NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

INFERRING MISSING INFORMATION ON A SOCIAL NETWORK

by

Ross Spinelli

March 2020

Thesis Advisor: Ruriko Yoshida
Second Reader: Samuel E. Buttrey

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2020	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
INFERRING MISSING INFORMATION ON A SOCIAL NETWORK

5. FUNDING NUMBERS

6. AUTHOR(S) Ross Spinelli

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

Networks have long been a subject of study, but with an explosion in the amount of available data to describe them, machine learning (ML) methods have become a popular complement to traditional network analysis techniques. When the challenge of uncertainty enters the picture, many difficulties can be attacked with ML methods. Social network analysis is an analysis of networks among people using machine learning techniques. For example, social networks can be networks of friends or followers in social media, academic collaboration networks, or networks among workers. In this thesis, we propose a novel machine learning method to analyze social networks among people from a personal database via connecting similarities among people in order to get a better characterization of their relationships. This machine learning method is a non-parametric method so that we can utilize data sets containing categorical variables as well as data sets with small sample sizes or sparse networks. Under our method, we assume that nodes in a social network represent people of interest, and an edge is defined as a connection between two people in the group. To verify our methodology, we apply our method with the social network data collected in the Teenage Friends and Lifestyle Study and demonstrate its efficacy.

14. SUBJECT TERMS social network analysis, machine learning	**15. NUMBER OF PAGES** 45
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

INFERRING MISSING INFORMATION ON A SOCIAL NETWORK

Ross Spinelli
Lieutenant, United States Navy
BS, Rochester Institute of Technology, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 2020

Approved by: Ruriko Yoshida
Advisor

Samuel E. Buttrey
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Networks have long been a subject of study, but with an explosion in the amount of available data to describe them, machine learning (ML) methods have become a popular complement to traditional network analysis techniques. When the challenge of uncertainty enters the picture, many difficulties can be attacked with ML methods. Social network analysis is an analysis of networks among people using machine learning techniques. For example, social networks can be networks of friends or followers in social media, academic collaboration networks, or networks among workers. In this thesis, we propose a novel machine learning method to analyze social networks among people from a personal database via connecting similarities among people in order to get a better characterization of their relationships. This machine learning method is a non-parametric method so that we can utilize data sets containing categorical variables as well as data sets with small sample sizes or sparse networks. Under our method, we assume that nodes in a social network represent people of interest, and an edge is defined as a connection between two people in the group. To verify our methodology, we apply our method with the social network data collected in the Teenage Friends and Lifestyle Study and demonstrate its efficacy.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1 Introduction **1**
1.1 Motivation 1
1.2 Objective 2
1.3 Structure. 3

2 Background **5**
2.1 Machine Learning Methods 6
2.2 Literature Review 8

3 Models and Methodologies **11**
3.1 Predicting Connections (Edges) in Social Networks 11

4 Analysis and Results **17**
4.1 Analysis 17
4.2 Results 17

5 Conclusion **21**
5.1 Summary 21
5.2 Future work 21

List of References **23**

Initial Distribution List **25**

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

Figure 2.1 Overview of Data Science. 6

Figure 3.1 Workflow 12

Figure 3.2 Network Visualization of Girls Data 13

Figure 4.1 ROC Curve for Female Students 18

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Table 3.1 Simulated Data 13

Table 3.2 Real-World Data 14

Table 4.1 Real-World Data Confusion Matrix 18

Table 4.2 Feature Importance 19

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AdaBoost	Adaptive Boosting
AUC	Area Under the Curve
LogReg	Logistic Regression
ML	machine learning
NN	Neural Network
RF	Random Forests
ROC	Receiver Operating Characteristic
SNAP	Stanford Large Network Dataset Collection
SVM	Support Vector Machine

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Networks and graphs have long been a subject of study, but with an explosion in the amount of available data to describe them, machine learning (ML) methods have become a popular complement to traditional network analysis techniques. This is particularly true when the challenge of uncertainty enters the picture, but can be overcome with the application of ML methods with large amounts of data. Understanding a social network among workers can be used for modeling social relationship factorial analysis to improve connectedness of sailors.

In this thesis, we develop a novel non-parametric method to analyze social networks among people of interest. The input of this method is a social network and personal data. Attributes of a person help define their relationships with one another, and this thesis makes use of these attributes to identify links. To study which attributes most influence a connection, a response variable is created from connectedness in an observed social network. Using personal attributes from the observed personal data as features (explanatory variables), we predict connectedness (edges) among people (nodes) on a social network using random forests (RF), a non-parametric ML model.

We apply our method to a dataset based on friendships among primary school females in West Scotland. This dataset creates edges by linking two people if they say they are friends. From this, we examine personal attributes and try to correlate which attribute factors aid in the development of friendships among peers on the network.

