EXPLORING THE PRINCIPLE OF PROVENANCE WITH SOCIAL NETWORK ANALYSIS

by

Kathryn Suzanne Chandler

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARCHIVAL STUDIES

in

The Faculty of Graduate and Postdoctoral Studies
(Library, Archival and Information Studies)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2016

Abstract

Traditionally, an archival fonds is conceptualized as an aggregate of records which are mutually relevant. This mutual relevance is often attributed to the origin of member records in a common context – with this context typically understood as the context of an organization, and more specifically, a department.

It is considered difficult to identify mutually relevant records in modern organizations. This difficulty is often attributed to frequent administrative changes which disrupt departmental contexts. This thesis tests a technique that aims to use the information within the records to identify a context common to a set of records. It involves extracting the name of the creator and the name of the modifier from each record, then subjecting this information to a community detection algorithm. It was hypothesized that groups of individuals who frequently modify one another's records constitute a common context.

After applying various community detection algorithms to the records of an organization, the resulting groups of records were presented to the staff of the organization for feedback. Staff clearly indicated that groups of records produced by the community detection algorithms were not mutually relevant.

These results can be explained with reference to the works of Jenny Bunn, who argued that an autonomous community only comes into existence when constituent members engage in both "being" and "doing." During the interviews with staff, it was clear that some algorithms produced groups of people characterized by established relationships ("being") while others produced groups in pursuit of a joint activity ("doing"). The absence of overlap suggests there were no autonomous subcommunities in this study, and therefore, no common context by which records can be bound.

Mutually relevant records can also be formed by employees in their attempts to keep records orderly. To explore this further, it was argued that constructing a folder structure is akin to constructing a narrative, with the narrative components taking the form of records. When numerous employees attempt to organize the same records using different narratives, the aggregate may seem disorderly. This thesis suggests that disentangling these narratives is a method by which order may be restored.

Preface

I, Kathryn Chandler, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. The fieldwork reported in Chapter 4 was approved by the University of British Columbia Behavioural Research Ethics Board. The project was previously known as “A Solution to Disorderly Backlogs of Digital Files? Evaluating the Effectiveness of Social Network Analysis in File Reclassification Projects.” It was covered by Ethics Certificate number H15-01278.

Table of Contents

List of Figures

Acknowledgements

I am deeply indebted to my supervisor, Jennifer Douglas, for her steadying influence as she guided me through the new terrain of writing a thesis. She created a space of connection during our Wednesday meetings, while gently opening my eyes to other ways of thinking about the archives. As I wrote the first draft, I saw for the first time the subtle way she had pointed me forward and realized how lucky it is to have an supervisor who can create an experience of both support and freedom.

Luanne Freund immediately understood what I was trying to do in this project, and mobilized her connections and her resources to help me. Her support was the turning point of a project which had previously struggled to get off the ground. Luanne also played a crucial role in the planning stages, as her astute feedback streamlined the project in a way that made implementation possible. Her feedback on thesis drafts was invaluable.

Giovanni Michetti opened my eyes to the puzzles of this discipline when I was a student in his ARST 515 class, and his unwillingness to accept a superficial conceptualization of an archives infected me and sustained this project. He generously replied to the many emails I sent him over the past two years, unmasking flaws and possibilities each time.

The hardest part of this project was finding a host department. Given the risk that is involved in letting a stranger access work records, and the fact that most departments are so busy such that taking on one more task seems impossible, I am very grateful to the manager whose willingness to support the advancement of records management overrode these concerns. As the project progressed, I came to realize just how lucky I was: each person I interviewed, including the manager himself, was prompt in setting interview times, courteous during interviews, and thoughtful in their responses.

I would like to give a special thanks to Rick Kopak, who moved this from being a side project I worked on in my spare time, to a 12-credit thesis I could work on full-time. If it wasn't for him, this project would likely not exist.

For an extended period of time, I didn‘t know if it was possible to extract the metadata that was necessary for this project. Luckily, a highly-knowledgeable IT professional who I had met once during a Co-op position replied to my email, detailing several possibilities for extracting the metadata, and connecting me with a SharePoint specialist. Many thanks to Matthew Ellis for the crucial role he played in making the project possible.

Michael LeBlanc, Sherryll Catindig, and Waii Chow wrote the script to extract the metadata. When they sent me a spreadsheet with the relevant metadata as proof that the script worked, I spent the rest of the day on cloud nine. If this wasn't enough, they also covered the cost of writing the script, put a call out to a community of SharePoint users to help me find a host, and helped me gain permission to access sites shared across departments. Encountering their professionalism, and kindness was a highlight of this project.

