2}$$
Figure 11: Example graph for calculating the closeness centrality

<table>
<thead>
<tr>
<th>Path number</th>
<th>Starting vertex</th>
<th>Ending vertex</th>
<th>Geodesic path(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>B</td>
<td>(A,B)</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>C</td>
<td>(A,E,C)</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>D</td>
<td>(A,E,D)</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>E</td>
<td>(A,E)</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>A</td>
<td>(B,A)</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>C</td>
<td>(B,E,C)</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>D</td>
<td>(B,E,D)</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>E</td>
<td>(B,E)</td>
</tr>
<tr>
<td>9</td>
<td>C</td>
<td>A</td>
<td>(C,E,A)</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
<td>B</td>
<td>(C,E,B)</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>D</td>
<td>(C,D)</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>E</td>
<td>(C,E)</td>
</tr>
<tr>
<td>13</td>
<td>D</td>
<td>A</td>
<td>(D,E,A)</td>
</tr>
<tr>
<td>14</td>
<td>D</td>
<td>B</td>
<td>(D,E,B)</td>
</tr>
<tr>
<td>15</td>
<td>D</td>
<td>C</td>
<td>(D,C)</td>
</tr>
<tr>
<td>16</td>
<td>D</td>
<td>E</td>
<td>(D,E)</td>
</tr>
<tr>
<td>17</td>
<td>E</td>
<td>A</td>
<td>(E,A)</td>
</tr>
<tr>
<td>18</td>
<td>E</td>
<td>B</td>
<td>(E,B)</td>
</tr>
<tr>
<td>19</td>
<td>E</td>
<td>C</td>
<td>(E,C)</td>
</tr>
<tr>
<td>20</td>
<td>E</td>
<td>D</td>
<td>(E,D)</td>
</tr>
</tbody>
</table>

Table 3: List of Geodesic Paths

If there were two geodesics between two agents, its path length would be the length of one of those geodesics. The lower this number, the lower the average path length, which
implies higher closeness. So we take the inverse of this average, and Agent D has
closeness centrality of \( C'_{CD} = \frac{1}{2} = \frac{2}{3} \).

Compare Agent D to Agent E. Using the same method, we find that the average path
length of Agent E is

\[
\frac{1 + 1 + 1 + 1}{n - 1} = \frac{4}{4} = 1
\]

So its closeness centrality is \( C'_{CE} = \frac{1}{1} = 1 \), a perfect centrality score, implying that Agent
E is the most central that a agent can be within a network.

**Definition 2.2.6.** The **eigenvector centrality** is a way to measure how connected an
agent is to other greatly connected agents. It essentially denotes the “extent to which an
individual is a big fish connected with other big fish in a big pond” [21]. It assigns
relative scores to all agents based on the assumption that connections to higher-scoring
agents contribute more to the score of an agent than the same amount of connections,
but to low scoring nodes. As defined by [9], the eigenvector centrality of an agent \( i \) is
calculated by

\[
x_i = k \sum_{j \in C_i} x_j = k \sum_{j=1}^{N} a_{ij} x_j
\]

where this denotes the eigenvector centrality of the \( i \)th agent of a network, \( a_{ij} \) is the
adjacency matrix, \( C_i \) is the set of agents connected to \( i \) and \( k \) is the proportionality
constant. We choose \( k \) so that \( k = \frac{1}{\lambda} \), and thus we have

\[
x_i = \frac{1}{\lambda} \sum_{j=1}^{N} a_{ij} x_j
\]
which can be rewritten in vector notation, and yields the familiar eigenvalue problem

\[ Ax = \lambda x \]

Typically there are many values for \( \lambda \) that solve this problem, and so there is an additional requirement that all values of \( x \) be positive and we desire the greatest eigenvalue. Thus with these additional requirements, the \( i \)th element of the eigenvector \( x \) gives the eigenvalue centrality of the \( i \)th agent in the network. However, by requiring that all entries in the eigenvector be positive, by Theorem 1 or the Perron-Frobenius theorem, only the greatest eigenvalue results in the desired centrality measure [22]. With these requirements satisfied, the \( i \)th element of the related eigenvector gives the eigenvalue centrality of the \( i \)th agent in the network.

Consider the following weighted graph.

![Weighted Graph](image)

Figure 12: A weighted graph for Eigenvector Centrality example

The adjacency matrix \( A \) of this graph would be
The characteristic polynomial of $A$ is given by

$$
\begin{vmatrix}
-\lambda & 2 & 3 \\
2 & -\lambda & 5 \\
3 & 5 & -\lambda
\end{vmatrix} = -\lambda \begin{vmatrix}
-\lambda & 5 \\
3 & -\lambda
\end{vmatrix} + 3 \begin{vmatrix}
2 & -\lambda \\
5 & -\lambda
\end{vmatrix} = \lambda^3 + 38\lambda + 60
$$

Solving for 0, the eigenvalues are $\lambda \approx -5.128, -1.711, 6.839$. Choosing the greatest of these values, 6.839, and appending a column of zeros gives the “rank-deficient” augmented matrix

$$(A - 6.839I)x = 0$$

or

$$
\begin{bmatrix}
-6.839 & 2 & 3 & 0 \\
2 & -6.839 & 5 & 0 \\
3 & 5 & -6.839 & 0
\end{bmatrix}
$$
Through basic row reductions and elementary algebra, we can find the eigenvector

\begin{equation}
\begin{bmatrix}
  a \\
  b \\
  c \\
\end{bmatrix} = k 
\begin{bmatrix}
  1 \\
  1.317 \\
  1.402 \\
\end{bmatrix}
\end{equation}

where \( k \) is any arbitrary scalar. We choose \( k \) so that the highest eigenvector centrality is 1, thus normalizing the measure, so for \( k = \frac{1}{1.402} \), the eigenvector contains the eigenvector centralities \( x = (0.713, 0.939, 1) \). Thus the most central agent by this centrality measure is agent \( c \), and the least central is \( a \).

**Definition 2.2.7.** The **cognitive load** of an agent \( i \) denoted \( L_i \) as defined by [23] uses the meta-matrix approach to network analysis that was previously described in Section 2.2.2. This particular measure looks at the networks formed by Agent-Task networks, \( AT \), and Resource-Task networks, \( RT \), where \( A \) is the agent network, \( T \) is the task network and \( R \) is the resource network.

For a particular meta-matrix network, this allows \( ATR = AT \ast RT^T \) and \( ATA = AT \ast AT^T \), where the superscript \( T \) denotes the transpose of the matrix.

Let \( x_1, x_2, x_3, x_4, x_5, x_6 \) be denoted as:

\begin{align*}
  x_1 &= \frac{\text{deg}(A_i)}{|A|} \\
  x_2 &= \frac{\text{deg}(AT_i)}{|T|} \\
  x_3 &= \frac{\text{deg}(ATA_i)}{(|A| - 1)(|T|)} \\
  x_4 &= \frac{\text{deg}(AR_i)}{|R|}
\end{align*}
\[ x_5 = \frac{\text{deg}(ATR_i)}{|T| \times |R|} \]
\[ x_6 = \frac{\text{deg}(AR_i) - \text{deg}(ATR_i)}{|R| \times |T|} \]

Where \( x_1 \) is essentially measuring the degree centrality of agent \( i \), \( x_2 \) is measuring the proportion of tasks that agent \( i \) is assigned to out of all possible tasks, \( x_3 \) is measuring the amount of agents who do the same tasks as agent \( i \) standardized for the total amount of tasks and agents, \( x_4 \) is measuring the proportion of resources \( i \) manages, \( x_5 \) is measuring the amount of resources that \( i \) needs to manage its tasks standardized for the total amount of tasks and resources and \( x_6 \) is measuring the amount of negotiation needs \( i \) must do for each task standardized for the total amount of possible negotiations.