With our methodology applied to this social network dataset, we computed a Receiver Operating Characteristic (ROC) curve to score accuracy rates of predicted connections among female students (Figure 1). As we can see in Figure 1, the estimated ROC curve is close to the $(0, 1)$ point. The Area Under the Curve (AUC) is a measurement between 0.5 and 1, and it indicates the accuracy of a classifier. The closer the AUC is to 0.5, the lower the accuracy of the classifier with respect to the input data set. Conversely, the closer the AUC is to 1, the better the accuracy. In our case, the AUC is approximately 0.911, a good indication that our method works well with this dataset.

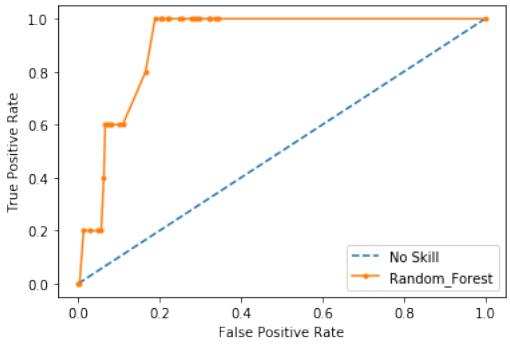


Figure 1. ROC Curve

(ROC) curve of the accuracy to predict connections between female students in a friendship social network.

We work closely with OPNAV N17, the 21st Century Sailor Office, to identify members in a command unit. As a future line of research, by identifying relationships and the connecting attributes, we may be able to apply our method to identify a potential sexual abuser. First, historical data from prior sexual assault cases may allow us to know the likely attributes linking a potential abuser to the victim. We can use this information to help rebuild network with missing data based on these known attributes.

Acknowledgments

I would like to thank the Navy for providing the opportunity to expand my knowledge and receive a master's degree. Next, I would like to thank my family and friends for supporting me in this endeavor with positive and reinforcing encouragement. Professor Ruriko Yoshida and Professor Samuel Buttrey guided me through this process better than anyone else could have done, and I thank them for their patience and mentorship. Lastly, I would like to thank my amazing girlfriend, Devin Collins, for always being patient with me and having a wonderful meal prepared after long days of thesis work. I could not have completed this thesis without her. Thank you to everyone else who I did not mention but helped me along my path to completing my thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

1.1 Motivation

A priority for OPNAV N17, the 21st Century Sailor Office, is identifying destructive behaviors among sailors in a command unit, ranging from suicide to sexual assault. We are working closely with OPNAV N17 to establish a method to identify at-risk personnel prior to any incidents. Command climate surveys are used in every command, but analyzing this data with current methods can be slow and time-consuming. We propose using predictive network analysis to identify at-risk individuals in a network based on their relationships to others.

Suicide and sexual assault continue to be two major concerns in the military. While there are several indicators to identify if a person is suicidal, there are not as many indicators or psychological tests that can be performed to identify someone as a potential sexual victimizer. We believe predictive network analysis, combined with historical data, can be used to identify a potential sexual abuser in a unit prior to any negative actions occurring. This thesis establishes the base model to identify an individual in a command unit network based on bonds or relationships among people.

To analyze social networks among people of interest using machine learning (ML) techniques and network analysis studies can be powerful. From these ML techniques, a unique pattern of relationships among people in military units can be established, and we can create algorithms to help identify links among individuals in a command unit. These links help form the basis for identifying how closely connected people are related. For example, using metadata, such as the frequency of calls people made and whom they called, we can predict relationships, such as their partners, friends, and coworkers (Grito 2016).

Links, or relationships, among people are not always reported, which leads to many problems when working with a social network. In social networks particularly, people may lie about their relationships to others or just not recall certain facts during the survey that establish links between two people. The amount of information not provided or missing from a

network is not known in advance.

Unfortunately, very little research has been completed to study the effects of incomplete data on network structures (Sparrow 1991). Most techniques for handling incomplete networks involve data imputation, the process of estimating unknown data from the observed data, which may incur unknown consequences to a network's true structure and ultimately affect classification (Little and Rubin 2014).

In this thesis, we propose our novel non-parametric method to predict connections between two people in the social network based on personal information, such as alcohol use, working hours, etc. With our ML method we might be able to predict "hidden"/"missing" connections among people in a social network.

Future work, not included in this thesis, will look at historical cases of sexual harassment and sexual abuse. We hypothesize there are certain characteristics that are prevalent in sexual abuse cases. With the use of historical data about the links between the potential abuser and the victim in sexual harassment and sexual abuse cases, we hypothesize that we can identify these connections in the unit, even if the connections were not initially reported.