Joy Rowe believed in this project during a time that I was discouraged, and showed her support by reaching out to the owners of various SharePoint sites to ask if they would host the study. I remember feeling incredulous that someone would go so far out of their way to help me.

In the process of reaching out for a host for this project, I was buoyed by the positive response of people who work in the field of archives and records. Many thanks to Susan Hart, Anne Rathbone, Alan Doyle, Barbara Towell, Dan Gillean, and Sarah Romkey.

One of the benefits of having a disability is being witness to multiple acts of care. I wish to thank every instructor who wore an FM microphone so that I could hear them, with special thanks to the five instructors at the end of my degree who facilitated the passing of the FM from student to student during class discussion. Several students reminded professors to do things on my behalf when I wasn't in a particularly assertive place, and showed a kind of profound inclusion by changing their habits in the classroom. I am grateful to them for shifting the way that I saw the world.

I have no way to adequately thank Lois Evans, who taught me ARST 575J. I hadn't communicated to her that I couldn't hear the other students, or that I was frustrated with the effort of trying. Her perceptiveness and inclusion came at a time when I needed it most.

I would like to thank my two brothers. My younger brother's legendary patience as he walked me through a complicated technological procedure was a tremendous gift in the early stages of this project. My older brother was the anchor of my graduate student life, bringing me back into balance every Saturday with his generosity of spirit and his careful listening.

I lucked out in this life, as I am the daughter of two parents who give so much love to their children. From a young age I witnessed their love of learning, and their sense of responsibility to others. Over the past two years, they helped me think through the planning of this project, gave me advice on navigating bureaucracies, edited my interview questions and the final draft, helped me make the visualizations more clear, learned archival theory so I could discuss my arguments, and propped me up during the setbacks. For their positive influence, and for enduring with me through the hard times, I am very grateful.

To my grandmother, Annette, for helping me when I was in middle school.

1 Introduction

1.1 Presentation of the Topic

An archives is traditionally defined as the aggregate of written documents belonging to a single organization, which may be further subdivided into smaller groups by the members of the organization.[1] Additionally, an archives is traditionally understood to constitute evidence of the past, as it was created as part of the organization's practical activity.[2] Thus, there are two major features of an archives according to traditional theory: it functions as evidence, and it involves boundary lines which demarcate groups of documents.[3]

This thesis was motivated by a deep sense of curiosity with regards to these fundamental features of an archives. It seemed to me that if evidence and boundary lines were fundamental, then these two features of an archives should be connected in some way. I am influenced in this belief by the work of Terry Eastwood, who notes that archival documents are "interdependent for their meaning and in their capacity to serve as evidence"[4] – a claim that explicitly ties the notion of bounded record groups to the notion of evidence. To make this abstract claim more concrete, Eastwood suggests that documents are bound together when they perform the same function. He makes clear, however, that this is an exploratory

[1] Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 52.
[2] Jenkinson, *Manual of Archive Administration,* 12.
[3] A note on terminology: I use the terms *record, file,* and *document* interchangeably, but what I mean is a *record* as it is defined in the InterPARES dictionary: "A document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference." *The InterPARES 3 Project Terminology Database*, s.v. "record," accessed March 27, 2015, http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41.
[4] Eastwood, "What is Archival Theory," 128.

hypothesis, noting that “we are far from understanding what we mean by function in archival science and how function governs creation of records.”[5]

Like Eastwood, I found the concept of function confusing and decided that it would be important to start from the basics. More specifically, I started from the articulation of the two kinds of boundary lines which demarcate groups of records. In the first type of boundary line, archivists group together the documentary output of a single organization. In doing so, they follow the principle *respect des fonds*; the record aggregate which results is called an *archives*, a *fonds,* or an *archival fonds*. Likewise, an archivist who retains the organization’s particular grouping of documents within a fonds follows *respect for original order*; the resulting documentary aggregates are called *series.*[6] [7]

In his article “The Last Dance of the Phoenix,” Peter Horsman makes clear that both *respect des fonds* and *respect for original order* are constituents of an overarching concept known as the *principle of provenance*.[8] Further, he shows that both *respect des fonds* and *respect for original order* have a long history of contesting definitions. Prussian archivist Adolf Brenneke, for example, argued that *respect des fonds* refers to the documentary aggregate of a community,[9] a contrast to Samuel Muller’s claim that it refers to the records of a single organization.[10]

[5] Ibid., 129.

[6] Wiersum, as cited in Horsman, “Last Dance of the Phoenix,” 10.