Thus the cognitive load for agent \( i \) is

\[ L_i = \frac{(x_1 + x_2 + x_3 + x_4 + x_5 + x_6)}{6} \]

This measurement is obviously quite complex and an example does not necessarily foster increased understanding, but it is important to note because it has been regularly used in prior research. The purpose of cognitive load is to measure the total amount of effort expended by each agent to do their tasks, so it is more applicable when analyzing graphs where each agent is a person and we are trying to understand the dynamics of that network. An example where cognitive load would be important is graphing leadership structures of states or the amount that different people are involved in terrorist attacks.
2.3 Text Analysis

Across many fields, interest is increasing in using text as a computational resource as compared to the standard data typically used to address questions in the social sciences [27]. Text analysis is a tool for discovery and measurement by uncovering patterns of language use, interpreted as prevalent attitudes, concepts, or events [27]. Network text analysis, or the crossover between network analysis and text analysis, is used to describe a wide variety of “computer supported solutions” that allow analysis through extracting networks of concepts from texts [37]. NTA operates under the assumption that words in a text may be modeled as a network. Generally, NTA assigns words to categories and then determines if there is a relationship between the words by forming a link. This linking forms a network, or a graph, as was previously explained. This method has numerous applications, from determining if sentiment in the media affects the stock market or how racial attitudes affects voting patterns. To understand how text analysis can be applied to SNA, as is done in NTA, a basic level of understanding of the process behind text analysis must occur.

2.3.1 Basic Definitions

A document is a collection of words bound together by some standard. A document can be anything, from a chapter of a book to the book itself. It is predetermined by the question trying to be answered, but must have some feature that binds the words.

A corpus is a collection of documents used for data. For the research presented herein, a corpus will be a collection of tweets where each tweet is a document.
document \( j \) can be represented by the vector

\[
x_j = (x_{j1}, x_{j2}, \ldots, x_{jk})
\]

where

\[
x_{jk} = \begin{cases} 
1 & \text{if } k \text{ is in } j \\
0 & \text{otherwise}
\end{cases}
\]

where each \( x_{jk} \) is a unique term in the corpus and \( k \) is the number of unique terms in the corpus, thus each document considered in the corpus has the same number of dimensions [24].

2.3.2 Parsing Text

In order to analyze text, particularly when there is a lot of it, we must parse or clean the data. By parsing it, we are changing the text so that it is most readable and processable by a computer. Parsing text can occur in many phases or steps, and can range from simply changing all text to lowercase so that \( Cat \) wouldn’t be read or processed as a different word than \( cat \), or parsing can be complex and involve highly complicated stemming algorithms.

**Definition 2.3.1. Tokenization** is the process of segmenting running text into words and sentences, or into tokens. A token must be linguistically significant and methodologically useful [25]. This does not solely rely on the ability to recognize spaces or other delimiters on either side of the word, but rather a sense of pattern recognition. Consider, for example, the following text:
where is meadows dr who asked

The word “dr” could refer to a Doctor, so “Doctor Who” or it could refer to a drive, so “Meadows Drive.” This discrepancy relies on collocation recognition and overall pattern recognition [25].

Definition 2.3.2. Stemming is the process of reducing a word or token into its word stem, or its base or root form. This stem does not have to be an actual word itself, but rather all related words simply need to map to the same stem. For example, argu is the stem form of arguing, argue, argued, argues.

Definition 2.3.3. A thesaurus is a two-columned matrix that links text-level words to higher level concepts. For example, the text dog play fetch, the word dog would be classified as an actor, where play and fetch could be classified as roles for the actor or jobs [37]. This approach to parsing text is particularly useful in a meta-matrix approach so then words have more meaning when considered in their class.

Definition 2.3.4. Stop words are words that are removed from text data before analysis or processing, and can be anything. Most frequently, these words are common words that hold little value as far as what the text is saying. For example, a, and, at, for, he, in, that, with are all in the basic level stop word package for most text analysis software, but this list can be modified for your own purposes. For example, in the research presented herein, available and via were both added to this list due to their frequency in Tweets without adding anything to the interpretability of the message. RT could also be added to the list of stop words if one isn’t concerned with who re-tweeted whom, and so stop words can change given the task at hand, even with the same data.
2.3.3 Representing and Analyzing Documents

**Definition 2.3.5. Sentiment analysis** is the identification of opinion through text analysis and computational linguistics. Through a complex framework involving numerous algorithms and large data, a classification tool is thus created to detect Positive, Negative, or Neutral sentiments in a given document.

2.4 Twitter Application

Twitter is increasingly being used as a platform to gather data for text analysis. It’s very popular, thus there are a huge volume of documents to analyze on a limitless number of topics, and it is easily accessible by creating a Twitter app and obtaining an API. Due to its ease of use and ability to create numerous profiles and tweets, ISIS has used Twitter to spread its message and, specifically, recruit in areas it may have not been able to otherwise. Using tweets as documents for NTA also gives large flexibility because each document is short (there is a 140 character limit) and the sheer volume of tweets. Tweets also provided information far beyond solely the text of the document. They also give information as to who “follows” whom, who “re-tweets” whom, at times the location of that tweet and also profile information, like the gender of the user.

2.4.1 R

While it would be ideal to sift through each document individually to analyze, much like how a person gains information by reading a book, the amount of data used makes a computer program necessary to try and stay current on tweets and data which is constantly increasing. R is a free-software environment for statistical computing and
graphs [28]. It runs on a wide variety of platforms, and is widely used among statisticians and data analyzers [29]. R’s usage has increased in recent years [29], and so there are numerous open-source codes available for similar text analysis projects and other large data problems. How frequently R is used, too, ensures reliability and that codes and the tool in general stays up to date on changes made by Twitter or other social media platforms.

3 Prior Literature and Work

There is no doubt that the twenty-first century and the parallel influx of technology that has occurred changes almost everything. As the world moves into a more globalized, interdependent society, security concerns are generally moving away from state vs. state conflict and towards the threat of terrorist organizations. These decentralized networks typically are not geographically oriented, but are instead focused primarily on an ideology [1]. Due to their covert nature, the U.S. can no longer use traditional, hierarchical tactics and must instead uncover new strategies. Social network analysis, although a new field, has been the primary tool used to attempt to destabilize terrorist networks as well as mapping the various networks to produce a clearer understanding of what these networks look like and, hopefully, how one can fight them. Current research using social network analysis to analyze terrorism is limited in quantity, but has exponentially increased in recent years and shows promise for further work.

