PRELIMINARY STUDY OF COMMUNICATION NETWORK CHARACTERIZATION TOWARDS IMPROVED ORGANIZATIONAL BEHAVIOR

THESIS

Taylor Flaxington, DAF

AFIT-ENV-MS-20-M-202

DEPARTMENT OF THE AIR FORCE

AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENV-MS-20-M-195

PRELIMINARY STUDY OF COMMUNICAIOTN NETWORK CHARACTERIZATION TOWARDS IMPROVED ORGANIZATIONAL BEHAVIOR

THESIS

Presented to the Faculty

Department of Systems Engineering and Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Engineering Management

Taylor Flaxington

March 2020

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENV-MS-20-M-195

PRELIMINARY STUDY OF COMMUNICATION NETWORK CHARACTERIZATION TOWARDS IMPROVED ORGANIZATIONAL BEHAVIOR

Taylor Flaxington

Committee Membership:

Lt Col Amy. M. Cox, PhD
Chair

Dr. Alice Grimes, PhD

Dr. Christine Schubert Kabban, PhD

AFIT-ENV-MS-20-M-202

Abstract

Nearly one third of the Air Force acquisition system's large programs are over cost and half are over budget; performance must improve. This research applies a systems perspective to this challenge and frames the acquisition system as a complex system of systems. It is a system composed of people in multiple organizations; organizations interacting with one another to develop, acquire and sustain weapon programs. The system's performance is an emergent behavior of its components (people), structure and processes.

Communication networks are a view of a system's structure, revealing the flows and interactions between components (people) as the system accomplishes its functions. Literature demonstrates that these networks are key to the effective performance of various system functions (ex. innovation). Further, established methods in organization behavior literature allow for characterization of these networks. Yet, limits to existing methods reduce their utility. This research validates a method to characterize communication networks within large technical organizations.

This research compares communication network mapping with two separate data sources: interviews (existing) and archival e-mail log files (new). Ego-centric networks for five volunteers are characterized with both data sources and compared via case study methods. This research makes three contributions. First, it demonstrates the effect of archival data inclusion on the observed network completeness. Second, it compares the content of the networks observed with both data sources. Third, it compares established network measures obtained with both data sources. Future research can leverage this method validation to explore the use of e-mail-based network characterization to improve organizational performance.

For those I love.

Acknowledgments

I would like to express my sincere appreciation to my thesis advisor, Lt Col Amy Cox, for her unparalleled insight, enthusiasm, and experience throughout the course of this thesis effort. I would also like to thank my committee members, Dr. Alice Grimes and Dr. Christine Schubert Kabban, for their participation and help over the last eighteen months.

I would like to thank Dr. David Long for his mentorship and support, and being the "sanity check" to every step of this research. As well as the rest of our weekly 20M thesis group, Capt Evan Gist and Lt Ethan Blake, who had to sit through presentation after presentation on social networks.

Table of Contents

Abstract.....IV

I. Introduction.....1

1.1 Background.....1

1.2 Problem Statement.....2

1.3 Research Questions.....3

1.4 Methodology.....4

1.5 Assumptions/Limitations.....5

II. Literature Review.....7

2.1 Organizational Behavior.....8

2.1.1 United States Air Force.....8

2.1.2 Organizations as a System.....9

2.2 Communication Networks.....10

2.2.1 Innovation.....11

2.2.2 Network Stabilization.....12

2.2.3 Newcomer Socialization.....12

2.3 Social Network Analysis.....14

2.3.1 History.....16

2.3.2 Analysis.....18

2.4 Data Collection.....21

2.4.1 Methods.....21

2.4.2 Method Selection.....25

III. Methodology.....29

3.1 Research Setting.....29

3.2 Case Study Selection.....30

3.3 Data Collection.....30

3.4 Data Analysis.....33

3.4.1 Phase 1: Characterization of Archival Data.....33

3.4.2 Phase 2: Extent of Similar Data.....35

3.4.3 Phase 3: Comparison of Methods.....35

IV. Results and Analysis.....38

4.1 Archival Data Selection 38
4.1.1 Completeness 38
4.1.2 Glocalization 44
4.1.3 Summary 45
4.2 Extent of Similarity of Data 45
4.2.1 Test: Do the names from the interview appear in the email data? 46
4.2.2 Test: Do the connections from the interview appear in the email data? 46
4.2.3 Summary 49
4.3 Comparison of Methods 49
4.3.1 Presence, Placement, and Position 49
4.3.2 Test: In its totality, how does the email data compare to the interview data? 51
V. Conclusion 54
5.1 Conclusions of Research 54
5.2 Significance of Research 57
5.3 Recommendations for Action 58
5.4 Recommendations for Future Research 59
VI. References 60
VII. Appendix A 63
VIII. Appendix B 65

List of Figures

Figure II.1 Structure of Literature Review 7
Figure II.2 USAF Demographics [18] 9
Figure II.3 Network Diagram Nodes 15
Figure II.4 Network Diagram Edges 16
Figure II.5 Low Density 20
Figure II.6 High Density 20
Figure II.7 Network Modularity 21
Figure II.8 Size of an organization versus effort required of data collection [10] 25
Figure III.1 Research Design 29

List of Tables

Table II.1 Phenomena Constructs for Data Collection Method Comparison 26
Table II.2 Setting Constructs for Data Collection Method Comparison 26
Table II.3 Validity Constructs for Data Collection Method Comparison 27
Table II.4 Data Collection Method Comparison 27
Table III.1 Example Data Output 31
Table III.2 Data Analysis 33
Table III.3 Breakdown of Archival Data 34
Table III.4 Phase 1 Construct Definition 35
Table III.5 Phase 3 Construct Definitions 36
Table III.6 Phase 3 Construct Definition (continued) 37
Table IV.1 Sent Data Cross Case Analysis 39
Table IV.2 Inbox Data Cross Case Comparison 41
Table IV.3 Archival Combination Data Cross Case Comparison 42
Table IV.4 Percentage of Nodes in Each Data Source 43
Table IV.5 Percentage of Edges in Each Data Source 43
Table IV.6 Cases with Greater than 250 Nodes 44
Table IV.7 Node Matching 46
Table IV.8 Edges subgraph comparison 48
Table IV.9 Presence of Interview Data in Top Ten of Email Data 50
Table IV.10 Placement of Interview Data in Top Ten of Email Data 51
Table IV.11 Position of Interview Data in Email Data 51
Table IV.12 Phase 3 Construct Comparison 52
Table V.1 Archival Data Organization 55

Preliminary Study of Communication Network Characterization Towards Improved Organizational Behavior

I. Introduction

1.1 Background

The former Secretary of the Air Force Heather Wilson established "advancing at the speed of relevance" as an objective for Air Force technology development [1]. This need for speed is further highlighted in the annual review of the Air Force's top 50 programs; the Air Force's acquisition system is not meeting its established performance objectives. Of the weapon system programs with established baselines, nearly one-third are over cost and one-half are over schedule [1]. The Air Force's organizations responsible for technology development are not performing at the desired rate and are not meeting established delivery timelines.

The challenge of improving the performance of the Air Force acquisition system serves to motivate this current research. This challenge of effective innovation and technology development is not new for military organizations [2]. While there are unique setting characteristics that can influence innovation (ex. market structure), there are more general challenges associated with the performance of large technical organizations. This general area of research is active.

Concepts from Systems Architecture are leveraged in this current research to frame the problem of organization performance. Technical organizations can be modeled as social systems in which people are the parts and their communication is the interactions between the parts [3], [4]. Systems, whether of technologies or people, can be modeled as networks of nodes and the connections between those nodes. In Systems Architecture, the capabilities and performance of

collaborative systems are emergent behaviors that result, in part, from the components (nodes) and their interactions (connections) [3]–[5]. This concept from the field of Systems Architecture, the influence of structure on performance, underlies this research. This research provides a means to observe this collaborative social system as a network of components and their interactions.

One method to characterize communication networks is social network analysis. Social network analysis (SNA) is a multidisciplinary approach to understanding the social context and behavior of relationships between people [6]. In a manner similar to how doctors use an X-Ray to see the underlying structure of their patients, SNA is a means to visualize the underlying communication structure among groups of people [7].

1.2 Problem Statement

The United States Air Force (USAF) is a large, dynamic, technical organization facing persistent disruptions. Nearly one third of military members move each year and the number of experienced professionals in acquisitions is declining; nearly 50% of the acquisition workforce is within 5 years of retirement [8]. Each year, federal acquisitions gets more and more complex and is facing huge shortfalls when it comes to mid-career professionals [8].

Modeling technical organizations after complex systems provides one way of studying the behavior of such a dynamic environment. However, how and what communication data is collected has an effect on the outcome of the characterization of communication networks [9]. Much of the criticisms of SNA is that it is often based on weak data, with strong analysis [9].

Of the vast number of data collection methods used in SNA, sociometric questioning is the most widely used [10]. Sociometric Questioning is the process of asking every member in a defined set of people who they interact with most [6], [7], [10]. This can take a lot of time and effort to produce one network diagram [10]. It involves surveys and participation from everyone in a program or on a team [10].

Archival data is an alternative to sociometric questioning that relies on the collection of stored communication data [10] [11]. Sources of archival data may be internal mail, phone records and email. The focus of this research is on archival data found in email. Recent examples of conducting a SNA from email data can be seen in the MIT Immersion Tool [12] and an analysis of the Clinton email network [13]. However, few strictly methodological contributions outline the application of archival data for an organizational improvement purpose [9], [10]. There is a need to compare the traditional data collection method, sociometric questioning, with archival data from email traffic.

1.3 Research Questions

This research aims to answer the following questions towards characterization of communication networks in technical organizations

- What is the effect of archival data inclusion on the completeness of the observed network?
- How similar is the structure obtained with sociometric questioning to the representative subset of archival data (e.g. subgraph isomorphism)?
- What effect does the data collection method have on the observation of predefined network characteristics?

1.4 Methodology

This research focuses on a case study based analysis of a large technical organization [14]. This population was simulated using faculty and staff at the Air Force Institute of Technology (AFIT). From this pool, five volunteers were asked to complete an online questionnaire and data from their email was collected. Each research question then drove a phase of analysis for this research.

Phase one characterized the email data using a test for completeness. Completeness was defined as representing the greatest number of nodes and edges. This phase also looked to identify glocalization, a phenomena previously identified in the study of email communication [15]. This phase identified the sources of archival data that provided the most complete network representation. This phase informed the data selection for the archival-based characterization that was the point of comparison between archival and interview methods.

Phase two identified the presence of the interview data in the email data using graph matching; how similar is the network structure obtained with the two data sets? Based on earlier experimentation with the techniques it was known a priori that archived data provided a larger set of nodes and edges (e.g. multiple orders of magnitude). A subgraph of archival data, inclusive of the nodes represented in the interview data, was compared to the graph from the interview data to determine the degree of isomorphism or similarity. This comparison informed on whether, or to what extent, both methods observed the same structure.

Phase three of the research compared the totality of the data collected from each method. The first test was to identify the presence, placement and position of the interview data in the whole of the email data. It is not enough to say that the interview data is present but also where

that data is in terms of importance in the email data. The two methods were then compared through network measures that included the number of nodes and edges, modularity and density.

1.5 Assumptions/Limitations

There are two major assumptions of this research. First, the archival is representative of work interactions and communications (internal validity). The archival data considered in this research is work email. While there is the possibility of social interaction via work e-mail, this assumption of the predominance of professional content is validated by the literature. Work e-mail is a good indicator of work-related communication. Email is the least likely form of communication to be used socially relative to other forms (e.g. discussions by the coffee pot) [15], [16].

The second assumption is tied to the generalization of the data in this research to our setting of interest (external validity). The goal of this research is to study a large technical organization, similar to the Air Force acquisition system (or components thereof). This population was not accessible for the duration of this research, so a simulated population was used. This population engaged in collaboration across multiple organizations with sustained interaction over time and provided both available data and data representative of a complex technical organization. Faculty and staff engage in collaboration within AFIT as well as across various collaborating research organizations and sponsors.

This research pursued analytical versus statistical generalization regarding individual networks; each of the five cases provides for literal replication of observed patterns in the ego-centric networks of individuals in a technical organization [Yin, 2014]. Two levels of analysis were considered, full networks traits (e.g. the density or modularity of the entire network) and node/edge level presence. Due to the numbers of nodes and edges considered, this research does

perform some statistical analyses. Returning to analytic generalization, further research might target populations of informants that are expected to behave in dissimilar ways (ex. comparing student to faculty networks) to allow for a more robust theoretical replication of network traits. The case selection only considered a common population with common expected traits. This study did not select dissimilar cases towards the end of theoretical replication. Future research should pursue more diverse informants, relevant to the environment of a complex technical organization, to provide for theoretical replication.