Identifying a person as a potential sexual abuser poses moral implications, as an incorrect identification based on machine learning and historical cases could negatively impact a person's career or life. This thesis does not explore the moral implications the results may produce. The goal of this study is to develop a tool that can be tailored to many different areas of interest. Other potential applications of this methodology are identifying links among the poor performers in the unit and among personnel who outperform others.

1.2 Objective

This research project attempts to create a robust method for predicting connections in a social network from personal information. In this analysis, we set the connection between two people in the group of interest as the response variable and use other personal information, such as working hours, units, alcohol use, etc., as the explanatory variables. Then we apply Random Forests (RF) to predict connections among people. Future work will supplement this thesis by examining historical traits of a potential abuser.

1.3 Structure

The thesis is organized as follows. Chapter 2 reviews definitions and literary work performed in network analysis. Chapter 3 describes the multiple machine learning methods used for classification and the methodologies used for creating data. Chapter 4 presents results from classifying the percentage of information missing and describes how well we can rebuild a network. Chapter 5 summarizes the results and provides future applications to follow-on work.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
Background

This chapter gives a background of the basic terminology used for the analysis along with definitions of common terms used for social network analysis. This chapter relates the scientific definitions to the social terms used for the basis of this project.

The information in this chapter relates specifically to network analysis, and is not specific to social networks or human interactions. Throughout this analysis, the words *graph* and *network* are used interchangeably. Many studies make a subtle distinction between the two words, but in the context of this thesis, they are identical.

Graph

A graph is a combination of nodes and edges. All of our data can be described as a graph with connections based on some relationship between data points.

A graph is "an ordered pair of disjoint sets (V,E) such that E is a subset of the set $V^{(2)}$ of unordered pairs of V. . . The set V is the set of vertices and E is the set of edges. . . An edge $\{x, y\}$ is said to join the vertices x and y and is denoted by xy. Thus xy and yx mean exactly the same edge" (Bollobas 2013).

Nodes and edges of a graph can contain additional information, "such as names or strengths, to capture more details of the system" (Newman 2010). A graph consists of "set of nodes (vertices) and a set of edges (arcs) whose elements are pairs of nodes" (Ahuja et al. 1993).

Network

According to Newman (2010), "[a] network in its most elementary form is a collection of points joined together in pairs" by connections. "[It] is a simplified representation that reduces a system to an abstract structure capturing only the basics of connection patterns" (Newman 2010).

2.1 Machine Learning Methods

There are two major branches of machine learning: unsupervised learning and supervised learning (Figure 2.1). Unsupervised learning treats all of the input variables equally and tries to detect relationships among the variables, or among the observations. In supervised learning, the goal is to predict a particular response variable from some combination of the predictor variables (also called features). For more details, see James et al. (2013).

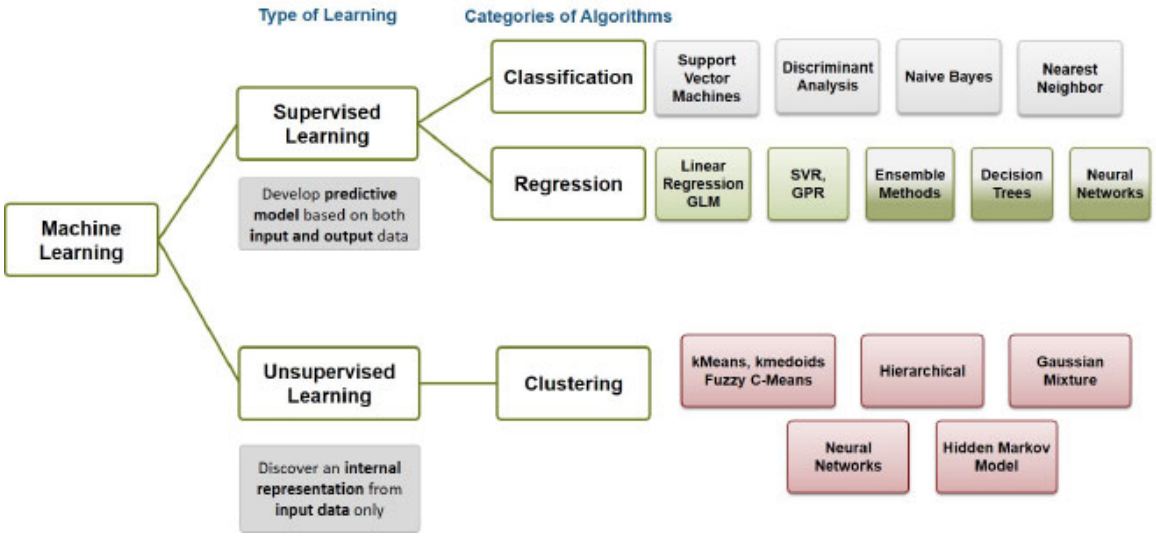


Figure 2.1. Overview of Data Science.
Source: Pilotte (2016).