3.1 Framework for Application

Most literature on the topic doesn’t provide examples of actually using social network analysis to map terrorist organizations, but more often provides a framework and
theoretical base. Dr. Kathleen M. Carley, a professor of computer science at Carnegie
Mellon, has written much of this literature. In [26] Carley uses dynamic network analysis
(DNA) as the primary technique. This method evaluates destabilization strategies for
networks that are currently evolving and that also typically do not have all information
provided. DNA “extends the power of thinking about networks to the realm of large
scale, dynamic systems with multiple co-evolving networks under conditions of
information uncertainty with cognitively realistic agents” [26]. This paper uses the
meta-matrix approach, as previously defined, treats ties as “variable,” thus having a
probability associated with them, and combines social networks with cognitive science
giving agents, or nodes, the ability to adapt.

Carley proposes DNA as a better way to destabilize terrorist groups, given that they do
not follow the typical hierarchical structure. Instead, terrorist networks are cellular and
distributed. By destabilizing the network, one can more easily fight the network as a
whole and hopefully the destabilization is done in an efficient manner so that time isn’t
wasted. To destabilize the network, Carley proposes to identify key players by a few
criteria, as established below, and connections among the players, then to identify the
process by which connections are added or dropped, and how that changes the strength
of the players. Next, a collection of data would ensue so that the performance of the
existing network as well as the performance of an optimal system by location of
vulnerabilities, where performance is a measured by using DyNet, a multi-agent network
analysis program that assesses destabilization strategies on changing networks. Finally,
one can test destabilization strategies on the network.

After the primary background steps are completed, one of the four strategies of
destabilization must be chosen. These four strategies are to eliminate the person with
the highest degree centrality (the number of edges that are connected to that node),
betweenness centrality (the number of shortest paths from all vertices to all others that
pass through that particular node), cognitive load (the node with the most possible connections), or task exclusivity (determines the uniqueness of the connections of that node). In [26], Carley uses data collected on an embassy bombing in Tanzania, and yields mixed results using the four proposed strategies. For example, if judged solely by efficiency it seems that removing one node increases efficiency, but if judged by the flow of communication, a different node is more disruptive. Thus, what exactly is optimal is unclear, and so the researcher cannot draw significant conclusions.

Carley provides a table that shows the results of the removal of agent 5 of the network. All differences are significant, but none of the removals substantially change the shape of the optimal design. The removal of agent 5 reduces the presence of unnecessary resources thus making the organizational design "leaner" [26]. While the removal thus does make the organization more efficient, it also makes it less adaptive. The table from [26] is shown below in Table 4.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Original Design</th>
<th>After Removal of 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamming from Optimal</td>
<td>88</td>
<td>83</td>
</tr>
<tr>
<td>Resource congruence</td>
<td>.475</td>
<td>.525</td>
</tr>
<tr>
<td>Performance as Accuracy - Initial Impact</td>
<td>78.5625</td>
<td>78.22</td>
</tr>
<tr>
<td>Performance Recovery - Percentage Increase in Performance</td>
<td>95.55</td>
<td>89.72</td>
</tr>
<tr>
<td>Diffusion - Initial</td>
<td>21.62291</td>
<td>14.70212</td>
</tr>
<tr>
<td>Diffusion Recovery - Percentage Increase in Diffusion</td>
<td>71.23304</td>
<td>89.05325</td>
</tr>
</tbody>
</table>

Table 4: Impact of Agent Removal from [26]

Carley relies heavily on two software tools. The first, [32], is a meta-network assessment and analysis tool used to contrast the resource congruence, or, as defined by [34], the ability of an agent to match what resources were needed for a task and the availability of those resources in the network, of the organization with and without the individual.

The second, [33], is a multi-agent network system (a loosely coupled network of software
agents that interact to solve problems beyond the scope of one software or individual [35]), which assesses the destabilization strategies on the changing networks, and is used here to contrast relative change in performance of the network with and without the node being removed. While Carley’s method does provide insight, to actually use this method and gain significant conclusions, a clearer definition of optimization must be used including what criteria to consider. Furthermore, given the need of these organizations to stay secretive, a key concern is the reality that false information is probably being flowed more often than correct information, and so any results from the analysis must be taken in such context.

Carley also wrote a paper with Jana Diesner [37] on predicting organizational structure of a network by using network text analysis. In [37], Carley demonstrates the process of determining the structure of a network through a given set of texts where the covert networks are represented in the texts. Carley uses Network Text Analysis (NTA), a method for encoding the relationships of words into a structured network of those linked words that has been previously discussed in greater detail.

In this particular paper, map analysis is used to determine the relationships between words and concepts in NTA. By assigning texts to maps, one can better discern the meaning of the word. In map analysis, a concept (a word or group of words) is equivalent to a node and the link between the two concepts, a statement, is equivalent to an edge. The union of all statements within the text forms a map, which is equivalent to a network [37]. The data used is provided by 247 texts collected by the Center for Computational Analysis of Social and Organizational Systems of Carnegie Mellon (CASOS), and then is analyzed by the same programs as previously mentioned ([32] and [33]), and also by AutoMap [36]. AutoMap is a software tool that uses map analysis, which systematically extracts links between words in a text to create a “mental map” as a network of these words [37].
This technique is not unique to just Carley and Diesner’s paper, [37]. Furthermore, in [37], the maps tend to be terribly complex. Thus to facilitate understanding of the basics of map analysis, a simple example is borrowed from "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis," also by Carley, [38] based on the following two passages:

1. Joe’s a gnerd who always studies in the library.

2. Joe’s a gnerd who always studies in the library and doesn’t fit in. His door is never open, he works too hard.

The maps were then created based on whether there is an “implicit semantic connection between the concepts,” so that the connection is present if the text reads “if \( a \) then \( b \)” or “\( a \) (verb) \( b \).”

As shown in Figure 14, the simple coding style used tends to “generalize relationships and emphasize the degree of similarity between the texts” [38]. This mapping also uses two types of connection, a positive relationship denoted by the solid line or a negative relationship denoted by the dashed line. Obvious precautions must be taken when using this sort of analysis in order to not draw conclusions where none exist, like the conclusion from Map 2 that if someone studies in the library, then his/her door is never open.
The use of NTA rests on the creation of two thesauri. The creation of the first type of thesaurus is essentially a clustering technique, as was previously defined in Section 2.3. These groupings contain co-occurring terms and thus typically have some relevance [39]. In general, a thesaurus is a two-columned collection that “associates text-level concepts with higher-level concepts” [37]. The text-level concepts represent simply the content of the data set being analyzed and the higher-level concepts represent the text-level concepts, but in a more generalized way. From this general thesaurus a meta-matrix thesaurus is created that links the concepts into the entity classes used in the meta-matrix. Note this thesaurus may be larger than two columns because a concept may fall into multiple entity classes.

In [37] and [31], the text was first tokenized and stemmed. This was done by translating phrases into a single unit that is recognizable, so a terrorist’s name would be changed into the group as well as changing common misspellings of concepts into one recognizable concept, like Hizbullah into Hezbollah. The meta-matrix thesaurus created is too large to show for our purposes. The actual thesaurus in [31] contained 2083 concepts and numerous classes, but Carley provides a frequency table of the thesaurus as shown below in Table 5.