II. Literature Review

The foundation of this research lies on concepts. First, a definition of organizational behavior. In this context the organization to be studied is the United States Air Force (USAF). This organization is vast in both its size and scope. The USAF, with any technical organization, can be modeled as a social system, with people as the components. When defining an organization as a social system, performance, whether good or bad, is an emergent behavior of that organization. Second, the importance of the communication network between people in an organization as a means of influencing the emergent behaviors of organizations. The third, a method of visualizing and characterizing that communication network informs understanding. And finally, the data that is necessary to allow for such visualization. These concepts and how they relate are depicted in Figure II.1.

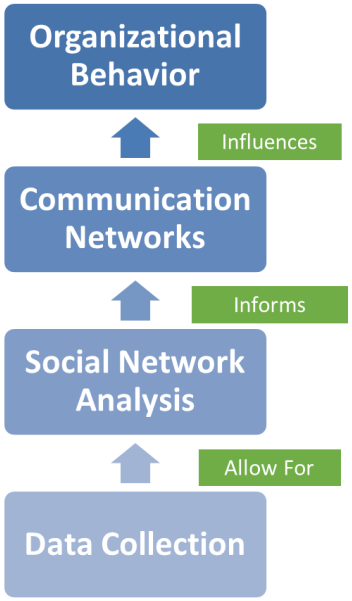


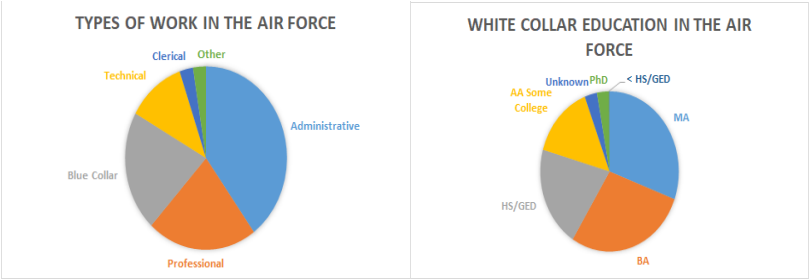
Figure II.1 Structure of Literature Review

2.1 Organizational Behavior

2.1.1 United States Air Force

When looking at the possible tools for analyzing the underlying communication network of an organization, it is critical to first characterize the environment to which the tool will be employed. The United States Air Force (USAF) is a very large, multidisciplinary, highly skilled workforce that tackles a wide variety of problems every day. Like other large public organizations the USAF is able to concurrently "develop and transform small, relatively autonomous organizations conducting ad hoc research efforts to a large, focused, geographically distributed centrally managed system" [17].

According to the 2017 Demographics Report, the DoD employees roughly 3.5 million people [18]. Within the USAF there are 325,100 Active Duty and 176,533 civilian personnel [18]. Figure 2.2 shows the breakdown of types of work, and education for civilians and officers within the USAF. Of the civilian personnel who work for the USAF, the majority work in administrative and professional jobs with a bachelor's degree or higher. Officers in the USAF are required to have a bachelor's degree or higher [18] [1].



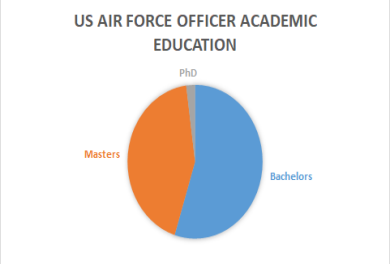


Figure II.2 USAF Demographics [18]

Of that workforce, each year one third of the military moves to different geographic locations all over the world as part of the Permanent Change of Station process (PCS) [1], [8], [18] [19]. Military members move every three years in order to get a breadth of knowledge in different fields within the USAF. On top of these moves, military members and their home organizations also have to deal with frequent disruptions from deployments. As a function of deployments, military members are tasked to serve in alternative locations for 180 or 365 day durations.

In addition to these persistent disruptions, nearly 50% of the civilian acquisition workforce is eligible to retire in the next five years [8]. This will lead to the number of acquisition professionals declining and with them the experience and knowledge that is necessary to tackle complex acquisition problems. A study done by MITRE identified that this massive outflow of experienced professionals will lead to huge shortfalls regarding mid-career professionals [8]. This working environment is dynamic and experiencing persistent disruptions.

2.1.2 Organizations as a System

If Aristotle is to be believed that “many things have a kind of plurality of parts and are not merely a complete aggregate but instead some kind of whole beyond its parts” [20], then a technical organization may be more than just the sum of its parts. The Systems Engineering

Body of Knowledge (SEBOK) describes this as holism; that ideas, people or things must be considered in relation to the things around them to be fully understood [4]. In order to gain insight to the USAF, the argument will be made that a technical organization is a system.

General Systems Theory (GST) defines a system "as a set of elements in interaction" [3]. The International Council on Systems Engineering (INCOSE) expands this definition of a system as "an arrangement of parts or elements that together exhibit behavior or meaning that the individual constituents do not" [3].

Systems can be either open or closed systems [3], [4]. Open systems exchange inputs and outputs with the environment [3], [4]. Bertalanffy categorized open systems into natural systems, social systems and technological systems [3] . Natural systems are outside of human control, and technical systems are man-made [3], [4]. Social systems are made up of human elements and often are representative of social constructs or groups [4].

The semi-autonomous organizations within the USAF are each their own social system. These social systems, based on the definition of a system, are an arrangement of people (parts). Social systems are more than just an aggregation of parts, but rather experience an emergent behavior that is unique to the relationships and interactions between those parts [3], [4]. This behavior is emergent from the system and is meaningful from the view of the system as a whole and not at the component level [3], [4].

2.2 Communication Networks

Communication networks within an organization represent the interactions between components, or people in the network. The following examples are independent areas in an organization that illustrate the emergent behaviors of the communication networks of

organization. Addressing areas that are impacted by communication set the groundwork for why this research is important.

2.2.1 Innovation

Where do ideas come from? In most cases people seek out the easiest sources of information [15], [16]. Information is costly [21] and in situations where the cost of information is high, people will first look for local information. This hyper-local information is first found within themselves; a lifetime of "idiosyncratic knowledge" builds up within every person given their unique experiences [21]. This unique, experience-based knowledge allows people to recognize potential that others would not have [21].

When that hyper-local information is no longer enough, people will move on to the next least costly form of information gathering - the people around them [21], [22]. Thomas Jefferson believed ideas spread in the same way that light spreads from a candle, "He who receives an idea from me, receives instruction himself without lessening mine; as he who lights his taper at mine, receives light without darkening mine [23]." It is under this premise that social networks play a huge role in knowledge transfer and innovation [21], [22]. These close ties, or ties that are not geographically separated allow for knowledge transfer to happen at the least cost to those in the transaction.

Within the social network of an individual, it is not just the close ties that play an important role in knowledge transfer but also that of the weak ties [9], [15], [16], [24], [25]. Weak ties are the connective tissue of the social network. The importance of weak ties is especially clear in the work done by Stanley Milgram and the six degrees of separation. This study identified that any two people, through a network of "friends of friends" or weak ties, are only six people separating everyone in the United States [26]. Ties can be defined as the

combination of time, emotional intensity, and reciprocity of services [9], [15], [16], [24], [25]. Strong ties are the people you interact with most often. Strong ties typically form between nodes that are within the same group and have access to similar pools of information. Weak ties represent the bridges between groups, thus linking pools of information together [9], [15], [16], [24], [25].

2.2.2 Network Stabilization

Similar to how destabilized nation states are ineffective in providing basic goods, so are destabilized organizations in providing basic services [11], [27]. These basic services in the context of a technical organization are the products and systems that exist to produce. Destabilization stems from an organization being fragile, weak or collapsed due to missed opportunities in the underlying informal network [11], [27]. An analysis of the structural relationships in that organization can help inform and influence the growth of organizations by identifying these missed opportunities. These missed opportunities are identified as structural holes in the network [11], [27].

Structural hole theory is the concept concerning the capacity of a network for mutually beneficial collective action [11], [27]. When structural holes are filled knowledge can be transferred among groups in turn, stabilizing the network and allowing for an increase in productivity [11], [27].

2.2.3 Newcomer Socialization

On-boarding is the action or process of integrating a new employee into an organization [28]. During this time of onboarding, an organization can lose anywhere from 1-2.5% of its total revenues to a lack of productivity [28]–[30]. Employers have accepted months of lag time

between when a new employee joins an organization to when they actually become productive [28].

This loss of productivity is due to the new employee spending time learning the knowledge, skills and most importantly behaviors associated with their new organization [28]–[30]. This learning period is defined as organizational socialization [28] [31]. The outcome of this socialization is ranging levels of organizational satisfaction, commitment, and performance [28]–[30].

To successfully meet the goal of newcomer socialization a new employee must have a number or behaviors that encourage adjustment to the organization. Bauer wrote, “Relationship building was found to be one of the important antecedents to socialization outcomes such as performance and satisfaction” [28].The ability of a newcomer to build relationships accelerated the process of on-boarding and led to faster results of performance [28], [29]. Newcomers need to build relationships to learn the organizational roles and gain acceptance [28]–[30].

People who have been in the organization know the “web of ropes”, or relationships gained through years of experience [28]–[30]. The critical information on how the organization works on a daily basis is found with the people who have been there [28]. Unfortunately, for a newcomer everyone in the organization at first glance seems the same; there is no differentiation between who to go to for what kind of help [28]–[30].

The key to new employees reaching those organizational outcomes of performance, commitment and satisfaction depend on the newcomers ability to identify key roles in the social structure of the organization [28]–[30]. Newcomers need to know who has the information. It is a fiction that newcomers are able to do this effectively by themselves [28]–[30]. In a survey done

by Rollag, nearly one third of newcomers attributed their inability to quickly get up to speed on an insufficient introduction to members of their organization [31]. In any organization it is who you know that counts [31], [32] .

According to a study on the DoD conducted by the Rand Corporation, a nonprofit, nonpartisan research organization, nearly one third of service members move each year [19]. Each year, one third of the military is adjusting to new jobs and faced with the daunting task of becoming a productive member of a program office as quickly as possible [19]. In the challenge of getting newcomers up to speed and becoming productive, the key is knowing the informal communication structure of the new organization [28], [29], [31], [32]. This gives newcomers the knowledge of what roles people play in the organization and who to go to for help [32].

2.3 Social Network Analysis

Social network analysis (SNA) is a multidisciplinary approach to understanding the social context and behavior of relationships between people [6], [16], [33]–[35]. SNA is one method of studying the communication network that underlies an organization. This communication network is responsible, in part, for the emergent behavior of a social system [3], [4].

SNA can focus at the microscopic, mesoscopic, or macroscopic levels. At the micro level the focus is on the relationships of one node, an actor, to small groups [35]. A dyad is a group that consists of two nodes and a triad is three nodes [6], [34]. At the macro level, the focus is on the outcome of interpersonal interactions [6], [25], [35] . This thesis is a study of organizational structure at the macro level. Meso is the connection between the micro and macro.

At any level SNA can be broken down into two groups: the study of a whole network, sociocentric, or of an ego-centric network[6], [25], [34]. A whole network study is used for small

populations, when the complete roster of people participating in that organization is defined [6], [10], [34], [35]. In other cases when the group cannot be so easily defined, the focus is on ego-centric networks [10], [16], [25], [33], [34]. Ego-centric networks have one person as the focus of the analysis and the connections that they are able to recall and therefore face the limitations of human bias and memory [10], [16], [25], [33], [34]. Ego-centric networks further differ from sociocentric networks in that they are focused on individual level outcomes like health, or voting behavior as opposed to resource distribution, or information diffusion studied in a sociocentric network [10], [16], [25], [33], [34]. The biggest benefit to egocentric networks are that they can be easily scaled to a large organization, and use individuals as cases [10], [16], [25], [33], [34].

Nodes and edges, depicted in Figure II.3, make up a network diagram. Nodes are representations of people in the graph [10], [16], [25], [33], [34]. In ego-centric networks the focal person is called the ego and the people to which the ego has connections with are called alters[10], [16], [25], [33], [34]. The nodes in a graph are connected by edges [10], [16], [25], [33], [34]. These tie or link two or more nodes in a graph. An example of a graph with people represented as nodes and edges representing their connections can be seen in Figure II.4**Error! Reference source not found.**. In social applications these edges are the relationships people share.