[7] A more complete definition for a series is as follows: “dossiers, file units or individual documents that are arranged in accordance with a classification or filing system or that are maintained as a unit because they result from the same accumulation or filing process, the same function or the same activity, and that have a particular form or because of some other relationship arising out of their creation, receipt or use.” *The InterPARES 3 Project Terminology Database*, s.v. “record series,” accessed April 12, 2015, http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. See also Cook, *The Management of Information from Archives,* 110.

[8] Horsman, “Last Dance of the Phoenix,” 3.

[9] Ibid., 15.

[10] Ibid., 9.

Horsman suggests that the principle of provenance is contested precisely because there is no definitive way to group records.[11] Ultimately, he makes the case that the principle of provenance is not rooted in a theory, and that its long endurance in the archival discipline should be attributed to its usefulness in keeping the holdings of an archival institution well organized.[12]

To better understand the claims presented by both Horsman and Eastwood, I carefully studied previous descriptions of the principle of provenance (chapter two), discovering that provenance refers to mutually relevant record aggregates. To explore this idea further, I carried out a research experiment which used social network analysis to identify the mutually relevant records within an aggregate produced by a department (chapter three and four). Social network analysis is a strategy for investigating social structures characterized by a focus on the relationships between individuals, rather than the individuals themselves.[13] In this study, the individuals were members of a department, and the relationship under study was record modification.[14] I explored the viability of these generated series by presenting them to the members of the department, and soliciting feedback.[15] In analyzing the interview transcripts

[11] Ibid., 17-18.

[12] Ibid., 22.

[13] Otte and Rousseau, "Social Network Analysis," 441-442.

[14] I focused on document modification as I cannot see how other relationships could be used to identify mutually relevant records. Document modification also has the benefit of being rooted in the practical activity of the organization. In the archival discipline, records formed during practical activity are considered authentic, as it is believed that records bear the trace of this activity and nothing more. By extension, it could be said that modifying documents is a practical activity, and hence the information is authentic. See Jenkinson, *Manual of Archive Administration,* 12.

[15] The astute reader will notice that this study could be one way to instantiate Greg Bak's vision of applying a web 2.0 philosophy to the archives. As far as I can understand, a web 2.0 philosophy involves making use of item-level metadata reflecting user interaction. The purpose of identifying meaningful patterns is to use it as a means of classification. The tool under examination in this thesis involves similar elements: it captures item-level metadata reflecting the record interactions of users, and it does so with the intent of exploring a new method of classification. See Bak, "Continuous Classification," 309-310.

and relating the findings back to the literature (chapter five), a nuanced justification for the principle of provenance emerged.

1.2 Why this Study?

This study investigates provenance and evidence, arguably the fundamental features of the archives. Understanding the nature of the archives is fundamental to any discussion of the archives. When archivists discuss appraisal – that is, the planned destruction of documents – their strategies should be informed by the nature of the thing that is being destroyed.[16] Similarly, when archivists discuss the process of arranging and describing documents, their decision-making should reflect the nature of the thing which is being arranged and described. While acknowledging that the nature of the archives is somewhat elusive as an abstract concept, it seems to me there are less complete and more complete articulations of this nature. In short, this thesis is important because it sheds light on a concept which is fundamental to the archival discipline.

This thesis is also important because it explores a new tool capable of quickly sorting an aggregate of documents into series. As many organizations have accumulated disorderly aggregates of records, there is a clear need for such a tool. The Provincial Government of British Columbia, for example, has 12,000 record aggregates stored on SharePoint sites[17] which

[16] Luciana Duranti's "The Archival Bond" makes a case for an appraisal strategy which respects the nature of the archives as an interconnected body of records. See Duranti, "Archival Bond," 217.

[17] Microsoft. "What is SharePoint?" Accessed August 14, 2015. https://support.office.com/en-us/article/What-is-SharePoint-97b915e6-651b-43b2-827d-fb25777f446f. To clarify, SharePoint is an online platform for storing, sharing and organizing records. SharePoint can be distinguished from other record-sharing platforms by its use of web technology, which enables employees to access the site from any device, provided they have the requisite credentials.

are likely disorganized because they are not integrated into the records management program.[18] Such disorderly aggregates make it difficult for employees to retrieve records when needed, resulting in frustration and wasted time. For governments, the significance of a disorderly aggregate may extend beyond the walls of the institution by presenting barriers to citizens who wish to access records in accordance with *Freedom of Information* legislation. By investing in the development of a tool which automatically sorts records into meaningful groups, this thesis represents an attempt to achieve greater efficiency and democracy in institutions that rely on records.