<table>
<thead>
<tr>
<th>Category</th>
<th>Accumulated sum of assignment of concepts to the entity classes in the meta-matrix thesaurus</th>
<th>Cumulated sum of appearance of entity classes in texts after application of meta-matrix thesaurus</th>
<th>Cumulated sum of linkage of concepts associated with entity classes into statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>48</td>
<td>569</td>
<td>434</td>
</tr>
<tr>
<td>Location</td>
<td>81</td>
<td>404</td>
<td>404</td>
</tr>
<tr>
<td>Agent</td>
<td>54</td>
<td>250</td>
<td>217</td>
</tr>
<tr>
<td>Resource</td>
<td>75</td>
<td>261</td>
<td>188</td>
</tr>
<tr>
<td>Task-Event</td>
<td>27</td>
<td>168</td>
<td>146</td>
</tr>
<tr>
<td>Knowledge</td>
<td>41</td>
<td>134</td>
<td>128</td>
</tr>
</tbody>
</table>

Table 5: Frequency table of meta-matrix thesaurus from [31]
[33] and [32] provided visualizations of the meta-matrix thesaurus, and an example of the sub-matrices from the text is presented below in Figure 14.

![Figure 14: Example visualization of meta-matrix thesaurus from [31]](image)

This visualization is rather disconnected, as there is not a lot of overlap between concepts and class entities. This visualization also tells little about the structure of the terrorist group.

Key actors were determined by their cognitive load, degree centrality, and betweenness centrality, as previously defined in section 2.2.3. The graphs provided are based on general text data of the network, and so after the Khobar Tower Bombing in Saudi Arabia in 1994, the connections shown in the graph give who was significantly connected to any other person and thus can determine who had significant ties with whom were arrested.

This application of NTA could yield significant results like the example used within the paper, but has its downfalls as well. The process used is lengthy and also not very interpretable, and a comparison is needed given empirical data later on. A significant concern again, though, comes from the validity of the text analyzed and also text availability. A covert network’s top priority is often secrecy, and so text that is truthful and has enough substance to draw conclusions from is rare. Furthermore, this text
becoming available is likely available only significantly after an event, and so this method loses the predictive ability that other SNA methods hold.

3.2 Previous Applications

Of the articles published that actually provided an example of mapping a terrorist cell, Valdis Krebs provided the clearest of these in [40]. Krebs was able to map the terrorists and their contacts involved in the 9/11 attack using public information and open-source data, largely using the New York Times, the Wall Street Journal and the Washington Post for his sources. Following 9/11 an obvious shift occurred in newspapers and media at large in that everything seemed to shift focus to the attack. Thus matrices of the attackers, including which flights they were on, was released. On this matrices also was an inclusion of how these people were connected. An example of this is provided in Figure 15, a very preliminary matrix that was published on the web site of the Sydney Morning Herald.

Using this information, Krebs was able to make a preliminary network graph that connected the actors, or nodes, by their similarities. Krebs defined initial similarities as to their prior trusted contacts, the ties formed through living and working together [40]. After more information surfaced by other trusted news sources, like the Washington Post, Krebs was able to create two new graphs. The first of which was a network graph where ties, or edges, represented prior trusted contacts and the second of which built on the first graph and added a variable representing if the two hijackers met prior to the actual attack on 9/11, and thus a weighted graph was created with two variables. The final network graph Krebs used had three levels of tie strength based on how much time was spent together prior to the attack and also on that day. The final graph Krebs created is presented in Figure 16.
Krebs’ techniques were actually not difficult at all. Instead, Krebs created simple graphs that contained a weighting system that solely utilized data from major newspapers. Krebs hypothesized various shortcuts that could have occurred after information surfaced of a publicized meeting in Las Vegas between the members, but these shortcuts were mere speculation. Furthermore, a finalized network neighborhood graph was created where the pilots seemed to be the central actors, but this too lacks a completeness of information.

Krebs primary finding was that covert networks are not efficient by any means and often trade efficiency in order to ensure secrecy, which could be the reason that these graphs are often in a strange shape instead of the typical hierarchical structure associated with states. Typically a hierarchical graph would be resemble Figure 17, and as one can see this varies greatly from Krebs’ graph that he notes is in the shape of a “serpent” [40].

Furthermore, the hijacker’s network had massive redundancy in that the relationship of
each actor was formed at an early childhood age. This, and that their training was primarily in Afghanistan, made this network resilient, or able to morph back into shape even after being destabilized. While Krebs does draw significant conclusions through his social network analysis, his findings can only be applied to prosecution instead of the prevention of attacks. This information may be useful if other covert networks act in a similar manner, but the immediate applications are not clear because each terrorist group attempts to act in a way that is unpredictable.

Stuart Koschade was able to accomplish similar goals in mapping the Jemaah Islamiyah cell responsible for the Bali bombings of 2002 [43]. Koschade provides a general framework as to how one can actually apply SNA to mapping a real terrorist group, although his framework, like Krebs’, focuses on the need for publicly available relational data. Without such data, little can be done. Koschade used SMS texts that the
members sent between each other and thus was able to establish two adjacency matrices, one representing simply a binary relationship indicating a relationship being present between the two or not, and the second of which took into account two interactional criteria: frequency and duration of content and transactional content and thus established an ordinal level ranking system, based on frequency and duration.

Koschade used visualization software like UCINet [44] to create graphs based off of these adjacency matrices, and then characterized the network graph by seven characteristics: size, density, degree of connection, centrality, closeness, betweenness, and clusters. Formulas are presented in this article, but the data needed for actually using any of these formulas is immense. In contrast with Krebs, however, the final graph yielded is actually quite centralized, as shown below in Figure 18.

The graph being so central could be in part due to the fact that there were 17 actors
present, a larger network size than was really needed for this operation [43]. At least four actors did not play a role necessarily in the operation at all, thus indicating that efficiency was not a priority for Jemaah Islamiyah like it was for Al Qaeda. This may also indicate that Jemaah Islamiyah had made arrangements for contingency plans if the cell were to be disrupted.

Koschade is able to reach incredible conclusions given his methodology, including who the central actors were in the attack, the priority of effectiveness over secrecy, the member that was the weakest point and thus whose capture would have led to the isolation of the most active portion of the network, and also that covert networks have the ability to heal themselves in the event of the loss of a singular node. The primary weakness in Koschade’s findings as well as methodology mirrors the weaknesses found in Krebs’. Both methods need an enormous amount of verified data, but this data probably isn’t available until a significant amount of time after the event. This makes both
applications insignificant in their predictive ability, unless for some reason the terrorist group makes similar information public thus going against their typical goal of covertness. Thus although limited, the examples of SNA being applied to real terrorist groups do exist, although at a pace that makes these applications irrelevant for counterterrorism attempts.

3.3 Twitter and ISIS

3.3.1 Why Use Twitter?

The Islamic State of Iraq and the Levant, ISIS, is an Islamic extremist group that has gained overwhelming attention in the past few years. For the past year, however, policymakers and the media have been engrossed by ISIS’ successful use of social media, particularly Twitter [45]. ISIS uses numerous techniques to exploit its popularity through Twitter, notably using an Arabic language Twitter app “The Dawn of Glad Tidings” as well as “organized hashtag campaigns” [46]. ISIS’s use of Twitter serves as a platform to promote strength and engagement within their community of supports, and it has thus far been quite effective. A study done in February of 2014 found over 10,000 mentions of a particular hashtag in a single day while ISIS’s biggest competitor only managed to receive at most 5,000 [46].