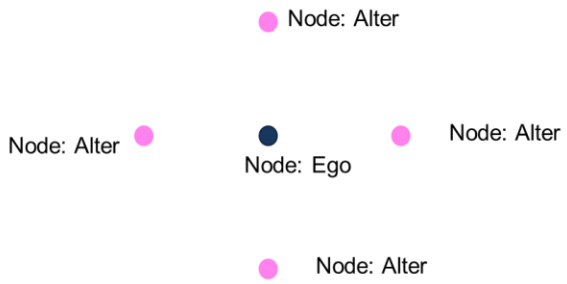


Figure II.3 Network Diagram Nodes

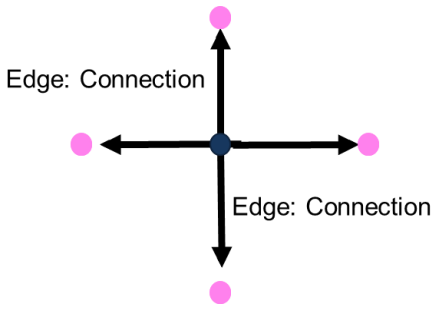


Figure II.4 Network Diagram Edges

2.3.1 History

SNA originates in the 1800s and the work done by Emile Durkheim and Ferdinand Tonnies [16], [25], [33], [36]. Durkheim believed that society consists of individuals who interact with one another, therefore, to generalize about society an analysis of the individual was insufficient [33]. The interaction and cooperation, or lack thereof, between individuals is what defined a society. Tonnies added to this theory with a definition of social groups as people who share values and beliefs. Soon after, Georg Simmel looked at the size of those social groups in the context of social networks [9], [16], [33], [34].

Jacob Moreno is widely considered to be the father of SNA [34], [35]. In the 1930's he researched one of the first methodologies for collecting data on social groups, defining a new academic concept that previously was loosely characterized. This methodology is called the "Sociometry Method" and became the principle of quantitative analysis of social groups [9], [10], [25], [34], [35], [37], [38]. Moreno believed relationship ties helped to determine the structure of an organization [9], [10], [25], [34], [35], [37], [38]. Those relationship ties were based on a pattern of observable interpersonal choice [9], [10], [25], [34], [35], [37], [38]. Consequently, social configurations had definite and discernible structures [9], [10], [25], [34],

[35], [37], [38]. With knowledge of structures, researchers could then begin to understand what influenced people and the underpinnings of society [9], [10], [25], [34], [35], [37], [38]..

The "Sociometry Method" consisted of a series of questions to determine the positive and negative relationships people had in a defined group [9], [10], [25], [34], [35], [37], [38]. Once the relationships had been identified and recorded, a diagram called a "Sociogram" was created to visualize those connections [25], [34], [35], [38], [39].

After Moreno's contribution, from the 1940s-1960s, little was done to advance the general acceptance of the theory [9], [10], [25], [34], [35], [37], [38]. It wasn't until the "Harvard Renaissance" in the 1970s that SNA became more widely accepted [9]. Under Harrison White, a group of students published several papers on social networks that became well known in the community of social science [9], [16], [34]. In the 1980s SNA was applied to economic phenomena; the general idea being that economics is rooted in social structure. [39] [9], [16], [34]. The 1990's brought the analysis of a new form of capital, social capital, and structural hole theory [40].

More recent applications of SNA can be seen in the 2016 investigation of Hilary Clinton's private email network [13]. When the news reported a cache of thousands of emails on a private server, the public wanted a way of understanding and learning from such a large pool of data. Therefore, numerous examples exist that use the information from these emails (i.e. sender and receiver) to build a visualization of those connections [13].

One of those examples came from Céser Hidalgo, leader of the Collective Learning Group at the MIT Media Lab, who created a tool called Immersion [12]. The Immersion tool was initial developed as a means for "self-reflection, art, privacy, and strategy" [12]. However, with

the Clinton data readily available, this was a seemingly perfect, practical application of the tool. This visualization was a first in that it provided a way of understanding the data that was clearer and faster than sorting through thousands of emails one by one. Hidalgo said:

> "The tool represents a different form of data reporting, or data journalism: one where people are provided with a tool that facilitates their ability to explore a relevant dataset, instead of being provided with a story summarizing a reporter's description of that dataset." [40]

By visualizing the large network of communication, people were able to form their own, potentially well informed, opinions on the situation [41]. The visualization of Clinton's email network sparked a conversation, continued in this thesis, about the potential of email archives as a source of data for social network analysis.

2.3.2 Analysis

Analysis of a social network is the basis for understanding what those connections mean for the structure of the organization. For this research, the analysis of SNA will be used to characterize the data collection methods. Analysis is broken down into four categories; transactional content, nature of the links, structural characteristics and network roles [34], [35].

These categories can be separated into qualitative and quantitative observations. Transactional content is the qualitative observation of what is being exchanged [34], [35]. This content can include experience, money, services, goods, or favors [34], [35]. The nature of the links is also a qualitative connection indicating the strength of the connections [34], [35]. This can include observations on intensity, reciprocity, clarity of expectations and multiplicity [34], [35].

Structural characteristics are quantitative observations of the network. Density is the general level of linkage among nodes in a network [34], [35]. Below is the formula for calculating density as a metric. Given n number of nodes, there is E number of potential edges. Density is then calculated by dividing the number of actual nodes to the E number of potential edges.

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

$$n - number\ of\ nodes$$

$$E\text{ -}Maximum\ number\ of\ potential\ Edges$$

Equation 1

$$Density = \frac{e}{E}$$

$$e - number\ of\ nodes\ in\ graph$$

$$E - maximum\ number\ of\ potential\ edges$$

Equation 2

A network with a density of 0% will have no connections, whereas a complete graph in which all points are connected to every other point will have a density of 100% [34], [35]. Figure 2.5 shows an example of a low density network which is dependent on the number of links that have been formed. Given five nodes the potential number of edges based on Equation 1 is ten. With four edges amongst the five nodes the density of this figure is 40%. In contrast Figure 2.6 represnts a graph with a higher density. With the same number of nodes (five), but eight edges, the density for Figure 2.6 is 80%.

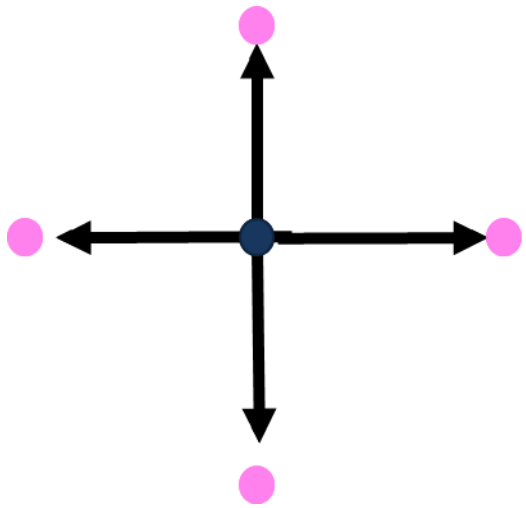


Figure II.5 Low Density

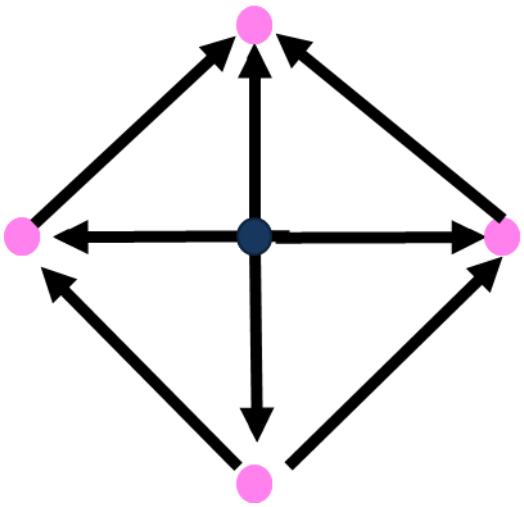


Figure II.6 High Density

In the context of social networks this represents a greater number of relationships in an organization [34], [35]. Knowledge of the density associated with an organization can help identify key roles in the network, as often, the person with the highest density in a network has the greatest impact on that network [22], [31], [33], [34]. The impact in the case of work relationships could be that the person with the highest density is the person who people go to most often for help, or they are a source of unique or important of knowledge.

Modularity is a measure of the presence of groups within a network [24], [34], [35], [42]. The presence of modularity by itself is not good or bad; context matters. Complex systems use modularity to reduce complexity and allow for adaptability [24], [34], [35], [42]. Figure 2.7 is an example of a modular network. Networks that have a high modularity have nodes with greater number of connections between nodes in the same module and few connections between modules. Gephi calculates modularity using the Louvain method for community detection in large graphs. It is a complex algorithm that associates the propensity for nodes to be in groups by leveraging the density of random groups of nodes. The higher the density of between random nodes the more likely they are to be in a module [43]. The example in Figure II.7 shows a module of a network circled in yellow.

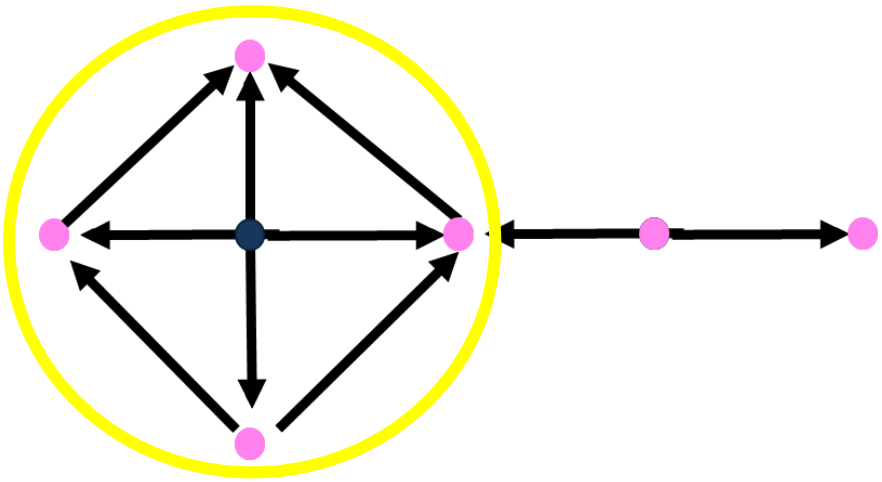


Figure II.7 Network Modularity

2.4 Data Collection

2.4.1 Methods

Analysis can only be made possible with data. Data collection focuses on two aspects of a person's connections: name generation and name interpretation. Name generation is recalling name in the person's social network, while name interpretation is giving meaning to those names

[9], [32], [36]. A balance is needed when collecting data; give people too much to think about and they will be overwhelmed, too little and the data will not be complete [10].

The observational method involves a researcher observing an organization's communication over a period of time [10].The observer typically imbeds in an organization to watch and record the communication happening in that organization over a defined timeframe. Due to the time limitations of the researcher, this would be anywhere from one hour to one month [10]. This method is great for small organizations as it is limited by what the researcher can observe. In large organizations, the researcher would be unable to observe all the communication happening as the researcher cannot be in more than one place [10]. This methodology is great for observing communication patterns, groups, and communication roles. Analysis is based on qualitative observations [10]. In terms of validity, this tool does not rely on self-reporting of individuals on their communication therefore reducing that bias [10]. However, this method is highly susceptible to the Hawthorne Effect, or the alteration of behavior by the subjects of a study due to their awareness of being observed [10].

The diary method asks each member of an organization to keep a diary/log of communication over a specific period of time [10]. The preponderance of effort falls on people within the organization as they must record all communication themselves. The diary entry is a very structured form respondents will either fill out after each interaction or at set times throughout the day [10]. While there is no limitation on the size of an organization, there is a greater amount of effort required from the respondents [10]. This can cause respondents to become overwhelmed, leading to incomplete data being collected and a shorter time span of data collection [10].

Sociometric questioning, the most widely used methodology, is the process of asking every member in a defined set of people who they interact with most [10]. This questioning is usually conducted in the form of a survey or interview [10]. This method is preferred due to the large amounts of data that can be collected about the stable lines of communication rather than communication over the course of a defined set of time [10]. Name generation is solicited by providing a roster or fixing the number of names a person can specify [10], [44]. Name interpretation is gathered through either ordering the connections based on the question asked or rating each individual connection on a set scale [10], [44]. Size of an organization in this method is not necessarily a limitation, however, each member of that organization needs to be interviewed or surveyed [10], [44]. In terms of validity, this method is highly dependent on the questions that are asked and how those questions are interpreted by the respondents, including respondents truthfulness [10].

Archival data is the final method discussed in this research of collecting information on a social network. This method focuses on ego-centric networks, and data is gathered through previously recorded modes of communication like phone logs, email traffic, or internal mail [10]. It is considered an unobtrusive way to gather a large amount of data[10]. The data can cover a large number of people over a wide range of time [10].