2.1.1 Supervised Learning

There are two thrusts in supervised learning: classification and regression. In regression, the response variable, the one being predicted, is numeric. In classification, the response variable is categorical. Here we list some of the major classification techniques considered in this work.

- **Classification** – the response variable is categorical. Under classification, there are algorithms like logistic regression, support vector machine, linear discriminant analysis, classification trees, random forests, adaboost and etc.

- **Regression** – the response variable is numerical. There are algorithms like linear regression, regression trees, lasso, ridge regression, random forests, adaboost and etc.

2.1.2 Classifiers

In this thesis, we only look at supervised learning models in ML. More specifically, we focus on classifiers, i.e., ML models for classifications (see in Figure 2.1).

Random Forest

According to Brieman (2001), a "RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest."

Adaptive Boosting

Adaptive Boosting (AdaBoost) has been historically used for classification (Freund and Schapire 1999). Like a RF, this method uses trees as well. Initially, all classes are given the same weighting. After each round of classification, the weaker, or incorrectly identified observations are weighted more heavily for the subsequent iterations. A new tree is established and this process continues until the error function is minimized. The final step is a weighted majority vote of the trees (Freund and Schapire 1999).

Neural Network

A Neural Network (NN) works in fashion analogous to the real neurons in the human brain (*U.S. News* 2018). The neural network is used to help find patterns in data. These patterns are developed by going through multiple layers of weighting each piece of information about the end result.

Logistic Regression

Logistic Regression (LogReg) compares the odds ratio when looking at multiple predictor variables. Unlike in linear regression, we use a binomial variable to act as our response. In the model the log of the odds produces predicted probabilities, and since they are between 0 and 1, the response curve has a logistic shape (Drakos 2012).

Support Vector Machine

A Support Vector Machine (SVM) is similar to LogReg in that it attempts to find a line or plane to split the data into distinct categories. Once a split of the data is made, the algorithm attempts to create the largest distance or margin from this hyper-plane (Drakos 2012). The margin must be the largest possible for each N-dimensional data cluster. This also means a SVM can have data in any dimensional space. This method is excellent for increasing the margin of each the classifier (Drakos 2012).

2.1.3 Classification Performance Assessment

In this subsection, we discuss methods to assess the performance of a classifier.

Receiver Operating Characteristics

A Receiver Operating Characteristic (ROC) curve is used to measure the accuracy of a binary test (Mariani et al. 2017). When classifying binary data there can be 4 outcomes: True-positive, True-negative, False-positive, and False-negative. From these four features, we can calculate the accuracy, sensitivity, and specificity (Mariani et al. 2017). Specifically defined, the "ROC curve is a graphical plot of sensitivity vs $(1 - \text{specificity})$, where each point of the curve represents a different value for the cutoff to classify a statistical unit"(Mariani et al. 2017).

Area Under the Curve

When analyzing the ROC curve, another feature to look at is the Area Under the Curve (AUC). The AUC looks at the overall ROC accuracy (Mariani et al. 2017). Mariani et al. (2017) state "[AUC] ranges from 0 to 1 (perfect classification) and takes value 0.5 for a random test."

2.2 Literature Review

For our project, we developed a non-parametric method to predict connectedness (edges) in social networks from personal information. There has been much done in predicting connectedness in synthetic networks. In this subsection, we discussed how we can predict connectedness between nodes as well as predict nodes with edges adjacent to these nodes from synthetic networks.

2.2.1 Missing Information on Synthetic Networks

Prior to determining the amount of missing information in a network, we must know the type of network we are analyzing. Chia (2018) shows that we can sometimes determine the network class with very little information about the graph. She uses multiple machine learning algorithms to classify the synthetic networks. The work shows there is no single graph feature that can classify the graph; rather, multiple features are needed to accurately classify the network (Chia 2018). While this work was performed on synthetic networks, we do not expect more than half the information to be missing in a real-world network. Even if more information is missing, "only about 20% of the information about a network is needed in order to achieve better than 90% accuracy in network classification" (Chia 2018).

We begin our work by assuming the graph type is known and only the amount of information is unknown.

2.2.2 Missing Information on Real Networks

Work done by Vu (2019) takes the analysis by Chia (2018) one step further, by using real-world networks. These include technological, social, information, and biological networks (Vu 2019). Vu further shows that real-world networks with missing information can be classified accurately with a focus on using the correct statistical graph features. Once again, prior work shows only a small percentage of information is needed to classify the network. Our thesis makes reference to knowing the correct graph classification and assumes we will always know the graph type.