The media draw to ISIS’s unprecedented use of technology to recruit and encourage allegiance simply adds to the cyclical nature, particularly the focus placed on Twitter. Twitter, a social media site, allows users to send messages of a maximum of 140 characters and is created in a way that focuses on keywords, like hashtags. The platform that Twitter represents is thus perfect for Network Text Analysis, a method that seeks to find key words and concepts through a multitude of information. Previous findings of the
specific effectiveness of ISIS when using Twitter is limited, but it is also a relatively new concept. Through the twenty-first century given the probable continued rise in technology and media, more terrorist groups may see ISIS’s success of using social media to gain support and power and may follow suit. Thus it is important to understand just how and why ISIS uses Twitter and if it changes through their maturing an an organization.

3.3.2 Qualitative Analysis of ISIS’s Use of Twitter

For states and throughout the traditional perspective use of international relations, states seek recognition by diplomatic initiatives and incentives that result in either the recognition of a state’s regime or a state’s statehood. The ultimate goal is accomplishing recognition on both standards, but in the case of new states this may not be the case. This Eurocentric view uses nation-states as the actors, and sees international relations as the transfer of power from actor to actor as the primary movement that shifts the system. While this analysis is clean, it does not completely describe the current state of the international spectrum. Instead of states being the sole actor that this viewpoint claims is the case, there are numerous outside groups that are now attempting to gain recognition and power like terrorist groups. Twitter offers the perfect platform for ISIS and other groups to “recruit, radicalize and raise funds” that otherwise may not be available [46].

Berger describes ISIS use of Twitter as a twofold technique in [46]. ISIS uses the app The Dawn of Glad Tidings as previously mentioned as well as a methodological hashtag process that yields trending results. The Dawn of Glad Tidings, nicknamed as Dawn, is advertised as a way to stay up to date on the jihadi group [46]. The app, which is available on Android phones through the Google Play store, is able to post tweets to an
account after the account enrolls in the program while maintaining a certain amount of space between tweets to avoid Twitter’s spam-detection algorithm. The Twitter account enrolled functions as it normally would except for these postings, thus making it a minimal commitment. The app was first created in April 2014, and has since been able to post 40,000 tweets in one day as ISIS “marched into the northern Iraqi city of Mosul last week” declaring “We are coming, Baghdad” [46].

As Berger notes, the sheer volume of the tweets posted were enough to result in any search of “Baghdad” on Twitter immediately result in the image posted by Dawn. The image, shown below in Figure 19, would certainly provoke fear in the area that ISIS was moving towards, thus giving ISIS more power due to its growing perception.

![Figure 19: Image posted by Dawn on numerous Twitter accounts warning of Baghdad invasion, from [46]](image)

ISIS does not solely use Dawn and apps to promote its content. As previously mentioned, ISIS also uses a calculated hashtag campaign to encourage participation and publicize its message. The group does this by volunteering hundreds and sometimes
thousands of group members to repeatedly tweet a hashtag so they trend on social media sites [46]. An Arabic Twitter account @ActiveHashtags often falls prey to this method and unintentionally promotes ISIS’s materials. When a hashtag is recognized by @ActiveHashtags, it is retweeted on average 72 times per initial tweet which continues the cyclical nature by assuring that the hashtag gains even more popularity and momentum [46]. As the hashtag gains popularity, more users are exposed to ISIS’s message. The exposure is not confined to a particular region, either, because Twitter is an international social media platform, so ISIS is able to gain informal recognition and thus power in areas it might not otherwise.

Twitter can also serve as an environment to test public opinion on decisions or “policies” that ISIS may use. For example, in early 2014 there was talk of changing the name of ISIS to signal the rebirth of an Islamic caliphate as opposed to the Islamic state it is currently attempting to create [46]. ISIS was able to track public opinion of the controversial suggestion by creating what appeared to be a grassroots like initiative hashtag demanding the path towards establishing a caliphate. The proposal, as expected, was responded with enormous “angry and divisive discussion” that ISIS probably used given it has not since announced the pursuance of such caliphate [46].

Twitter thus has obvious appeal to non-state actors as an alternative mean to gather strength and power within the international community.

3.3.3 Quantitative Analysis of ISIS’s Use of Twitter

The Brookings Institution, a reputable US based think tank, published an elaborate report, [47], earlier this year that attempted to answer how ISIS is spreading its message by analyzing the population of ISIS’s supporters on Twitter. Among numerous findings, the Institution founded that from September to December of 2014, at least 46,000
Twitter accounts were used by ISIS supports, although not all were active at the same time [47]. Most of these Twitter accounts were located in Syria and Iraq and other territories claimed by the organization, although a significant number were also located in the United States and outside ISIS’s territory, as shown in Figure 20.

![Location Claimed in Profile](image)

**Figure 20: Location claimed in ISIS supporter profiles from [47]**

Among the language preferences selected of ISIS supporting accounts, three quarters selected Arabic while one in five selected English. These accounts also had significantly more followers than a typical account, each averaging about 1,000 followers each. Twitter suspended at least 1,000 of these accounts between the months studied, but those with the most followers were most likely to get suspended ignoring the majority of the still active accounts [47].

What this investigation found that is probably most interesting, however, is that much of ISIS’s success on social media and specifically Twitter can be attributed to a relatively small group of users, most likely between 500 and 2,000 accounts, which tweet in hyperactive bursts, as shown in Figure 21.

The list of findings included in the extensive 70 page report clearly show the success of ISIS’s techniques. The authors of the report recommended social media companies and US state officials work together to counter the successful extremism on Twitter and other social media sites [47].
3.3.4 Attempts to Counter

With so much media and public attention being drawn to ISIS’s use of Twitter, it is obvious that there have been efforts to counter ISIS’s success. As described in [48], on March 31, 2015, an anonymously created list of more than 25,000 Twitter accounts was released to the public. The release is aligned with growing campaigns carried out by the hacktivist collective Anonymous that hopes by providing the public with these accounts will increase awareness of ISIS’s use of Twitter as a platform for recruitment [48].

The database was created by using data mining tools not detailed in the article, although at the time the database was shown to International Business Times UK the database was tracking 26,374 accounts, 10,408 of which were still active [48]. ISIS is able in part to maintain a steady following on Twitter despite accounts being constantly deleted by creating usernames that are very similar to prior known names used. For example, an account could be the exact same name as an old account with the inclusion of the number ‘1’ at the end. The database is said to be created autonomously by inputing prior known account names that are associated with ISIS and sifting through recently created account names in an attempt to isolate ones associated with ISIS [48].
Although the programmer responsible for the creation of the database wishes to remain anonymous, the list was published through the cyber counter-terrorism activist XRSone, who was responsible for a release of 9,200 accounts earlier in March of 2015 [49]. XRSone describes the network that ISIS’s Twitter accounts creates as “hydra-like” and adaptable given that despite accounts being deleted, ISIS can still thrive by creating new accounts. This is in large part to the sheer volume of accounts that ISIS is using, making the network harder to decentralize than a small network with obvious accounts that would be easy to decentralize by removing the most important account.