Recent examples of building network diagrams from email traffic can be seen in the MIT Immersion Tool and an analysis of the Clinton email network [12], [13], [41]. These tools can harvest existing data (e.g. e-mails or log-files) to generate network diagrams. In the case of the MIT Immersion Tool, the network is generated within minutes [12], [41]. However there are few true methodological contributions to the study of archival data collection [10].

In the current technical work environment, email is a good source for archival data [15], [45]–[48]. It is scalable and practical for most settings as it low cost, meaning it does not require a lot of time and effort to gather the data [15]. Most email servers log email data for years and this data is readily available for use [15], [45]–[48]. Research also supports that email is a good basis for work communication as email is the least likely form of communication to be used socially [15].

In a study by Wellman in 2005, an organization of 80 people was asked about their communication habits both face to face and computer mediated communication (i.e. email), the researchers through observations, surveys and interviews identified the following phenomena: hyperconnectivity, local virtuality, and glocalization [15].

Hyperconnectivity is the availability of people for communication anywhere and anytime [15]. Using email and other forms of computer mediated communication (CMC), a person's network is constantly within reach [15]. This expands their network to bigger than just their local communities. Local Virtuality is the pervasive use of CMC for interaction with physical proximate people [15]. People will use email as a mode of communicating with people who are geographically close to them; whether that be in the same building, office or room [15]. Glocalization is constraint-free communication combining global and local connectivity[15]. Networks that are glocal have representation from a global scale but also the local level. Glocalization allows people access to new and varied information[15]. Access to information can lead to innovation as the cost of information is high [15], [21], [22]. These phenomena about email communication were observed through interviews and have yet to be observed through exploration of email data [15].

2.4.2 Method Selection

Due to the especially large size of the USAF, the data collection method must be applicable to large organizations and require the least amount of effort to use. A method that is difficult to use will never be implemented or will produce valuable results. From the research, Figure 2.7 was created to visualize how the size of an organization impacts the effort required of the method. The graph is broken up into four quadrants that depict the target area for the application to the USAF. The green square includes data collection methods that are good for large organizations and require minimal effort. Yellow squares represent methods that are either for small organizations with minimal effort or large organizations with maximum amounts of effort. And the final red square represents maximum effort required for small organizations. From this graph the small world, archival and sociometric questioning fall into the green target quadrant.

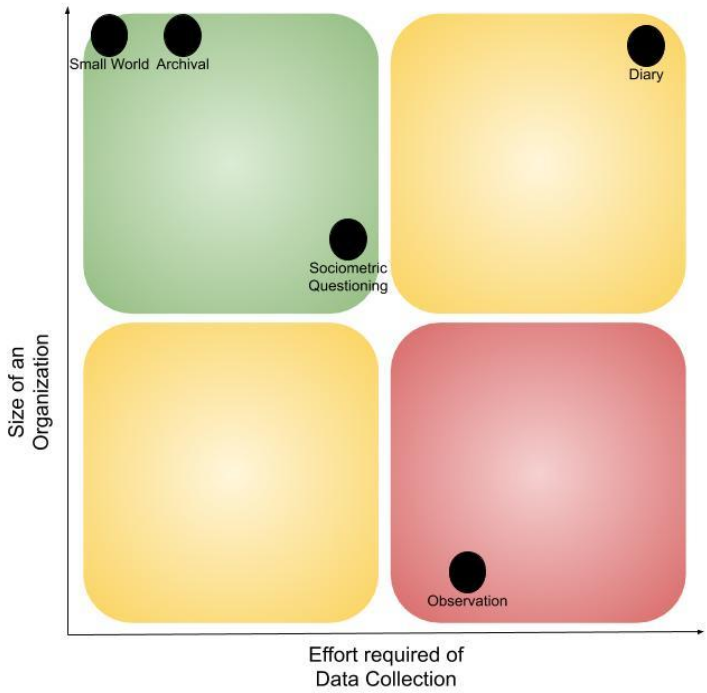


Figure II.8 Size of an organization versus effort required of data collection [10]

In addition to the large size of the USAF, a number of constructs were identified as being critical to this environment. These constructs are divided into three categories: phenomena, setting and validity. The constructs in Table II.1 will be used to observe specific phenomena. These phenomena were identified as they critical to what we want to observe. These constructs must be met for inclusion in the research.

Table II.1 Phenomena Constructs for Data Collection Method Comparison

Construct	Definition	Importance
Name Generation	Names actual persons in a network without limitation [44]	Builds the nodes included in the network
Name Interpretation	Meaning behind names	Builds the connections included in the network
Longitudinal	Study of group of people over time	Collect the most data possible
Work Communication	Data collected represents work communication	Studying work environment

The following constructs are setting specific. These setting characteristics describe the environment to study in this research. Methods that meet these criteria will be considered for inclusion.

Table II.2 Setting Constructs for Data Collection Method Comparison

Construct	Definition	Importance
Large Population Size	The scale to which the network represents a population	USAF is a large organization
Minimal Organizational Awareness	The amount of information needed from an organization to get results	It is not practical to gather information on all organization before conducting research
Discreet	Not readily seen or noticed	Data collection must be robust without interfering with daily work

The following constructs address validity.

Table II.3 Validity Constructs for Data Collection Method Comparison

Construct	Definition	Importance
Limited Hawthorne Effect	Effect of the observer on the observed [15]	Unbiased data

These constructs were then used to compare the methods of data collection used in SNA. These methods are observational, diary, small world, interview and archival data collection. Each method received a check mark for the corresponding construct that it supported.

Table II.4 Data Collection Method Comparison

Construct	Observational	Diary	Small World	Interview	Archive Data
Phenomena					
Name Generation	✓	✓	✓	✓	✓
Name Interpretation	✓	✓	✓	✓	✓
Longitudinal	✓	✓		✓	✓
Work Communication				✓	✓
Setting					
Large Population	✓	✓	✓		✓
Minimal Organizational Awareness		✓	✓	✓	✓
Discreet			✓		✓
Validity					
Limited Hawthorne Effect					✓

From this comparison the interview method and archive data meet the requirements for the phenomena. While the small world method meets all the constructs for the setting the phenomena that it observes does not support this research. Therefore, the research will move forward to characterize the interview and archival data methods.

III. Methodology

The purpose of this preliminary study is to provide a comparison social network analysis data collection techniques, sociometric questioning and archival data, toward the improvement of organizational behavior. This methodology takes an iterative approach based on research by Eisenhardt on building theories from case studies [14]. Case study research helps to understand the dynamics present in a single environment [14]. Specifically, it can help provide description, test theory or build theory. The process is as follows: identify research setting, case identification and data collection, within case analysis, cross case analysis, building theory, and enfold literature [14]. The within case analysis through cross case analysis is an iterative process.

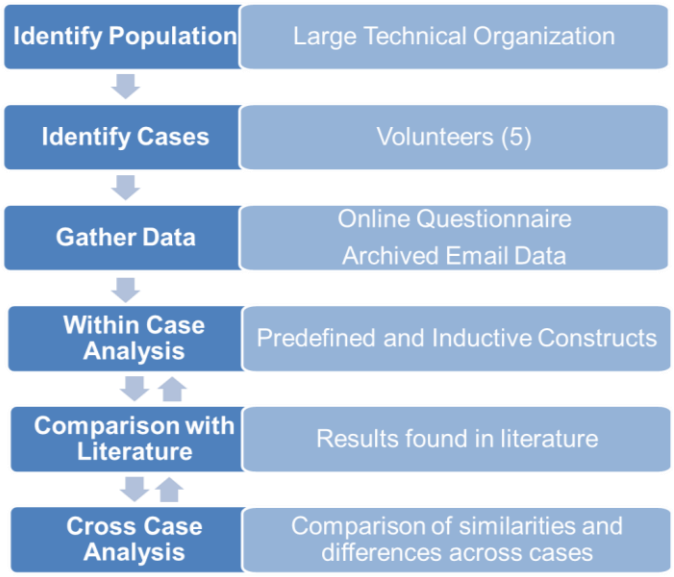


Figure III.1 Research Design

3.1 Research Setting

The intended population was a large technical organization, the USAF. A technical organization, AFMC/A1M, the organization for Manpower, Personnel and Services, was approached to be the research subjects. Unfortunately, due to the time constraints of the Master's

program a simulated environment was used in the place of a technical organization. To simulate that environment this research used the pool of faculty and staff at AFIT.

Faculty and Staff at AFIT were candidates for this research because of their position at AFIT was less transient then students and it was assumed that they would use their email in a similar manner to people in a program office in which daily responsibilities include project management. The pool of faculty and staff collaborate within AFIT, across department boundaries as well as outside of AFIT. This pool of participates was also easily accessible should there be any questions or concerns during data collection.

3.2 Case Study Selection

Case studies were identified from a call for volunteers. Further, self-elimination occurred after the potential pool of participants were emailed with a description of the research and data that was required. Therefore, five participants volunteered. The names of the volunteers are masked for privacy by a singular letter; A-E. For each case data was collected for both sociometric and archival data.

3.3 Data Collection

For each case, data was collected using the sociometric questioning technique and archived data. The sociometric questioning method, an established method in the literature, involved sending each volunteer an online questionnaire. The volunteer was asked a series of questions from the literature:

- "Think back over the past month. Consider all members of the organization here that you have contacted during business hours. Which ones have you spent the most time with on a business basis? With whom do you spend the most time with getting work done?" [37], [49]

- “Name five assistants/subordinates”
- “Name five superiors/associates at the same level” [37], [49]

Then to generate an interpretation of those names the volunteer was then provided with the following instruction:

- “Rank all members in accordance to how much time you have spent communicating with them.” [37], [49]

This method will generate a list of recalled members of the volunteer’s organization ranked according to frequency of conversation. The form that the volunteers filled out is referenced in Appendix A.

For archival data collection two files are needed. Steps to generate the files can be found in Appendix B. In general, Outlook was opened. Then, under file and options the advanced tab was selected. Then, the export feature was selected. This feature exports Outlook folders to a file for use in other programs. By using this feature, it is possible to extract the information to a Comma Separated Values (CSV) file which was later processed in Excel and Gephi, an open source software for social network analysis. The “from”, “to”, and “Carbon Copy (CC)” fields the names of the people were extracted from Outlook and imported to a CSV file. This process was repeated to collect data on the Inbox and Sent folders for each volunteer. Table 3.1 provides the data from three separate e-mails, each coming from Jane Doe.

Table III.1 Example Data Output

From: (NAME)	To: (NAME)	CC: (NAME)
Jane Doe	John Smith	Sally Brown
Jane Doe	Tony Stark	
Jane Doe	Tom Brady	

Further data conditioning may be necessary depending on the format of the downloaded archived email data. E-mail formats can be organization specific and this influences the format of data in the CSV file. For example, in the case of the information that was pulled from Outlook for this study, the name that was exported from Outlook was in the format of "LAST NAME, FIRST NAME Title Organization."

From this data, whether it be from sociometric questioning or from the archived data a visualization of that data was created using an open source network visualization program, Gephi. There are several tools available for social network analysis. However, unlike many available tools, this is a free tool to use with a clear front end that reduces the need for additional coding by the user. It was developed in 2008 by a group of college students in France [50].

To visualize the data in Gephi, a force direct algorithm was used for node placement. In visualizations of social networks, nodes are not tied to Cartesian coordinates, but physical laws of attraction and repulsion to inform network appearance. Nodes are repelled from one another and edges are attracted to their nodes like springs [51]. Force directed algorithms are the ideal algorithm for social network visualizations [51]

The visualization of the edges in Gephi was weighted to underscore the frequency of communication between two nodes. Heavier lines that connect two nodes represent the strength of the connection. These are connections where information/resources flow between the two people often. This can mean that these people have the knowledge that other people are looking for or that they are leaders in their organizations, independent of the hierarchy of the organization.

For ease of visualization the weight of the graph was levied to increase the size of dominant nodes. Node size in the archival data is based off weighted in degree for the sent data, weighted out degree for the inbox and weighted degree for the combination. For example, in the sent data, a larger node means that person received more emails than others. Color of the nodes is based on modularity.

3.4 Data Analysis

Data analysis took a phased approach. First, a comparison of the sources of archival data collected provided a basis of comparison for the next step which was a comparison of the sociometric questioning technique to the archival data.

Table III.2 Data Analysis

Phase	Research Question	Comparison Constructs
1	What is the effect of archival data inclusion on the completeness of the observed network?	Completeness Glocalization
2	How similar is the structure obtained with sociometric questioning to the representative subset of archival data (e.g. subgraph isomorphism)?	Graph Matching Completeness
3	What effect does the data collection method have on the observation of predefined network characteristics?	Presence Position Placement Number of Nodes Number of Edges Modularity Density

3.4.1 Phase 1: Characterization of Archival Data

Phase one of the research addresses the question: what is the effect of archival data inclusion on the completeness of the observed network? This phase is accomplished using two tests.