2.2.3 Rebuilding Networks with Missing Information

A method proposed by Rafailidis and Crestani (2016) is initially investigated. In this thesis we call their method the "Laplacian." Their research describes a method that makes use of linear algebra applications, such as calculating the Laplacian, performed on networks. When looking at a social network we have people that represent the nodes and edges that represent their relationships. We can then define a dataset that contains the attributes about a certain node, or person, in the node list. Rafailidis and Crestani propose creating two distinct adjacency matrices. The first matrix is the classical adjacency matrix between nodes and edges. Any missing data is treated as a non-existent edge between two nodes. The second matrix is constructed by looking at each pair of nodes. For each pair, "we

compute x similarities based on the x attributes. In the case of a categorical attribute, we set the similarity equal to 1, if two nodes have the same attribute and 0 otherwise" (Rafailidis and Crestani 2016). Rafailidis and Crestani go on to explain that by clustering each node based on the attribute matrix, the original network with missing information can be filled in completely. This method will be used in future work.

The first step to network completion is proposing a method removing only nodes. The Laplacian method will look at loss of information from both the nodes and edges.

All terminology from this chapter will be applied and called upon in the following chapters. Since these terms are not specific to any single type of network, they can be applied to the main idea of this thesis: determining the most likely potential abuser in a command unit from relational links among people in the unit.

CHAPTER 3:
Models and Methodologies

In this chapter, we outline our novel method to predict connections (edges) in social networks. In addition, we describe a dataset collected in the Teenage Friends and Lifestyle Study for testing our method to predict connections in the friendship network among teenage females in Scotland(SIENA 1997).

3.1 Predicting Connections (Edges) in Social Networks

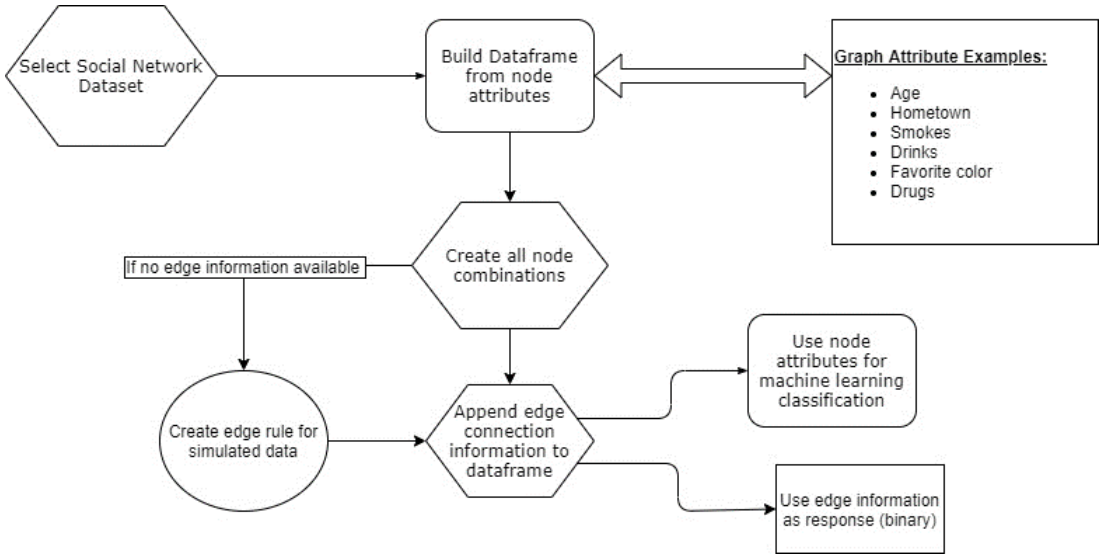
Our goal in this thesis process is predicting edges in the graph from personal information/attributes. This allows us to estimate the relationships among the nodes that were previously missing or not reported. We take two approaches to help rebuild the network based on simulated data before moving on to real-world data.

Our method here to predict edges in the network does not use the graph features (such as in Vu (2019)), but rather the node attributes. The node attributes equate to the information about a specific person. In Section 3.1.2, we see how demographics can be used as node attributes.

For both methods, the network inputs are information about a certain person that we use to link one person to another.

3.1.1 Network Completion with Algorithms

The first method involves determining a feature that connects two people in a network. First, we determine a dataset to use for the social network. This thesis uses a friendship network. Figure 3.1 shows the flow of determining an edge between two people in a network.



This diagram shows the workflow for determining whether an edge exists between two people in the network.

Figure 3.1. Workflow

We look at the known connections among people in the network and determine which attribute most strongly links the two people in the network. Once the connection is determined we can rebuild a graph based on the most important attributes.