XRSone, however, has been criticized by intelligence agencies who claim the publication of account names makes the accounts no longer a key source for gathering intelligence [49]. Nonetheless, these arguments are dismissed by these and numerous other hackers who claim the information published does not contain valuable information in terms of military operations or potential terrorist attacks. Instead, ISIS’s use of Twitter, while diverse, is focused on the recruitment of people to its cause, particularly in regions they would not have access to otherwise. As expected, ISIS is not overtly able to recruit or communicate with citizens in many Western states. Twitter serves as a way to discretely carry out conversations and hopefully gain support and thus power. By gathering support, ISIS continues to legitimize its cause securing its community and network.

4 An Analysis of ISIS’s Recruitment Techniques

4.1 Example of Methods and Techniques on “#greekweek”

Although the goals of this research are to gain understanding as to how ISIS recruits using Twitter and why it is effective, modeling the methods used on other data is
beneficial to gain how social network analysis techniques look when applied to real data. For our trial, we obtained 200 tweets that contain the hashtag “greekweek.” To understand how these tweets look, a sample of the first 5 is provided below:

1. “caraboucoffee: just so happy to party in leggings and a tshirt tomorrow
   #greekweek”

2. “RomanMarasco: RT @JustCaptnMorgan: 3:24 on gin buckets
   \xed\xa0\xbd\xed\xb2\xa9 #GreekWeek”

3. “PanhellenicAtWU: RT @golferbp: Couldn’t have asked for a better start to
   #GreekWeek @PanhellenicAtWU @WUIFC”

4. “golferbp: Couldn’t have asked for a better start to #GreekWeek
   @PanhellenicAtWU @WUIFC”

5. “soRAWRity: TSM. #greekweek #oldheads https://t.co/mWw6aHAesH”

Note the text before the colon for each tweet is the username that posted that message. Also note that text similar to “\xed” is an encoding issue because the search was set up to return Latin characters. These could be anything from an emoji because the tweet was posted by a cellphone, to characters written in a different language. Regardless, we are interested in analyzing the text and need English words and phrases, so we can ignore these for now.

Clearly these tweets need to be parsed given their mixed form. To do this, we use the tm package [50] for R to remove all non-english letters, make all letters lowercased, remove punctuation, remove numbers, remove URLs or weblinks, remove all stop words using the generic stopwords list given by the package with the inclusion of “via” and
“available,” and then stem the words. After this parsing, the first five documents previously shown in their original form were transformed to

1. “just happy party leggings tshirt tomorrow greekweek”
2. “rt justcaptnmorgan gin buckets greekweek”
3. “rt golferbp couldn’t asked better start greekweek panhellenicatwu wuife”
4. “couldn’t asked better start greekweek panhellenicatwu wuife”
5. “ttsm greekweek oldheads tcomwwahaesh”

From the remaining words in the documents, a corpus is created as well as a term document matrix. The term document matrix is the adjacency matrix for the corpus where each column denotes a unique word. The term document matrix for this particular corpus has the following characteristics:

<table>
<thead>
<tr>
<th>Terms</th>
<th>635</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>200</td>
</tr>
<tr>
<td>Non-sparse entries</td>
<td>1729</td>
</tr>
<tr>
<td>Sparse entries</td>
<td>125271</td>
</tr>
<tr>
<td>Sparsity</td>
<td>99%</td>
</tr>
<tr>
<td>Maximal term length</td>
<td>24</td>
</tr>
<tr>
<td>Weighting</td>
<td>Term frequency</td>
</tr>
</tbody>
</table>

Table 6: Characteristics of the term document matrix for tweets containing “#greekweek”

As expected, this matrix has a high level of sparsity, meaning that most terms are in a small number of the documents as opposed to being in all of the documents, and the number of characters in the longest term considered was 24. The terms with frequencies higher than 15 were: brothers, go, greek, greekweek, latechphimu, rt, songfest, tonight, week, will, winning. A complete frequency chart of terms with frequencies greater than 10 is shown in Figure 22.
A visual representation of this can be shown with a word cloud, as shown in Figure 23. For this word cloud, all terms are considered with frequencies greater than or equal to 3. The term “greekweek” was omitted from this word cloud because it has a frequency of 200 and so its proportion compared to the other terms would have made the graph difficult to interpret.
Figure 22: Frequency chart displaying terms with frequencies greater than or equal to 10 for tweets containing “#greekweek”
Figure 23: Word cloud displaying terms with frequencies greater than or equal to 3 for tweets containing “#greekweek”

Now we must graph the terms in the network. Because of how many terms there are in this network and for clarity purposes, we have removed the least frequent terms and reduced the sparsity of the original term document matrix to 89%, thereby considering only 24 terms. The new term document matrix has the following characteristics:

<table>
<thead>
<tr>
<th>Terms</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>200</td>
</tr>
<tr>
<td>Non-sparse entries</td>
<td>546</td>
</tr>
<tr>
<td>Sparse entries</td>
<td>4254</td>
</tr>
<tr>
<td>Sparsity</td>
<td>89%</td>
</tr>
<tr>
<td>Maximal term length</td>
<td>15</td>
</tr>
<tr>
<td>Weighting</td>
<td>Term frequency</td>
</tr>
</tbody>
</table>

Table 7: Characteristics of sparse frequent term document matrix for tweets containing “#greekweek”

This term document matrix was used to create the network graph. Edges are weighted by their frequency count, but for visual purposes all edge weights were divided by the
maximum frequency count, 29, and then multiplied by 7. The vertex set for this graph are the terms: all, brothers, cmnhospitals, congrats, greek, greekweek, helped, latechphimu, latechpikes, night, penny, phimualpharho, phimufraternity, phimupda, raise, saelarho, sigma, songfest, to, tonight, wars, week, will, winning. Thus there are 24 agents to consider in our analysis. There are 120 edges in this graph whereas a complete or maximally connected graph of this order would have 276 edges. The network shown in Figure 24 uses the Fruchterman-Reingold layout function of the igraph package [51] on R, and thus uses the Fruchterman-Reingold algorithm [5] when drawing the network.

![Network Diagram]

Figure 24: “#greekweek” network with 89% sparsity drawn by the Fruchterman-Reingold algorithm

Note that by the Fruchterman-Reingold algorithm, and as was clear in our frequency counts and data, “greekweek” has the highest degree and is thus located in the center of the graph. This is further clarified in the following table, which provides the normalized centrality values previously defined for each agent in the network.
The term “greekweek” has the highest degree, betweenness, closeness, and eigenvalue centrality, which is to be expected given its placement on the graph with the Fruchterman-Reingold algorithm and its degree. As for the other term’s centrality scores, they are not quite as obvious simply given the graph. For example, the term “night” has a high betweenness centrality score, suggesting it is an important term for communication between high and low degree agents. None of the terms have a particularly high closeness centrality score, although as previously noted “greekweek” does have the highest, which implies that the network is dispersed and there are numerous connections needed for any two terms to appear in the same document.