1.3 Research Questions

1. Are groups of records formed by social network analysis mutually relevant?
2. Do communities formed by social network analysis represent the working groups in an organization?
3. Assuming that employees will explain the social network visualization using narrative, to what extent do these individual narratives converge with one another?
4. What does the application of social network analysis to records tell us about the principle of provenance?

1.4 A Note on the First Person Voice

This thesis is strongly influenced by the dissertations written by Jenny Bunn[19] and Jennifer Douglas,[20] who described their research process using the first person voice. In my view, writing in the first person is the most accurate way to describe a research project. After all, research is essentially an attempt to make sense of something, which is to say, it is a personal

[18] Gillean, “The Consequences of Ignoring Records Management,” 12.
[19] Bunn, “Multiple Narratives, Multiple Views: Observing Archival Description.”
[20] Douglas, “Archiving Authors: Rethinking the Analysis and Representation of Personal Archives.”

process. Additionally, it could be argued that works written in the personal voice are easier to understand because the researcher's perspective contextualizes the meaning of a sentence, reducing ambiguity. However, in some cases, excessive contextualization is unnecessary and distracting. Therefore, this thesis is written in a mix of the first and third person voice, with an emphasis on the first.

1.5 Soft Theory

My decision to rely primarily on the first person voice is also reflective of the philosophical worldview that informs my approach to research. This approach is best expressed by Wolfgang Iser's concept of 'soft theory,' which can be understood by bringing it into comparison with its counterpart, hard theory. Hard theorists, according to Iser, are in search of a "keystone idea."[21] This keystone idea functions to explain various phenomena, with an example being Newton's three laws of motion. By contrast, soft theorists welcome multiple ways of seeing the object of study.[22] As a soft theorist, I believe that the concept of the archives benefits from multiple viewpoints. By using the first person, I make clear that I view the archives through a lens, and that my lens is one of many.

My approach to research is also influenced by the Greek philosophers Plato and Socrates, who used metaphor as a means of exploring abstract concepts. To discern the nature of justice, for example, Plato and Socrates posit that the just city is a metaphor for the just person. This metaphor enabled them to eliminate from their discussion the superfluous elements of justice

[21] Iser, *How to do Theory,* 5.
[22] Ibid.

specific to a city or a person. Theoretically, what remains is justice in the abstract.[23] For me, the elusive abstract concept under study was not justice but an archives. In the next section, I describe my search for a metaphor which would best explain the abstract features of an archives as set out by traditional archival theory.

1.6 Applying Metaphors to an Archives

In my earliest attempts to make sense of the archives, I proposed that the archives could be understood as an abstract system of logic. My idea for this came from the concept of the archival bond, which is a theoretical relationship existing between a pair of records.[24] A network of archival bonds exists between records belonging to the same series, and between records belonging to the same fonds[25] – making clear it is an alternative method of conceptualizing the boundaries implied by provenance. Unlike a boundary line, however, the archival bond enables one to hypothesize the nature of the relationships within the archives. Luciana Duranti, for example, asserts that the archival bond is based on *cause-effect*.[26] I understand this to mean that if there are two events, each described by separate records, and the first event causes the second event, it could be said that the first and second records exist in a *cause-effect* relationship.

[23] "We thought that, if we first tried to observe justice in some larger thing that possessed it, this would make it easier to observe in a single individual. We agreed that this larger thing is a city, and so we established the best city we could, knowing well that justice would be in one that was good. So, let's apply what has come to light in the city to an individual, and if is accepted there, all will be well. But if something different is found in the individual, then we must go back and test that on the city. And if we do this, and compare them side by side, we might well make justice light up as if we were rubbing fire-sticks together. And, when it has come to light, we can get a secure grip on it for ourselves." Plato, *Republic,* 110.
[24] Duranti, "Archival Bond," 216.
[25] Ibid., 215-216.
[26] Ibid., 217.

This cause-effect relationship reminded me of a logical operator known as the conditional. The conditional is represented as an arrow (→), and it is placed between two statements to signify that if the first statement is true, then the second statement is also true.[27] I imagined that if records were generalized into statements, the conditional could represent the *cause-effect* connection between records. The two initial "statements" result in a new statement about both records which is also considered true – reminiscent of the way archival documents can be used to construct complex storylines. After an extended period of exploration, I realized I had assumed a record could be represented as a statement, but this was an erroneous assumption: records are very complex. In other words, if the purpose is to reconstruct the narrative represented by the record aggregate, logical operators are too simplistic a means of connecting these containers of information.