The archival data included the sender, receiver, and CC fields of any email in the Sent Folder and the Inbox. These email folders are also presented in a combined view, as well. The

data is broken down further within each box to two groups. The first data set includes just the sender and receiver field from the emails, the second data set includes the sender, receive and the CC field. This is expressed in the following table.

Table III.3 Breakdown of Archival Data

Source	Set	Field
Sent Folder	1	To, From
	2	To, From, CC
Inbox	3	To, From
	4	To, From, CC
Combined Data	5	To, From
	6	To, From, CC

For each case, data was collected and visualized for each data set. Visualizations for sets 1-2 were produced using the Fruchterman-Reingold Algorithm. This algorithm is a force directed algorithm that utilizes the spring force for attraction and electrical forces for repulsion [51]. Data sets 3-6 were visualized using the Force Atlas algorithm.

The first test uses the construct of completeness. Completeness is defined as having all the necessary or appropriate parts. A complete data source would represent the most amount of nodes (people) and their edges (connections).

The second test is based on literature on email usage in a technical environment. The following phenomena were observed in interview research on email usage. The test is to identify whether these phenomena are observed in the data sets. Of that phenomena, this research addresses the presence of the phenomena of glocalization [15]. In the scope of studying an organization's behavior, the presence of glocalization allows for a greater pool of knowledge and therefore increased potential for innovation.

To confirm the presence of glocalization, the simulated population of AFIT faculty and staff was considered. If the pool of faculty and staff at AFIT is 250 people, this means that if the networks only represent the local population, no network should have more than 250 people. If a network includes more than 250 people than we can conclude the presence of an outside or global population and therefore the presence of glocalization.

Table III.4 Phase 1 Construct Definition

Metric	Definition	Implication
Glocalization	Constraint-free communication combining global and local connectivity [15] Number of nodes greater than 250	Network reaches beyond local population of faculty and staff at AFIT and includes a "global" network

3.4.2 Phase 2: Extent of Similar Data

With the most complete archival data set phase two address the second research question: how similar is the structure obtained with sociometric questioning to the representative subset of archival data (e.g. subgraph isomorphism)? This will be accomplished through two tests.

The first test will identify if all the people identified in the interview are present in the email data. The second test will identify if all the connections identified in the interview exist in the email data. This is a simple comparison of the list of people and their connections in the interview to the list of people in the email data.

3.4.3 Phase 3: Comparison of Methods

Phase three will address the third and final research question: What effect does the data collection method have on the observation of predefined network characteristics? This will be accomplished through two tests.

Test one address the presence, position and placement of the interview data in the email data. When comparing the two data sets in their totality, these metrics are used to compare where in the email data does the interview data appear. These metrics are based on the network metric of degree. Degree is the extent to which the node has an impact on the network [24], [34], [35], [42]. It is the total number of edges associated with a given node. In-degree, in a directed network is the total number of edges coming into a node, whereas out-degree is the total number of edges leaving a node [24], [34], [35], [42]. In the context of emails, degree is the total number of emails sent and received to a given node, in-degree represents the number of emails sent to a node and out-degree represents the emails sent by a node.

Table III.5 Phase 3 Construct Definitions

Metric	Definition
Presence	The occurrence of the self-identified top ten in the top ten of the data set after being sorted by degree
Placement	The matching of the self-identified top ten to the particular place in the top ten of the data set after being sorted by degree
Position	The location of the self-identified top ten in the data set after being sorted by degree, this can be described as the rank position of the people that were self-identified

Test number two is based on literature on social network analysis, which guided the selection of communication network characteristics that were used as a basis for characterization of the data collection methods. These metrics were number of nodes, number of edges, modularity, density, glocalization and hyper connectivity. Further analysis was also conducted regarding the ranking of the people the volunteers said they work with. These constructs

consisted of presence, placement and position. These metrics of comparison are defined in Table III.6.

Table III.6 Phase 3 Construct Definition (continued)

Metric	Definition	Implication
Number of Nodes	A count of the number of people, represented as nodes, are present in the network [24], [34], [35], [42]	Size of the network
Number of Edges	A count of the number of connections between nodes, represented as edges, are present in the network [24], [34], [35], [42]	Size of the network
Modularity	A measure of the presence of groups within a network [24], [34], [35], [42]	
Density	The proportion of all links that are actually present to all possible links [24], [34], [35], [42]	The general linkage among members

IV. Results and Analysis

The analysis in this chapter is in three parts, beginning with an exploration for the archival analysis, establishing what data provides for the most complete network. Next, it is known a priori that the archival data contains at least an order of magnitude more nodes that the survey data. Knowing that the one network (archival) is larger than the other (survey), further analysis is accomplished to determine the degree of matching between the survey data and its representative subgraph within the archive data. This comparison establishes whether and the extent to which, the interview data is a subset of the archive data. Finally, the networks obtained via both methods are compared based on established network constructs. This last analysis, considers the effects of the data source on the higher-level network constructs observed.

4.1 Archival Data Selection

The email data was characterized using two metrics: completeness and glocalization. Completeness is a measure of the size of the network and glocalization informs on the local and global reach of the network. Both metrics characterize the social networks of a technical organization which can be used towards improved organizational behavior.

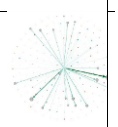
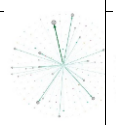
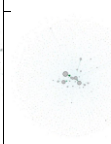
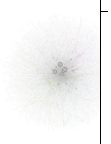
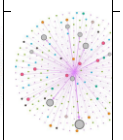
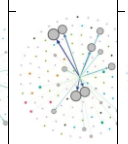
4.1.1 Completeness

Archival Data collection was obtained from two different outlook folders, sent mail and the inbox (received mail). This artifact of data collection allowed for a consideration of data source (inbound versus outbound) on the completeness of the network observed. Further, within each box there are two types of recipients, direct recipients and carbon copies (CC), people added to the correspondence outside of the direct line of communication. Between data source

(sent, inbox, combined) and recipients included (to/from, cc), it is possible to have six permutations of networks for each informant.

The networks attained with the different sources as well as recipients are considered. Each network was compared using the metrics number of nodes, number of edges, modularity, and density to determine completeness.

Table IV.1 Sent Data Cross Case Analysis

Case	A		B		C		D		E	
	To/From	+ CC	To/From	+ CC	To/From	+ CC	To/From	+ CC	To/From	+ CC
Visualization										
#Nodes	189	214	1542	1778	150	169	181	245	259	294
#Edges	189	214	2040	2849	150	168	181	244	259	294
Modularity	.002	.004	0.154	0.191	0	.004	0	.002	0	0
Density	.005	.005	.001	.001	.007	.006	.007	.004	.004	.001
Transactional Content	Sent Mail	Sent Mail	Sent Mail	Sent Mail	Sent Mail	Sent Mail	Sent Mail	Sent Mail	Sent Mail	Sent Mail

Nodes are colored and sized based on weighted in degree. The grey and large nodes are the people who have the highest weighted in degree, meaning the people who received the most emails

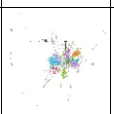
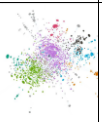
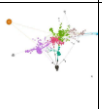
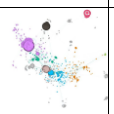
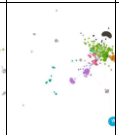
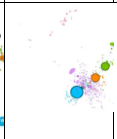
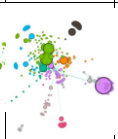
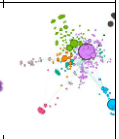
The common theme among sent data across the cases was that the typical force directed algorithm did not produce a useful visualization for any of the cases. The visualization that was used produced a concentric alignment of the nodes. The usual layout algorithm is based off of shared connections among the members of the network. Because this data is sent mail there is little to no modularity among the members of the network. Even with the addition of CC data there was little increase in the modularity across the cases.

For four out of five of the cases, the number of people represented as nodes on the visualizations was an order of magnitude larger than that of the sociometric questioning. For the leftover fifth case, the number of people represented was well over one thousand, significantly higher than the other cases. A number of causes could have contributed to this outcome, the first being that in any case the archival sent data is over a varied amount of time. A lesson learned from this preliminary study is that the date and time should be collected from outlook. This would insure that the same time period is covered amongst all the compared cases.

A second possibility is that Case B, the case with the highest number of people represented, could have a different role than the other Cases. This role may enable them to talk to more people more frequently. While there are roles associated with a social network analysis another limitation of this method and research was missing personal data from the volunteers. For example information regarding position, job, or years of service could have been useful. As stated in Chapter 1, our case selection considered literal replication, faculty and staff within a university. More deliberate and diverse case selection in future research will provide for theoretical replication and opportunities for more robust theory building.

As a departure from the transactional content of the sociometric data, which asked specifically for those the volunteer may work with, archival sent data represents only people the volunteer has contacted. This could imply that sent data only represents people you work with and that is it. This type of data is directional and does not imply a reciprocity of connection.

Table IV.2 Inbox Data Cross Case Comparison

Case	A		B		C		D		E	
	To/From	+ CC	To/From	+ CC	To/From	+ CC	To/From	+ CC	To/From	+ CC
Visualization										
#Nodes	1133	1460	3086	3709	1079	1312	1142	941	386	431
#Edges	2645	3621	7857	10911	2038	2581	1671	1661	806	1004
Modularity	.592	.607	.572	.553	.733	.722	.678	.633	.519	.522
Density	.002	.002	.002	.001	.002	.002	.001	.002	.005	.005

Visualizations based off of the force directed algorithm became more differentiated with the inbox data. With this differentiation we observe the emergence of groups. The force directed algorithm alters node positions based on mutual connections and frequency of contact, yielding clusters (groups) and patterns in the visualization.

Case B and Case E have similar moderate modularity; moderate as defined as between .4 and .6, whereas the other cases have high modularity, defined as .6 and above. The visualization of these cases is more tightly grouped, and each group is seemingly independent of one another.

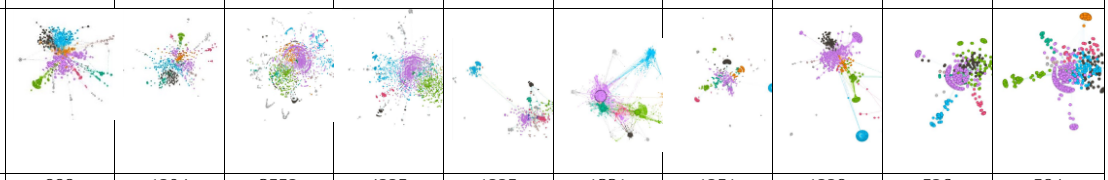
This is in contrast to Case B and E, with lower modularity and the groups become harder to discern without the coloring of the nodes. While B and E represent the furthest ends of number of nodes across the cases this pattern across the visualizations is hard to ignore. More cases would be needed to develop a theory on modularity.

Within the archival inbox data the number of people represented is more diverse. One case only represents approximately 400 nodes while another case represents close to 4000.

What becomes clearer with the inbox data is the effect of the cc field on the overall network. In each case the addition of the cc field brought with it a person to connection ratio of ¼, meaning that with each new person came four new connections on average.

The archival inbox data represents the connections that come to the volunteer for work related matters. Again, this in itself does not imply a reciprocity of connection. Therefore, this is still a directed network.

Table IV.3 Archival Combination Data Cross Case Comparison.

Case	**A**		**B**		**C**		**D**		**E**	
	To/From	+ CC	To/From	+ CC	To/From	+ CC	To/From	+ CC	To/From	+ CC
Visualization										
#Nodes	998	1204	3552	4225	1225	1331	1251	1229	526	584
#Edges	2374	3126	9460	13227	2212	2740	1832	2197	1064	1294
Modularity	.621	.641	.53	.516	.748	.6940	.666	.632	.482	.422
Density	.002	.002	.001	.001	.001	.002	.001	.001	.004	.004

Few differences exist between the inbox-only and the combination data. The visualizations, modularity and density are stable or consistent, with similar patterns across the cases being observed.

The combination data does not simply add the number of nodes from the sent and inbox. There are fewer nodes than would be represented had the two numbers just been added. One explanation is that there is an overlap between the people who email you and the people you email. This also implies that this data set is both who works with the volunteer and who the

volunteer works with, making this the most similar archival data to the sociometric questioning data.