<table>
<thead>
<tr>
<th>term</th>
<th>degree</th>
<th>betweenness</th>
<th>closeness</th>
<th>eigenvector</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.47826</td>
<td>0</td>
<td>0.04925</td>
<td>0.66928</td>
</tr>
<tr>
<td>brothers</td>
<td>0.39130</td>
<td>0</td>
<td>0.05569</td>
<td>0.31143</td>
</tr>
<tr>
<td>cmnhospitals</td>
<td>0.47826</td>
<td>0</td>
<td>0.04920</td>
<td>0.66928</td>
</tr>
<tr>
<td>congrats</td>
<td>0.39130</td>
<td>0</td>
<td>0.05913</td>
<td>0.28027</td>
</tr>
<tr>
<td>greek</td>
<td>0.17391</td>
<td>0</td>
<td>0.05882</td>
<td>0.23129</td>
</tr>
<tr>
<td>greekweek</td>
<td>1</td>
<td>0.52174</td>
<td>0.07302</td>
<td>1</td>
</tr>
<tr>
<td>helped</td>
<td>0.47826</td>
<td>0</td>
<td>0.04925</td>
<td>0.66928</td>
</tr>
<tr>
<td>latecphimu</td>
<td>0.39130</td>
<td>0</td>
<td>0.05736</td>
<td>0.29199</td>
</tr>
<tr>
<td>latechpikes</td>
<td>0.39130</td>
<td>0</td>
<td>0.06267</td>
<td>0.26394</td>
</tr>
<tr>
<td>night</td>
<td>0.47826</td>
<td>0.42688</td>
<td>0.07077</td>
<td>0.14528</td>
</tr>
<tr>
<td>penny</td>
<td>0.47826</td>
<td>0</td>
<td>0.04925</td>
<td>0.66928</td>
</tr>
<tr>
<td>phimualpharho</td>
<td>0.47826</td>
<td>0</td>
<td>0.04925</td>
<td>0.66928</td>
</tr>
<tr>
<td>phimufraternity</td>
<td>0.47826</td>
<td>0</td>
<td>0.05157</td>
<td>0.62906</td>
</tr>
<tr>
<td>phimupepa</td>
<td>0.47826</td>
<td>0</td>
<td>0.04925</td>
<td>0.66928</td>
</tr>
<tr>
<td>raise</td>
<td>0.47826</td>
<td>0</td>
<td>0.04925</td>
<td>0.66928</td>
</tr>
<tr>
<td>saelarho</td>
<td>0.39130</td>
<td>0</td>
<td>0.06101</td>
<td>0.26304</td>
</tr>
<tr>
<td>sigma</td>
<td>0.17391</td>
<td>0</td>
<td>0.06479</td>
<td>0.11054</td>
</tr>
<tr>
<td>songfest</td>
<td>0.39130</td>
<td>0</td>
<td>0.05569</td>
<td>0.29997</td>
</tr>
<tr>
<td>tco</td>
<td>0.47826</td>
<td>0</td>
<td>0.05011</td>
<td>0.63537</td>
</tr>
<tr>
<td>tonight</td>
<td>0.43478</td>
<td>0.03953</td>
<td>0.06667</td>
<td>0.28815</td>
</tr>
<tr>
<td>wars</td>
<td>0.47826</td>
<td>0</td>
<td>0.04925</td>
<td>0.66928</td>
</tr>
<tr>
<td>week</td>
<td>0.17391</td>
<td>0</td>
<td>0.05882</td>
<td>0.18322</td>
</tr>
<tr>
<td>will</td>
<td>0.47826</td>
<td>0</td>
<td>0.04545</td>
<td>0.68809</td>
</tr>
<tr>
<td>winning</td>
<td>0.39130</td>
<td>0</td>
<td>0.05736</td>
<td>0.29199</td>
</tr>
</tbody>
</table>

Table 8: Characteristics of more frequent term document matrix for tweets containing “#greekweek”
Finally, the terms “all,” “cmnhospitals,” “helped,” “penny,” “phimalpharo,” “phimufraternity,” “phimupda,” “raise,” “to,” “wars,” and “will” have eigenvector centrality scores of higher than .60. This high of a score, especially in a network this dispersed, infers that these terms are commonly associated with important terms. To look at this more closely, we can identify the other terms associated with “will” at a correlation level of 0.5.

The terms correlated with being in the same document as “will” at a 0.5 level are: all, cmnhospitals, helped, penny, phimalpharo, phimupda, raise, wars, phimufraternity, tco. These terms are correlated that strongly within the entire corpus, not just the sparred version used in the network graph for visual purposes.

As for the entire network whose graph is presented below in Figure 25, there are 635 agents and 3828 edges. Note for this network to be complete it would need 7324878 edges.

Figure 25: “#greekweek” network using entire data
Consider the term “all” as was previously analyzed. In the context of the entire network, all has a degree centrality of 0.01893, a betweenness centrality of 0.00000049835, a closeness centrality of 0.27697 and an eigenvector centrality of 0.44690. Note that the degree and eigenvector centrality went down when analyzed as the entire network graph as compared to the subgraph containing more frequent terms, but its betweenness and closeness centrality actually went up. It makes sense for the degree centrality to be a smaller value than in the subgraph because in the subgraph this agent already didn’t have a large degree centrality as compared to other agents, and so unless its edges were primarily outside of the subgraph’s edge set, the score would go down. The eigenvector centrality, too, makes sense. In a larger network, there is the possibility for important agents to become more important while less important agents become less important, thus creating a dichotomy, or agents that were important in the subgraph will have less relative importance given the influx of powerful new agents. The term “all” had some relationship power in the subgraph, but “greekweek” overtook a lot of the potential control. Thus when new agents were permitted in the network, “all” has more power, as shown by the increase in its eigenvalue centrality. Its betweenness and closeness centralities decreasing implies that while this node was important previously in its ability to connect with numerous agents and its ability to connect in a small amount of steps, with more nodes added to the network it would be much harder for it to maintain these scores.

The entire network is connected as the term “greekweek” is present in each document and thus an edge is formed with all other terms in the network. However, if you apply vertex deletion the term “greekweek” thereby removing that agent and all edges connected to that agent, the network now has 17 components. The largest component has 560 agents, and the other 16 components are of size 9, 7, 18, 5, 3, 7, 1, 5, 3, 4, 3, 2, 1, 1, 1, 4. The second largest component, the one of size 18, has the following terms:
anchors, announces, aoiiphbeta, ast, astesu, bowling, brother, first, greekaward, ifc, kdr, masoltan, place, president, tcomkfgfysco, tie, vice, way. Note that none of these terms were identified as being important earlier given their centrality or frequency count, and by forming their own network it means that the important terms already identified would not lead us to looking at these terms because they aren’t in the same document as any of those terms. By removing the term “greekweek,” though, and identifying this neighborhood, we can make the decision to explore these terms more and see if they are popular with different terms pertaining to greek week and just weren’t picked up in this data set and could thus lead us to new terms that may be important for greek week or it may be that these terms do not hold importance outside of their little neighborhood and we should delete them from the analysis.