3.1.2 Simulated Data

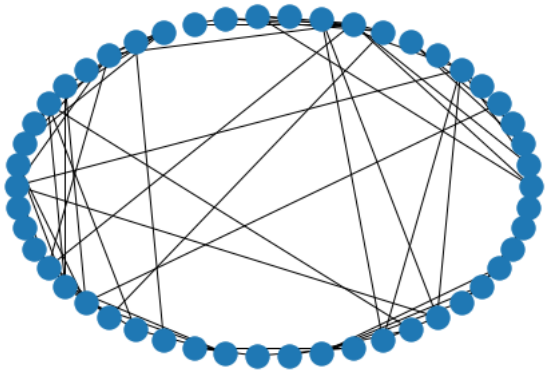
Ideally, we would like a network that has known connections (relationships) among the nodes (people). For this thesis we create a dataset of 100 nodes. We assign three attributes to each node: age, hometown, and favorite color. To establish edges between nodes, we create and algorithm to define an edge, as seen in Algorithm 1. Table 3.1 shows a sample of the simulated data we use in RF. From this dataset, we are able to start with a fully known network. This means we can delete information from the network and rebuild the network with a known end result. We use this dataset to test if an attribute can be used to determine the most likely relationship, before moving on to real world data.

Table 3.1. Simulated Data

Node	Age	Home	Color
0	23	NYC	Yellow
1	21	NYC	Yellow
2	21	NYC	Purple
3	30	PIT	Yellow
4	29	NYC	Green
5	30	MRY	Red

3.1.3 Real-World Data

The real-world data for this thesis is from COSOS Public Datasets. The data is a friendship network among girls attending primary school in Europe. The subset network used for this research is seen in Figure 3.2. Table 3.2 is a sample of responses given by a pair of girls. An edge is created for each friendship a girl reports.



The network displayed shows the connections among friends during the first year of school.

Figure 3.2. Network Visualization of Girls Data

Below is a description of the dataset as described by SIENA (1997).

> This is an excerpt of 50 girls from the Teenage Friends and Lifestyle Study data set . . . Friendship network data and substance use were recorded for a

cohort of pupils in a school in the West of Scotland. The panel data were recorded over a three-year period starting in 1995, when the pupils were aged 13, and ending in 1997. A total of 160 pupils took part in the study, 129 of whom were present at all three measurement points. The friendship networks were formed by allowing the pupils to name up to twelve best friends. Pupils were also asked about substance use and adolescent behavior associated with, for instance, lifestyle, sporting behavior and tobacco, alcohol and cannabis consumption. The question on sporting activity asked if the pupil regularly took part in any sport, or go training for sport, out of school (e.g. [soccer], gymnastics, skating, mountain biking). The school was representative of others in the region in terms of social class composition.

- **Smoking**: 1 (non), 2 (occasional) and 3 (regular, i.e. more than once per week).
- **Cannabis use**: 1 (non), 2 (tried once), 3 (occasional) and 4 (regular)
- **Alcohol**: 1 (non), 2 (once or twice a year), 3 (once a month), 4 (once a week) and 5 (more than once a week)
- **Sport**: 1 (not regular) and 2 (regular)

Table 3.2. Real-World Data

(i,j)	Node_i Alcohol	Node_i Smoke	Node_i Drugs	Node_i Sports	Node_j Alcohol	Node_j Smoke	Node_j Drugs	Node_j Sports	Edge
(5, 13)	3	1	1	2	3	1	3	1	0
(34, 35)	2	1	1	2	3	1	2	2	0
(11, 15)	5	3	2	2	4	2	1	2	1
(13, 37)	3	1	3	1	2	1	1	2	0
(12, 18)	5	3	3	2	4	1	1	1	0

Simulated versus Real-World Data

For the simulated data, a rule is needed to determine which nodes should be connected. There currently is no affinity for an edge connection between two nodes in the simulated data. Randomly assigning edges to each node pair will not produce realistic results. For the simulated data, we create a rule to establish an edge between node pairs as follows: *Create an edge between a node pair if their age is within 2 years of each other and if they like*

the same color or live in the same city. Algorithm 1 shows the pseudo-code used to create edges for the simulated data.