If the metric of completeness is considered, the combined data source with the to/from/CC fields is the most complete data. In every case this view represents the largest amount of nodes and edges, people and their connections, and therefore the most amount of data. The more sources of data that are included the more data that therefore is represented. In each case, one hundred percent of the data is represented in the combined source with to/from/CC data.

Table IV.4 Percentage of Nodes in Each Data Source

Case	Inbox		Combined	
	To/From	To/From/CC	To/From	To/From/CC
A	68.36	82.47	77.60	100.00
B	73.04	87.79	84.07	100.00
C	81.07	98.57	92.04	100.00
D	75.22	91.29	98.24	100.00
E	66.10	73.80	90.07	100.00

Table IV.5 Percentage of Edges in Each Data Source

Case	Inbox		Combined	
	To/From	To/From/CC	To/From	To/From/CC
A	65.56	86.33	73.05	100.00
B	59.40	82.49	71.52	100.00
C	74.38	94.20	80.73	100.00
D	75.60	76.06	83.39	100.00
E	62.29	77.59	82.23	100.00

4.1.2 Glocalization

Glocalization is defined as the presence of global and local connectivity. This phenomena was measured in the archival email data through a comparison of the potential pool of faculty and staff at AFIT. Because the volunteers were from a pool of 250 faculty and staff, if a network experiences glocalization than the number of nodes represented in the network will be greater than 250.

The following table compares the number of nodes (people present in the network) to the 250 benchmark. Without access to the names of the network participants an assumption is made that any number of nodes less than 250 does not experience glocalization. If a network does have more than 250 nodes than it definitely reaches beyond the potential local pool of faculty and staff at AFIT. For ease of comparison cells that are greater than 250 are highlighted in green.

Table IV.6 Cases with Greater than 250 Nodes

Case	Sent		Inbox		Combined	
	To/From	To/From/CC	To/From	To/From/CC	To/From	To/From/CC
A	189	214	998	1204	1133	1460
B	1542	1778	3086	3709	3552	4225
C	150	169	1079	1312	1225	1331
D	181	245	941	1142	1229	1251
E	259	294	386	431	526	584

The sent folder for cases A, C, and D do not experience glocalization as defined above. However, Case B and E do have more than 250 people present in their sent folder. While this is not true for all the cases we cannot conclude that the sent folder would be a good representation of glocalization. When looking at the networks from the inbox and the combined data, all the cases present with the potential for glocalization.

If the inbox and combined data are then isolated and the number of nodes put on a scale that identifies in what data set the most nodes are present, the combined data set for each cases is the has the greatest amount of nodes. From our presented definition of glocalization, this would imply that the combined data including the To/From/CC fields experiences the most glocalization.

4.1.3 Summary

Characterization of a communication network is the basis for understanding and then improving a technical organization. Getting the full picture is the first step.

To get the most complete picture of the data based on number of nodes and number of edges from the archival data sources it is best to use the combined data with the to, from, and cc fields. The combined data source with the to, from and cc fields also expresses the greatest potential for glocalization.

The data source with the most complete representation of a network will then in turn give the greatest characterization of that organization's communication network. Completeness is also important as a baseline for comparison. The interview data is complete in its scope so therefore the email data must be complete in its scope as well.

4.2 Extent of Similarity of Data

The extent of the similarity of data is determined using two tests. The first test determines if the same people identified in the interview are present in the email data. Then, looking only at the names identified in the interview, the network structure (edges) of those names in the email is compared to that of the network from the interview.

4.2.1 Test: Do the names from the interview appear in the email data?

Test one is a simple check of the names identified by the interview for each case is present in the email data. For each case, every name identified by the volunteer is also present in their email network with the exception of one name in Case C. While this test may seem obvious it helps identify the similarities and differences of the two methods. The email method still captures 98% of the names that were identified in the interview.

*Case E only identified 6 names in the interview.

Table IV.7 Node Matching

Name	A	B	C	D	E
1	✓	✓	✓	✓	✓
2	✓	✓	X	✓	✓
3	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓
6	✓	✓	✓	✓	✓
7	✓	✓	✓	✓	-
8	✓	✓	✓	✓	-
9	✓	✓	✓	✓	-
10	✓	✓	✓	✓	-

4.2.2 Test: Do the connections from the interview appear in the email data?

The next step was identifying if the connections identified between the ten people in the interview were present in the email data. Cases A, B and C all identified connections between the people identified in their interview. Case D and E did not respond to that question.

The following tables represent the edges from the interview and email data from all five cases. The interview data, represented in the top column and row, was ranked from one to ten.

From there, the connections were added to the matrix and then compared to the connections present in the email data.

When the edges from the interview data are identified in the email data, there is a gain of data that was not represented in the interview data. Specifically, with Case D and E the email data allows for an understanding of the connections among those identified in the interview data. There is also data found in the email data for Cases A, B and C that was not represented in the interview data.

The tables on the following page depict the edge subgraph comparison. For Cases A-C, duplicated data is represented with a “2” in a gray box. Boxes that only have a “1” in the cell were edges that were only represented in one method. The method is then identified by the color of the cell. Green cells are data gathered solely from the email data while red is data that is only represented in the interview.

For each case the majority of the data is duplicated or represented in both the interview and email data. With the exception of Case C, email data is the most complete view of the volunteer’s network. Even though slightly different edges appear in each case, the email data represents more connections than the interview data.

Table IV.8 Edges subgraph comparison

A	1	2	3	4	5	6	7	8	9	10
1										
2	2									
3	0	2								
4	2	0	0							
5	2	1	0	2						
6	2	1	0	2	2					
7	2	2	1	2	2	2				
8	2	1	0	2	2	2	2			
9	2	2	0	2	2	2	2			
10	2	0	0	2	2	2	2	2	2	

B	1	2	3	4	5	6	7	8	9	10
1										
2	2									
3	2	2								
4	2	2	2							
5	2	1	2	2						
6	2	2	2	1	1					
7	2	2	2	2	2	2				
8	1	1	0	0	0	2	0			
9	2	2	0	1	0	0	1	0		
10	0	0	0	0	1	2	0	1	0	

C	1	2	3	4	5	6	7	8	9	10
1										
2	1									
3	2	1								
4	2	0	2							
5	2	0	2	2						
6	1	0	1	0	1					
7	2	0	1	2	1	2				
8	0	0	1	0	1	0	0			
9	0	0	1	0	0	0	0	0		
10	1	0	0	0	0	0	0	0	0	

D	1	2	3	4	5	6	7	8	9	10
1										
2	1									
3	0	1								
4	1	1	1							
5	1	0	0	1						
6	1	1	1	1	1					
7	0	1	0	0	0	0				
8	0	1	0	1	1	1	0			
9	1	0	0	0	1	1	0	1		
10	1	1	1	1	1	1	1	0	0	

E	1	2	3	4	5	6
1						
2	1					
3	1	1				
4	1	1	1			
5	1	1	1	1		
6	1	0	1	0	0	

4.2.3 Summary

There are differences in the interview network and the representative sub-network attained with e-mail. The capture of nodes was high, where only one interview node was missing in the e-mail archive (1 of 46). While the email captures a majority of the total interactions represented in the interview data, a percentage of the connections were missing. The greatest strength of the email data is in its unbiased ability to collect data. When completing the questionnaire for the interview, Case D and E, did not identify connections among their network. There was necessary, missing network data for Cases D and E due to the bias, awareness and choices of the informants. The email data was not limited by human bias, or ability to answer questions but rather was a representation of email connections.

When trying to characterize an organization to understand an organization's behavior though email data, email data represent more data than the interview data. Our anecdotal indicates that the e-mail provides a more complete network representation that is not subject to bias, limits in interpersonal awareness (e.g. does coworker 6 interact with coworker 8) and choice to omit data.

4.3 Comparison of Methods

4.3.1 Presence, Placement, and Position

Not only is it important to identify the existence of the nodes and edges of the interview in the email data but where in the email data these nodes exist. By identifying where in the email data these nodes exist will speak towards the similarities and differences of the methods. To compare the nodes, the constructs of presence, placement and position were used. These variables are based on the metric of degree.

Degree is the measure of the total number of edges connected to a particular node [24], [34], [35], [42]. Degree informs on the impact of a node in a network [24], [34], [35], [42]. The top ten nodes ranked by degree would imply the ten nodes with the most connections in the network. These metrics are important because the interview data was specifically sorted by intensity of work communication [24], [34], [35], [42]. Each volunteer listed the people they work with in order from most to least communication to get work done.

Presence is defined as the occurrence of the interview nodes in the top ten of the email data set after being sorted by degree. When looking at the combined data source with the to, from and cc fields the interview nodes there is no case where all the interview nodes appear in the top ten. Case D was the only case where the majority of the interview nodes were present in the email top ten.

Table IV.9 Presence of Interview Data in Top Ten of Email Data

Case	Combined Data: To, Form and CC
A	4
B	5
C	2
D	6
E	2

Placement is defined as the matching of the interview data to the particular place in the top ten of the email data set after being sorted by degree. Once the interview data was identified as present in the top ten, the next question was whether the interview nodes were present in the same placement in the top ten. There is no case where the interview data matches the placement in the top ten of the email data.

Table IV.10 Placement of Interview Data in Top Ten of Email Data

Case	Combined Data: To, Form and CC
A	0
B	0
C	0
D	0
E	0

So, if the interview data does not fall within the top ten of the combined data source where does it fit into the email data? Position is defined as the location of the interview data in the email data after being sorted by degree. The position helps inform on where the interview data is located in the email data. The interview data does not represent the email data ranked by degree.

Table IV.11 Position of Interview Data in Email Data

Name	**A**	**B**	**C**	**D**	**E**
1	8	3	19	9	1
2	4	5	-	4	7
3	2	4	5	23	9
4	36	33	13	12	37
5	20	36	10	8	3
6	9	14	59	10	58
7	13	2	18	40	-
8	22	124	41	5	-
9	67	10	102	14	-
10	16	216	112	20	-

Position, placement and Presence give insight to the representation of the interview data in the email data. This can inform on the location of nodes in the email data.

4.3.2 Test: In its totality, how does the email data compare to the interview data?

The final test is to compare the all the data collected in its totality. This informs on the whole communication network of data that is collected by each network.

Table IV.12 Phase 3 Construct Comparison

Case	Interview				Email Archives			
	Nodes	Edges	Modularity	Density	Nodes	Edges	Modularity	Density
A	11	44	.052	.8	1204	3126	.621	.002
B	11	36	.083	.655	4225	13227	.516	.001
C	11	28	.041	.178	1331	2212	.694	.002
D	10	10	0	-	1229	2197	.632	.001
E	6	6	-	-	584	1294	.422	.004

The literature supports that individuals are best at identifying five to seven people in each category they were asked about [15], [16]. In this study, the volunteers were asked to identify five superiors and five subordinates. For the interview data, each case represents 11 people, the volunteer and 10 of the people they work with most often. However, not all the cases identified ten people, thus, there are holes in the data. Even when the volunteer did identify ten people they did not always identify the connections among the people.

When considering the other metrics gathered on the interview data, each case has a low modularity and a high density. A high density implies a high degree of connections between the people identified. And with a high density comes a low modularity, as people are more likely to be from the same group if they have a high density of connections. This could imply that the people that were identified were all from the same local group.

In small populations the high density and low modularity makes sense according to the literature. With only eleven nodes there is a low complexity to these graphs. The cost associated with maintaining connections between ten people is low [5], [9], [24], [40], [42], [52]. The greater the nodes the more complex a network can become. Simple systems experience this tight coupling [5], [9], [24], [40], [42], [52] .

This is something that can be observed in the email data. The size of the networks represented in the email data is orders of magnitude larger than that of the interview networks.

These graphs, as discussed in previous sections, expand well beyond the community of faculty and staff at AFIT. But with this increase in network size come an increase in modularity and a decrease in density [24], [42], [43], [48].

If technical organizations are representative or experience similar characteristics associated with complex systems then the decrease in density is a product of survival of these systems [24], [42], [43], [48]. Due to the cost of connections and maintain connections it is not feasible in such a large system for all the nodes to be connected to one another [21], [24], [42], [43], [48]. The phenomena observed is similar to that of complex computer systems, where developers must make tradeoffs between the diffusion of knowledge and the processing power required to get the same information to all nodes [42], [53].

V. Conclusion

5.1 Conclusions of Research

This research worked backwards from the high-level justification of studying organizational behavior to the data collection methods that support such an analysis. Communication networks were targeted as a source for realizing potential in an organization. This potential was defined through newcomer socialization, network structure and tracing innovation.