For a sentiment analysis of the tweets, the polarity function from the qdap package [52] is used. This package first uses the sentiment dictionary by Hu and Liu [53] that then identifies clusters of sentiment, so a minimum of four terms next to each other, and then tags terms within this cluster as amplifiers, neutral, or a de-amplifier. Terms originally identified as being neutral by the sentiment dictionary hold no value, but terms that were identified as positive or negative by the dictionary are then weighted based on the location of the words that were tagged as amplifiers or de-amplifiers. For example, the term “very” is not defined as being negative or positive by the sentiment dictionary, but is an amplifier. Thus the term “bright,” which is a positive word, located directly following “very” so the text reads “very bright” holds more value than just the word “bright.” Note that positive sentiment terms yield a positive number and negative terms yield a negative number, thus the more positive the document the larger the polarity score and the more negative the document the smaller the polarity score. The final polarity score is obtained by taking the sum of these scores and weights and dividing by the square root of the entire word count. For the five tweets previously mentioned, the
sentiment scores were as follows.

<table>
<thead>
<tr>
<th>parsed document text</th>
<th>polarity score</th>
<th>positive word</th>
</tr>
</thead>
<tbody>
<tr>
<td>“just happy party leggings tshirt tomorrow greekweek”</td>
<td>.378</td>
<td>happy</td>
</tr>
<tr>
<td>“rt justcaptwmorgan gin buckets greekweek”</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>“rt golferbp couldn’t asked better start greekweek panhellenicatwu wuicf”</td>
<td>0.333</td>
<td>better</td>
</tr>
<tr>
<td>“couldn’t asked better start greekweek panhellenicatwu wuicf”</td>
<td>0.378</td>
<td>better</td>
</tr>
<tr>
<td>“tttsm greekweek oldheads tcomwwahaesh”</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Polarity scores for sentiment analysis of tweets from “greekweek” data

Note that most terms in each document were not denoted as positive or negative, which could be a problem given that many words are not typical English terms and hashtags denote a separate issue as words are combined without spaces thus being unreadable by simple dictionaries. However, the polarity score given can be useful if more terms are recognized and it has the benefit of ignoring words that aren’t labeled in the dictionary as positive or negative, thus not adversely affecting the summation. However, the scores shown are as expected in that documents containing “#greekweek” are expected to be more positive than negative, which these scores display.

4.1.1 Issues and Future Work for “greekweek” Data

The primary issue with analyzing this dataset is that it is small. Terms can hold more power given their centrality when in reality they only occur in this particular subset of the masses of tweets that could be analyzed. If the dataset was larger, an analysis of usernames could be done to see which users are most frequently contributing to documents containing the initial term searched for, and who follows these users and who they follow. After looking at the initial term, those terms identified as being important given their centrality scores in the initial network should also be analyzed as well as the users who typically contribute to these terms as well as the users who follow them and the users they follow. The analysis done on this data, while not necessarily meaningful
for the entirety of “greekweek” tweets, does show that this method is successful for identifying terms that are important, and also provides a springboard for future analysis and next steps to take.

4.2 Future Work for ISIS Tweets

The “greekweek” example displays some complications that may occur. The first is determining which terms are important. Given the vastness of the network that will be analyzed that pertain to ISIS, there are numerous terms that will probably have high centrality scores. Isolating which terms matter in various components of the network will allow for a meaningful analysis in various components that may exist in the greater network. Furthermore, after identifying important terms from the original hashtag search’s network, we can look at tweets that contain those terms and see if those documents also contain the original ISIS search or if they more frequently pertain to other topics. If they primarily aren’t used within the ISIS network but instead other corpuses, these terms hold a different power than those terms that when after searching for solely documents containing them still have high correlation to ISIS tweets. This could signal these terms holding a different recruitment technique than those terms that clearly lead to uncovering more ISIS tweets.

The fundamental difference between the two analyses is the sheer amount of data. The example previously analyzed does not have much power in the trends of this type of hashtag or this topic. For ISIS, a stream like collection will need to occur so the data set will be enormous and thus representative of recruitment techniques and ISIS’s social media methods in practice. The retweet method described in [46] that ISIS uses will also be interesting to see in action, as well as how important these tweets are in terms of the terms deemed as important.
For ISIS, username analysis will not be an issue given the amount of tweets and ISIS’s method of using the same users to tweet at extremely high volumes. As previously described, ISIS uses the small number of accounts to tweet in immensely high volumes in their methodology as to attract potential supporters and ease in their recruitment. Thus ISIS will have users that hold power within their network, both in the entire ISIS network as well as components and neighborhoods. These users can be identified after the first search as well as their followers and who they follow, but this should also be compared to the users who are frequently contributing to the networks created by searching for the terms that were identified as being important. If components occur given the second search for these central terms, it is important to note which users contributed to these corpuses and if there is overlap in which user contributed to which component or if users focused on one component or even one neighborhood within a component.

Given the present focus on analyzing ISIS’s recruitment techniques, it will also be interesting to see how an important user’s followers and who they follow change. Equally important to note will be if twitter accounts are deleted as they gain followers or if there are other factors to consider. The correlation between when accounts are deleted and numerous factors hold important implications for countering terrorism and recruitment, especially given the new method of using technology to manipulate globalization into power.

Finally, the language aspect of this dataset will be important to display. As previously described in [47], the most popular Twitter accounts associated with ISIS thus far have used Arabic as their primary language (75%) and English (20%). These accounts also averaged more followers than any other account not using Arabic of English [47]. These accounts having more followers means it is probably more likely that they are more successful in their recruiting techniques, or at least are able to associate more than those that don’t use these languages. While this sort of question has already been explored by
[47] and others, what has not been analyzed is whether the networks created by these languages are connected to each other in some form of overlap or whether they constitute their own network entirely with separate recruitment techniques based on this language dichotomy as was present in congressional voting given a person’s username’s Twitter language preference [54] and was true in evaluating political sentiment [55]. Because ISIS and Twitter are constantly changing in nature, no matter the particular variable used the results will uncover more information as for how they are using Twitter, an invaluable resource in the event that other group’s take on this approach given ISIS’s success.
References


http://mathworld.wolfram.com/LaplacianMatrix.html

Connectedness of a Graph." University of Chicago, REU. 2013.


Behrend College. Chapter 10.


The Development of Social Network Analysis: A study in the Sociology of Science." 

[20] Hulst, Renée C. van der. "Introduction to Social Network Analysis (SNA) as an 

[21] "Who is central to a social network? It depends on your centrality measure." Activate 
Networks, Inc. Newton, MA. http://www.activatenetworks.net/who-is-central-to-a-social-
network-it-depends-on-your-centrality-measure/

Systems, University of Michigan.


[26] Carley, K., Reminga, J., Kamnev, N. "Destabilizing Terrorist Networks." Dynamics Networks project, CASOS. CMU.


[34] Carley, K., Kamnev, N. "A Network Optimization Approach for Improving Organizational Design." ISRI, CASOS. CMU.
    https://www.cs.cmu.edu/softagents/multi.html

    http://www.casos.cs.cmu.edu/projects/automap/

[37] Diesner, J., Carley, K. "Using Network Text Analysis to Detect the Organizational Structure of Covert Networks." CMU.


[48] Cuthbertson, Anthony. "Inside the ISIS 'social media war machine’ tracking 25,000 terrorist Twitter accounts" International Business Times March 31, 2015