Algorithm 1 Creating an Edge

```
1: procedure EDGE CREATION RULE
2:     for all (i,j) node pairs do
3:         if |Node i Age - Node j Age| ≤ 2 then
4:             if Node i Color = Node j Color or Node i City = Node j City then
5:                 pair (i,j) has an edge
6:             else
7:                 pair (i,j) does not have an edge
```

3.1.4 Network Completion with the Laplacian

The second method, we begin to use is based on the paper by Rafailidis and Crestani (2016). In this thesis, we call their method the "Laplacian." Their research describes a method that makes use of linear algebra applications, such as calculating the Laplacian, performed on networks. In a social network the nodes are people and the edges are relationships. We can then define a dataset that contains all the attributes about a certain node, or person, in the node list. Rafailidis and Crestani propose creating two distinct adjacency matrices. The first matrix is the classical adjacency matrix between nodes and edges. Any missing data is treated as a non-existent edge between two nodes. The second matrix is constructed by looking at each pair of nodes. For each pair, "we compute x similarities based on the x attributes. In the case of a categorical attribute, we set the similarity equal to 1, if two nodes have the same attribute and 0 otherwise" (Rafailidis and Crestani 2016). Rafailidis and Crestani go on to explain that by clustering each node based on the attribute matrix, the original network with missing information can be filled in completely. This method will be used in future work.

The first step to network completion is verifying a method of only node removal. The Laplacian method will look at loss of information from both the nodes and edges.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Analysis and Results

4.1 Analysis

To identify links between nodes in a network we apply our method to both simulated data and real-world data. Our method is implemented in Python.

The simulated data contains six categorical features (seen in Table 3.1), while the real-world data contains eight ordinal features (seen in Table 3.2). Both of these datasets use a binomial response to predict if an edge exists or does not exist. We use a RF algorithm to predict the missing edges and the most important features.

Chapter 3 discusses the datasets in more detail. For all ML predictions, a 25% random sample of the data is used as the test set.

4.2 Results

4.2.1 Rebuilding the Graph

The real-world data collection spanned three years. Each year is analyzed separately and also a fourth analysis is done with all three years of information combined. The analysis shows we can accurately classify whether two people are friends with 81% accuracy. Given this small amount of data, this accuracy is outstanding, but we investigate the analysis further. In Table 4.1, we only have a small number of correctly predicted friendships existing. The total number of friendships in this network is 55. This provides a limited amount of data to analyze. Although the confusion matrix provides important information for us, we take one more step to confirm our results. We look at the ROC curve to more closely analyze our false-positive and false-negative rates.

Table 4.1. Real-World Data Confusion Matrix

		True	
		No Edge	**Edge**
Predicted	**No Edge**	249	0
	Edge	59	5

We see no false-positive results, but a high number of false-negatives.

The ROC curve, in Figure 4.1, has an area under the curve of 0.911. This shows that our model performs well, even with a small amount of data. The model should perform even better with a larger network, but can work with command units as small as 50 service members.

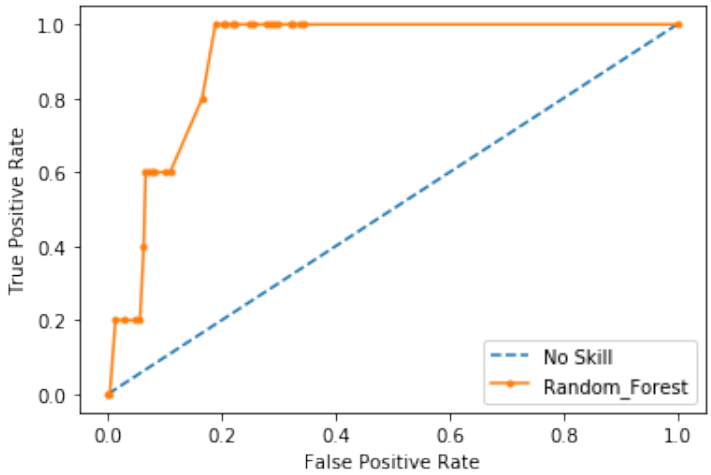


Figure 4.1. ROC Curve for Female Students

The results, from each of the four datasets, show alcohol as the most important classifier, or attribute of friendship. Table 4.2 orders the features used by the random forest by their importance, as computed in the algorithm. Whether a person uses alcohol and how often they use alcohol determines how likely two people are to be friends in this network. This is an extremely important result. We can now estimate the probability of a new person becoming friends with each of the existing members in the group.

Table 4.2. Feature Importance

Feature	**Importance**
Node_j Alcohol	0.2391
Node_i Alcohol	0.2369
Node_i Drugs	0.1407
Node_j Drugs	0.1112
Node_i Sports	0.0932
Node_i Smoke	0.0708
Node_j Sports	0.0633
Node_j Smoke	0.0444

The results are from the first year of data collection. We see that alcohol is the strongest indicator of friendship in this network.

The analysis shows we can accurately identify individuals in a network based on the most important classification features. If a new member were to joint the network, we should be able to predict their friends with high accuracy. If a new member were to join the network, we should be able to predict their friends with high accuracy.