The method for studying communication networks is through social network analysis. Unfortunately, much of social network analysis is known for "weak data, strong analysis" [9]. A major critique of work in social science and specifically social network analysis is the focus on the analysis component [9]. If research is to be represented as an equation it would be data plus analysis equals results and recommendations. Therefore, strong analysis on weak data do not yield productive results and recommendations [9]. Strong data along with a strong analysis can inform a strong result[9]. The following questions were answered as part of the aim to improve the weak data associated with social network analysis towards improved organizational behavior.

1. What is the effect of archival data inclusion on the completeness of the observed network?

The archival data that was collected was recorded work emails, specifically emails from the sent folder and inbox including the to/from and cc fields. Data from the sent folder and the inbox was then combined to create a "master" folder. Completeness of the network was defined as representing the highest number nodes and edges. With the addition of each data set, more data was included. With the inclusion of more data came a more complete view of the communication network. The most complete communication network, highlighted in Table V.1, was from the

combined data source, which included the to/from/cc data fields. Having identified the most complete archival data source, the two data collection methods could then be compared.

Table V.1 Archival Data Organization

Source	Set	Field
Sent Folder	1	To, From
	2	To, From, CC
Inbox	3	To, From
	4	To, From, CC
Combined Data	5	To, From
	6	To, From, CC

2. How similar is the structure obtained with sociometric questioning to the representative subset of archival data (e.g. subgraph isomorphism)?

The comparison of the two data collection methods first started with a subgraph comparison. The interview data presented a limited view on a communication network, in that it only represented a network of 11 possible nodes (people) and 45 possible, undirected edges (connections). This was both a factor of the question asked (to only identify ten people you work with most) and the limitation of people to identify successfully more than ten people in a given category. A comparison was conducted to identify if the nodes and edges represented in the interview data were present in the email data.

In every case, except for one instance in Case C, all of the nodes identified in the interview data were represented in the combined email data. Even this minor exclusion could be due to a mistake in the answering of the questionnaire. Potentially a nick-name was used or the person was incorrectly identified.

When turning to the subgraph comparison of the edges, most of the edges were the same in both methods. However, there were instances of edges identified in the interview that were not in

the email data and vice versa. Even with the differences in the edges in each method, though the lens of completeness, representing the most nodes and edges, the email data still represented a more complete view of a communication network, as with each case the email data represented more edges.

3. What effect does the data collection method have on the observation of predefined network characteristics?

The two methods were then compared in their entirety based on a number of predefined network characteristics. Due to the interview method relying heavily on the questions asked of the participants the number of nodes and edges that can be represented is greatly limited. In the case of an unbounded network, the volunteer is asked to generate names from memory. At most, a communication network based off of interview data from these cases can only represent at most eleven people and amongst those people 45 edges. With the low number of nodes and edges a high density and low modularity was observed. This could imply that all members identified for each volunteer were from the same group.

With the overarching aim to characterize an organizations communication network towards improved organizational behavior this method is limited in its depiction of communication networks. The success of this method may be in regards to newcomer socialization. In instances where new employees are being onboarded in small numbers, maybe one or two at a time, it would make sense to simply ask the person who previous held the newcomers role who they worked with most often.

Email data, in general represents a far greater number of nodes and edges. Email data has a high modularity and much lower density as it potentially represents data across groups. This is

a characteristic observed in large systems as a way of managing that complexity [24]. The size of the network from archival data is limited only by the bounds of the persons reach or at 1 million, the cap of the network processing software. Meaning this data source captures not only strong ties, but also weak ties, ties that are at best peripheral to the network. Weak ties are known to have an important role in networks and the ability to study these weak ties has been limited. Studying weak ties requires a lot of data

Difficulties of this method stem not from the data collection but from the processing. When pulling data from the Outlook server the names in the to/from, and cc fields are formatted as "Name, Rank, Organization." In multiple cases this formatting was an issue, especially in a military environment as rank changes and military members move organizations every three years. Organizations tend to change over time, condensing or splitting up to form new organizations. This posed a problem for archival data as the fields being pulled would change and that person would be represented twice in the network. This may not be a problem if the purpose of a study is to see the change in a person's network over time, using rank or organization to see how the networks differ in different positions/ranks could be informative. However, for this research of building a simple network of each volunteer, the formatting posed a challenge. Another challenge of this data is that is of a single mode of communication. Much of the data analysis relies heavily on the data that is collected from the archives. A challenge arises when people have "zero inboxes" or delete emails as they come in.

5.2 Significance of Research

This research provides a characterization of different data collection methods to support a social network analysis. Social network analysis is a method to study communication among members of an organization. This method is highly reliant on the data that is collected and then

analyzed. Usually data is collected through sociometric questioning, or interviews with the research participants or even simulated to prove an analysis technique. This data is multimodal and gives researchers data based on what the volunteer perceives their environment to be.

The other method studied is that of archival data from email data. This data characterizes a person's communication network using the to/from and cc fields archived in each email sent or received. The benefits to this method are that it is unbounded by human memory and therefore represents a larger number of people. Specifically, this method highlights the strength of weak ties in a volunteer's communication network.

With research heavily focused on analysis techniques little is often thought of the data that this analysis is done on. This research focuses on just that by characterizing two data collection methods for use in a social network analysis.

5.3 Recommendations for Action

With a way of characterizing and visualizing technical organizations through the use of email, one course of action would be to use this tool as a means of reducing the effects of turnover. Turnover, or newcomer socialization, was addressed in previous sections of this research. One way of mitigating the long-lasting effects of high rates of turnover is to help new employees realize their communication networks faster. Usually this is done by letting new employees socialize themselves and use trial and error to figure out who to go to, to get their work done.

By collecting data on exiting members of an organization the people they communicate with to get their job done can be passed along to the newcomer. This research has demonstrated that the email data can be sorted by degree which implies the level of connection that individual has to the network. It would be interesting to use this tool to help onboard a new employee to an

organization and to measure their rate of productivity with the communication network of the person who was previously in their position.

5.4 Recommendations for Future Research

The primary recommendation for future research would be to now use archival data collection, using the method characterized in this research, as the data source for a social network analysis. This research could address topics such as job satisfaction through perceived vs actual burden with email traffic, stress levels correlated to email traffic, or a comparison of people present in the sent box and the people present in the inbox.

VI. References

[1] H. Wilson, “Current Air Force Leaders,” p. 6.

[2] Smith, Adam, *An Inquiry into the Nature and Causes of the Wealth of Nations*, vol. 1. London: W. Strahan and T. Cadell, 1776.

[3] H. Sillitto *et al.*, “Systems Engineering and System Definitions,” p. 18, 2019.

[4] Adcock, Rick, “Introduction to System Fundamentals,” *SEBok Guide to Systems Engineering Body of Knowledge*. .

[5] R. M. Henderson, ; Kim, and B. Clark, “Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms,” 1990.

[6] J. Galaskiewicz and S. Wasserman, “Social Network Analysis: Concepts, Methodology, and Directions for the 1990s,” *Sociol. Methods Res.*, vol. 22, no. 1, pp. 3–22, 1993.

[7] “Analysis of Sociometric Data-Identification of Sociometric Groups.”

[8] C. Murphy and A. Bouffard, “Understanding Defense Acquisition Workforce Challenges,” pp. 1–19, 2017.

[9] Rogers, Everett, “Progress, Problems and Prospects for Network Research: Investigating Relationships in the Age of Electrongic Communication Technologies.” 1987.

[10] K. H. Zwijze-Koning and M. D. T. De Jong, “Auditing information structures in organizations: A review of data collection techniques for network analysis,” *Organ. Res. Methods*, vol. 8, no. 4, pp. 429–453, Oct. 2005, doi: 10.1177/1094428105280120.

[11] B. J. Bernardoni, R. F. Deckro, and M. J. Robbins, “Using Social Network Analysis to Inform Stabilization Efforts,” *Mil. Oper. Res.*, vol. 18, no. 4, pp. 37–60, Dec. 2013, doi: 10.5711/1082598318437.

[12] C. Hidalgo, D. Jagdish, and D. Smilkov, “Immersion Tool,” 2013.

[13] R. Rudis, “Visualizing the Clinton Email Network in R,” 2016.

[14] K. M. Eisenhardt, “Building Theories from Case Study Research,” p. 24.

[15] Anabel Quan-Haase and B. Wellman, “Hyperconnected Net Work,” p. 53.

[16] Marin, Alexandra and Wellman, Barry, “Social Network Analysis: An Introduction.” 2009.

[17] J. G. Casler, “Revisiting NASA as a High Reliability Organization,” *Public Organ. Rev.*, vol. 14, no. 2, pp. 229–244, Jun. 2014, doi: 10.1007/s11115-012-0216-5.

[18] “2017 Demographics Report.”

[19] C. Bond, J. Lewis, H. Leonard, J. Pollak, C. Guo, and B. Rostker, *Tour Lengths, Permanent Changes of Station, and Alternatives for Savings and Improved Stability*. RAND Corporation, 2016.

[20] Aristotle, *Metaphysics*, vol. 8, 14 vols. 350AD.

[21] C. Lüthje, C. Herstatt, and E. Von Hippel, “User-innovators and ‘local’ information: The case of mountain biking,” *Res. Policy*, 2005, doi: 10.1016/j.respol.2005.05.005.

[22] M. Fritsch and M. Kauffeld-Monz, “The impact of network structure on knowledge transfer: an application of social network analysis in the context of regional innovation networks,” *Ann. Reg. Sci.*, vol. 44, no. 1, pp. 21–38, Feb. 2010, doi: 10.1007/s00168-008-0245-8.

[23] “Jefferson’s Taper. A 200 Year Old Perspective on the Internet.” [Online]. Available: https://whatsthepont.blog/2013/02/17/jeffersons-taper-a-200-year-old-perspective-on-the-internet/.

[24] S. K. Ethiraj and D. Levinthal, “Modularity and Innovation in Complex Systems,” *SSRN*, 2003, doi: 10.2139/ssrn.459920.

[25] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, “Network analysis in the social sciences,” *Science*, vol. 323, no. 5916, pp. 892–895, 2009, doi: 10.1126/science.1165821.

[26] Morse, Gardiner, “The Science Behind Six Degrees,” *Febr. 2003*.

[27] Renfro, Robert and Deckro, Richard, “Social Network Analysis for the Iranian Government.” .

[28] T. N. Bauer and B. Erdogan, “Organizational socialization: The effective onboarding of new employees.,” in *APA handbook of industrial and organizational psychology, Vol 3: Maintaining, expanding, and contracting the organization.*, S. Zedeck, Ed. Washington: American Psychological Association, 2011, pp. 51–64.

[29] R. Korte, “The Socialization of Newcomers into Organizations: Integrating Learning and Social Exchange Process,” *Int. Res. Conf. Am. Acad. Hum. Resour. Dev.*, p. 8, 2007.

[30] D. Krackhardt, “Social Networks and the liability of Newness for Managers,” *Trends Organ. Behav.*, 1996.

[31] K. Rollag, S. Parise, and R. Cross, “Getting New Hires Up to Speed Quickly.”

[32] A. F. de Toni and F. Nonino, “The key roles in the informal organization: A network analysis perspective,” *Learn. Organ.*, vol. 17, no. 1, pp. 86–103, 2010, doi: 10.1108/09696471011008260.

[33] D. M.M. and F. K.A., “An Introduction to Social Network Analysis,” *New Dir. Eval.*, vol. 107, no. Fall 2005, pp. 5–13, 2005, doi: 10.1002/ev.157.

[34] F. U. Pappi and J. Scott, “Social Network Analysis: A Handbook.,” *Contemp. Sociol.*, vol. 22, no. 1, pp. 128–128, 2006, doi: 10.2307/2075047.

[35] N. M. Tichy, M. L. Tushman, C. Fombrun, and M. L. Tushman, “Social Network Analysis for Organizations,” vol. 4, no. 4, pp. 507–519, 2010.

[36] D. Kelley, J. H. Turner, and L. Beeghley, “The Emergence of Sociological Theory.,” *Contemp. Sociol.*, vol. 11, no. 4, pp. 466–466, 2006, doi: 10.2307/2068847.

[37] R. S. Weiss and E. Jacobson, “A Method for the Analysis of the Structure of Complex Organizations,” *Am. Sociol. Rev.*, vol. 20, no. 6, pp. 661–661, May 1955, doi: 10.2307/2088670.

[38] “Sociometry,” *International Encyclopedia of the Social Sciences*. pp. 390–392.

[39] C. Hollander, *An Introduction to Sociogram Construction*. Snow Lion Press, 1978.

[40] J. Nahapiet and M. Earl, “Creating Organizational Capital through Intellectual and Social Capital,” *Organ. Sci. Winter Conf. Keyst. CO*, vol. 23, no. 2, pp. 1–39, 2000, doi: 10.2307/259373.