This same idea can be applied to N17 data. Instead of looking at alcohol consumption, there may be another trait about a person that links them strongly to another individual. If we look at the links that have historically been seen in sexual assault cases, we may be able to identify a potential abuser before any incident occurs.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
Conclusion

5.1 Summary

Determining the links among people in a military unit is the first step to the complete network. Building the complete network allows us to see relationships among the people in the unit, including those not initially reported. By knowing the relationships, or connections, between people we can determine which are the most important. Not only do we determine the most statistically important features (attributes), but we also determine the nature of relationships among people. In future cases, our methods can be applied to the Alcohol Drug Management Information Tracking System and DEOMI Organizational Climate Surveys from OPNAV N17. Historical cases of sexual assault could then be used to determine which connections were present between the potential abuser and the accuser. Once these historical links are examined, we can look at our completed network and find the people (nodes) that have these connections and then try to apply these models to current and future commands.

This thesis used data from CASOS Public Datasets as a proof of concept. The proof of concept was to establish a network among a group of people based on information about each person in the dataset. Once relationships were established, information was deleted to simulate missing information, relating to the incomplete information provided by military units. We assume there will always be some amount of missing information. Machine learning methods were applied to the data to rebuild the network to completion and the data revealed a consistent result over multiple years. For each year of data, the categorical value of "Alcohol Use" proved to be the strongest link among friends. This attribute helped rebuild links between other members of the network with greater than 81% accuracy.

5.2 Future work

Although much work has been done on identifying incomplete networks and rebuilding those networks, much more work and analysis is needed to produce a robust model.

The next step to achieving this goal is to implement a network completion method introduced by Rafailidis and Crestani (2016). This method will allow for missing information to come from both the nodes (people not reported as being in a unit) and from edges (relationships not reported between people in a unit).

In addition to new methods, future work in this area can be completed on a wider sample of real-world datasets. Online network repositories, such as the Stanford Large Network Dataset Collection (SNAP), provide hundreds of datasets for public use. Performing this analysis on a wider array of social networks will help refine the process before these methods are applied to real-world data without a known amount of information missing.

The end state of this project is to develop a tool for OPNAV N17 to use to determine the likely people, if any, that could commit sexual assault or sexual harassment.

List of References

Ahuja RK, Magnanti TL, Orlin JB (1993) *Network Flows: Theory, Algorithms, and Applications* (USA: Prentice-Hall, Inc.).

Bollobas B (2013) *Modern Graph Theory*. Graduate Texts in Mathematics (Springer New York).

Brieman L (2001) Random forests. *Machine Learning* 45(1):5–32, https://doi.org/10.1023/A:1010933404324.

Chia XLP (2018) *Assessing the robustness of graph statistics for network analysis under incomplete information*. Master's thesis, Naval Postgraduate School, https://calhoun.nps.edu/handle/10945/58284.

Drakos G (2012) Support vector machine vs logistic regression. *Medium* Accessed 27 Jan, 2020, https://medium.com/@george.drakos62/support-vector-machine-vs-logistic-regression-94cc2975433f.

Freund Y, Schapire RE (1999) A short introduction to boosting. *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1401–1406 (Morgan Kaufmann).

Grito E (2016) Identifying relationships. Accessed March 18 2020, https://elgrito.witness.org/portfolio/metadata-relationships/.

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning : with Applications in R* (New York, New York: Springer).

Little R, Rubin D (2014) *Statistical Analysis with Missing Data.* https://ebookcentral.proquest.com/lib/ebook-nps/detail.action?docID=1775204.

Mariani P, Marletta A, Sciandra M (2017) Gamlss for big data: Roc curve prediction using twitter data.

Newman M (2010) *Networks: An Introduction* (USA: Oxford University Press, Inc.).

Pilotte P (2016) Analytics-driven embedded systems, part 2 - developing analytics and prescriptive controls. Accessed March 18 2020, http://embedded-computing.com/articles/analytics-driven-embedded-systems-part-2-developing-analytics-and-prescriptive-controls/.

Rafailidis D, Crestani F (2016) Network completion via joint node clustering and similarity learning. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* 63–68.

SIENA (1997) Teenage friends and lifestyle study. Accessed January 4, 2020, https://www.stats.ox.ac.uk/ snijders/siena/Glasgow_data.htm.

Sparrow MK (1991) The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks* 13(3):251 – 274, ISSN 0378-8733, URL http://dx.doi.org/https://doi.org/10.1016/0378-8733(91)90008-H.

US News (2018) Sexual assault reports spike in metoo era (December 27), https://www.usnews.com/news/national-news/articles/2018-12-27/sexual-assault-reports-spike-in-metoo-era.

Vu C (2019) *A Method for Classification of Incomplete Networks: Training the Model with Complete and Incomplete Information*. Master's thesis, https://calhoun.nps.edu/handle/10945/62313.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California