[41] Hidalgo, Cesar, “What I Learned from Visualizing Hillary Clinton’s Emails,” Nov. 2016.

[42] J. West and S. Gallagher, *Patterns of Open Innovation in Open Source Software*. 2008.

[43] V. D. Blondel, J. Guillaume, and E. Lefebvre, “Fast unfolding of communities in large networks,” pp. 1–12.

[44] V. Hlebec and T. Kogovšek, “How (not) to Measure Social Support Networks: the Name Generator vs. the Role Relation Approach,” p. 17.

[45] J. Tyler, D. Wilkinson, and B. Huberman, “Email as spectroscopy: automated discovery of community structure within organization. Communities and technologies,” *Inf. Soc.*, vol. pages, no. 2, pp. 81–96, 2003.

[46] S. Høyrup, “Employee-driven innovation and workplace learning: basic concepts, approaches and themes,” *Transf. Eur. Rev. Labour Res.*, vol. 16, no. 2, pp. 143–154, Apr. 2010, doi: 10.1177/1024258910364102.

[47] B. L. G. Sébastien Heymann, “Visual Analysis of Complex Networks for Business Intelligence with Gephi edicte Le Grand To cite this version : Visual Analysis of Complex Networks for Business Intelligence with Gephi,” 2013.

[48] B. Bengfort and K. Xirogiannopoulos, “Visual Discovery of Communication Patterns in Email Networks,” pp. 1–9, 2015.

[49] R. M. Stogdill, “The Sociometry of Working Relationships in Formal Organizations,” *Sociometry*, vol. 12, no. 4, pp. 276–276, Apr. 1949, doi: 10.2307/2785595.

[50] M. Bastian, “Gephi Features,” *Features*. [Online]. Available: https://gephi.org/.

[51] Jacomy, Mathieu, Venturini, Tommaso, Heymann, Sebastien, and Bastian, Mathieu, “ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software.,” *PLOS One*, 2014.

[52] M. S. Cole, H. Bruch, and B. Vogel, “Energy at work: A measurement validation and linkage to unit effectiveness,” *J. Organ. Behav.*, 2012, doi: 10.1002/job.759.

[53] S. Raisch, J. Birkinshaw, G. Probst, and M. L. Tushman, “Organizational Ambidexterity: Balancing Exploitation and Exploration for Sustained Performance,” *Organ. Sci.*, 2009, doi: 10.1287/orsc.1090.0428.

VII. Appendix A

Sample Questionnaire Sent to Volunteers

Name of people identified	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

Steps:

1. Think back over the past month. Consider all members of the organization here that you have contacted during business hours. Which ones have you spent the most time with on a business basis? With whom do you spend the most time with getting work done?
    a. Name five assistants/subordinates.

    b. Name five superiors/associates at the same level" (10 people in total)

2. Rank all members in accordance to how much time you have spent communicating with them. Place the Ranked names in the numbered row/column above.
    a. 1 – most time
    b. 10 – least time
3. Identify with a mark in the associated box if you believe a person in your list may work with another person in your list. SEE NEXT PAGE FOR EXAMPLE

EXAMPLE

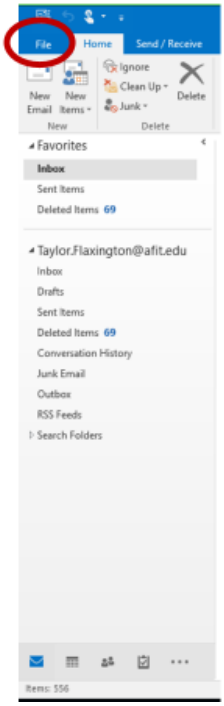
Name of people identified	1	2	3	4	5
1.					
2	1				
3	0	0			
4	0	1	0		
5	0	1	0	0	

1 – work with each other
0 – does not work with each other

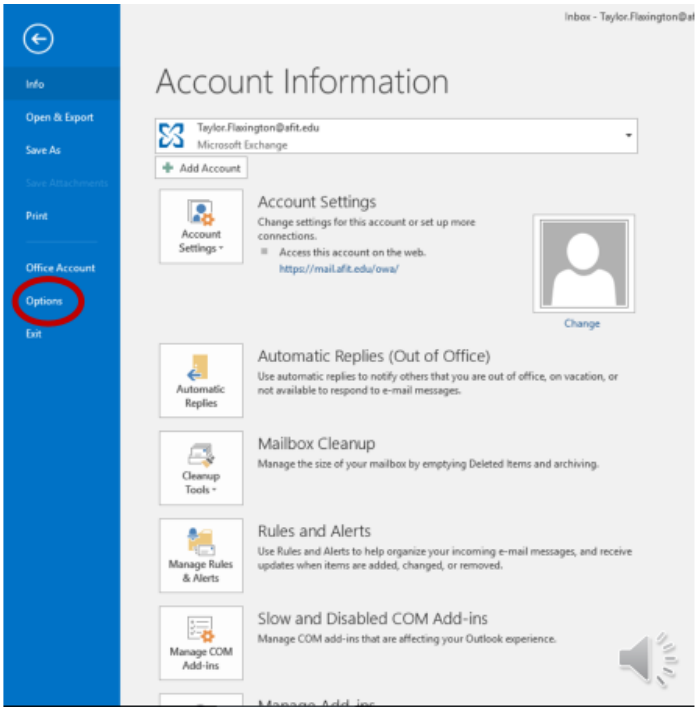
VIII. Appendix B

Email Data Collection Instructions

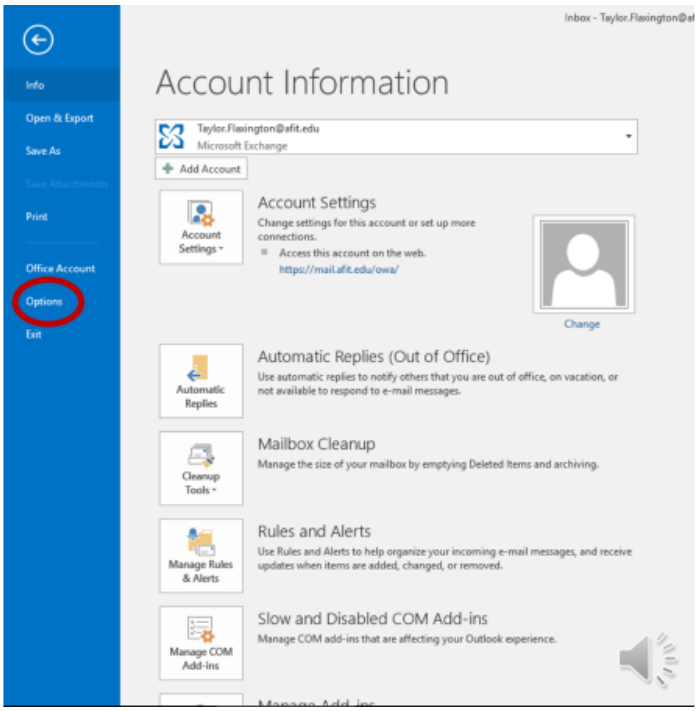
- Open outlook
- Select File



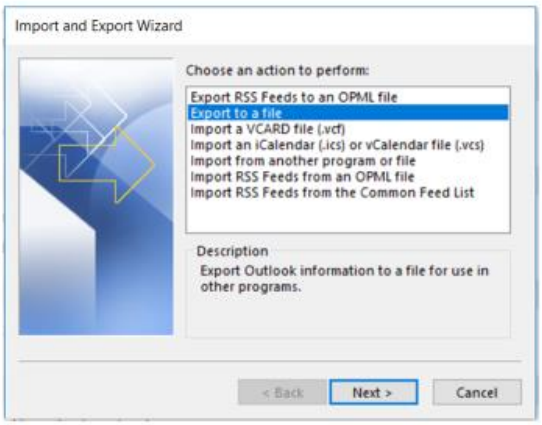
- Select Options
- A popup window will appear



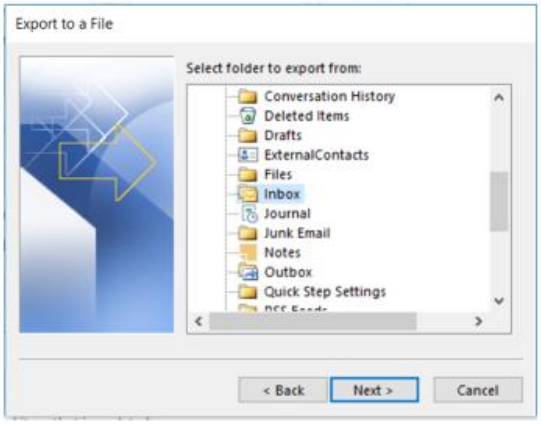
- Select Options
- A popup window will appear



- Highlight "Export to a file
- Hit next



- Highlight “Inbox” or “Sent”
- Hit next



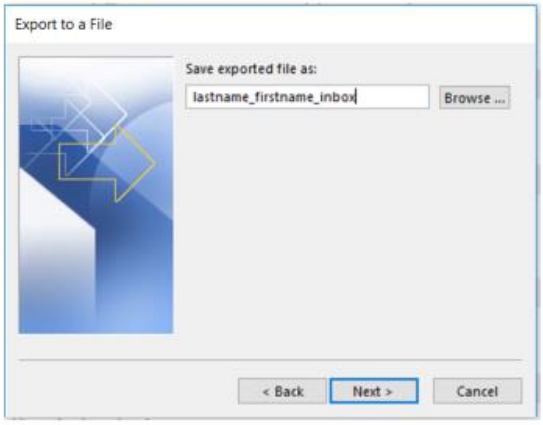
- Save exported file in desired location as:

“Lastname_firstname_inbox”

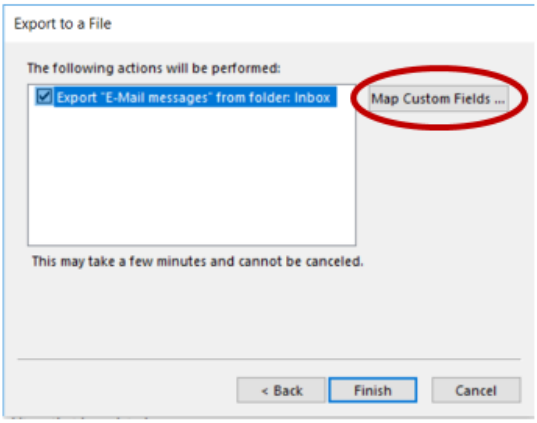
Or

“Lastname_firstname_sent”

- Hit next



- Highlight text in box
- Select “Map Custom Fields..”

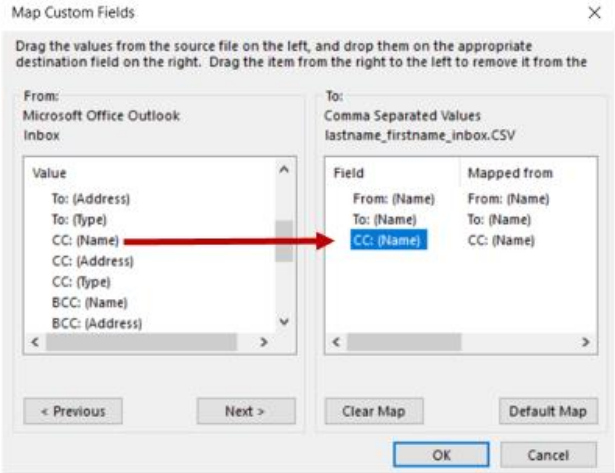


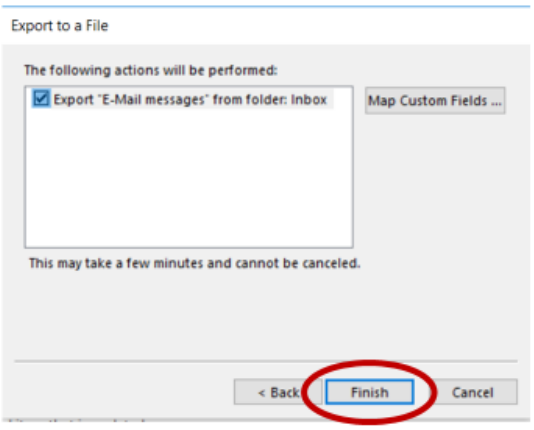
- Drag and drop

“From: Microsoft office Outlook Inbox” to “To: common Separated Values”

  - From: (Name)
  - To: (Name)
  - CC: (Name)

- Hit okay



- Hit “Finish”
- This will save your file in your specified location
  - Saving file may take a minute or two
- Send files via email

- **Repeat steps for “Sent” folder**