This claim runs somewhat counter to the diplomatics approach to records. Diplomatics is a discipline that identifies how the social, legal, and organizational context in which a document was produced influences its form.[28] One element identified in a diplomatic analysis is the document's "act" – with an example act being a "request for information."[29] This act is arguably a means to generalize the record into a simple statement. But if one looks closely at the definition for an act, it is not a straightforward concept. An act must "modify a situation"[30] – reflecting the origin of the concept in legally-binding documents. Duranti acknowledges this complexity when she notes that many contemporary documents may not have direct legal

[27] Luckhardt and Bechtel, *How to do Things with Logic,* 29-30.
[28] Duranti, *Diplomatics,* 41
[29] Ibid., 156.
[30] *The InterPARES 2 Project Dictionary*, s.v. "act," accessed March 21, 2016, http://interpares.org/ip2/display_file.cfm?doc=ip2_dictionary.pdf

implications, but suggests it is possible to discern an analogy to these legally-binding acts.[31] In my view, this need for creativity in discerning the act is an indication that a single act may not exist within a document.[32]

Given the complexity of an archival document, I realized I needed a metaphor which could represent the complexity of record content in a larger pattern. In July 2015, I came across the work of Gerald Edelman and Giulio Tononi, who argued that consciousness results when information is integrated, as it is in the brain of a human being.[33] They point out that sensory information - such as colour, spatial position, and depth - travel distinct neural pathways in our brains, and it is only at a later point in the journey that they fire closely together, giving us the perception of a coherent world.[34] What made this seem a viable metaphor for the archives – apart from the obvious fact that organizations also integrate information – was the notion that these sensory-specific pathways aggregate into functional clusters, each specializing in some kind of sensory information.[35] Likewise, the archives are often understood in terms of the functions of the organization. This approach seemed to avoid the myopia of my logic-based approach to an archives, as each neuron could be accounted for in a broader macroscopic structure.

[31] Duranti, *Diplomatics,* 68.

[32] It should be noted too, that Duranti's position may have changed since she wrote "The Archival Bond" in 1997. In the InterPARES online dictionary, several definitions are offered for the archival bond, but none that refer to its *cause-effect* nature, with emphasis instead on its networked nature. As a network evokes an entirely different structure than a linear sequence representing a series of causal events, my discussion may well be outdated. *The InterPARES 2 Project Dictionary*, s.v. "archival bond," modified March 21, 2016, http://interpares.org/ip2/display_file.cfm?doc=ip2_dictionary.pdf

[33] Edelman and Tononi, *Consciousness,* 54.

[34] Ibid., 116.

[35] In 1980, Anne Treisman was the first to explore how different visual features are integrated in the brain. See Ward, Grinstein, and Kelm, *Interactive Data Visualization,* 94.

According to Edelman and Tononi, neuronal clusters within the brain are not physically distinct, but are rather made distinct by the intensity of interaction among their constituent neurons. To identify these clusters, Tononi strategized a method for drawing boundaries around an intensely-interacting mass of neurons, hypothesizing that the ideal boundary would have few neuronal exchanges crossing the boundary, and large numbers of neuronal exchanges taking place within the boundary.[36] As it turns out, this articulation of a neuronal cluster is very similar to the articulation of a community within a large social network. That is, a community detection algorithm is able to identify groups of people who interact frequently amongst themselves, while interacting little with others in the network.[37]

As a record may be created by one person and modified by another, I realized that records capture the equivalent of a neuronal exchange. By extracting the name of the creator and the name of the modifier from each record, then subjecting this information to a community detection algorithm, it is possible to identify groups of people who frequently interact with one another. In my initial reading of the archival studies literature, it seemed clear that these informal working groups could constitute the provenance of the records.

I tested this idea by finding an organization that allowed me to extract the creator and modifier metadata from the records in their SharePoint site. I ran the community detection algorithm on this metadata, then showed the results to those who were familiar with the organization – the staff who produced the records. It was crucial that the network information was visualized to enable staff to easily understand the results and provide feedback. A network visualization,

[36] Edelman and Tononi, *Consciousness,* 120.
[37] Girvan and Newman, "Community Structure in Social and Biological Networks," 7821.

also known as a graph,[38] is generally comprised of two components: there are dots, known as *nodes* or *vertices*, and lines, which are known as *edges*.[39] In this project, nodes represent people, and edges indicate that one person modified the files of another. The network visualizations in this project additionally display coloured shapes encircling nodes, which are the communities identified by the algorithm. An example social network analysis visualization is shown in *figure 1—1*.

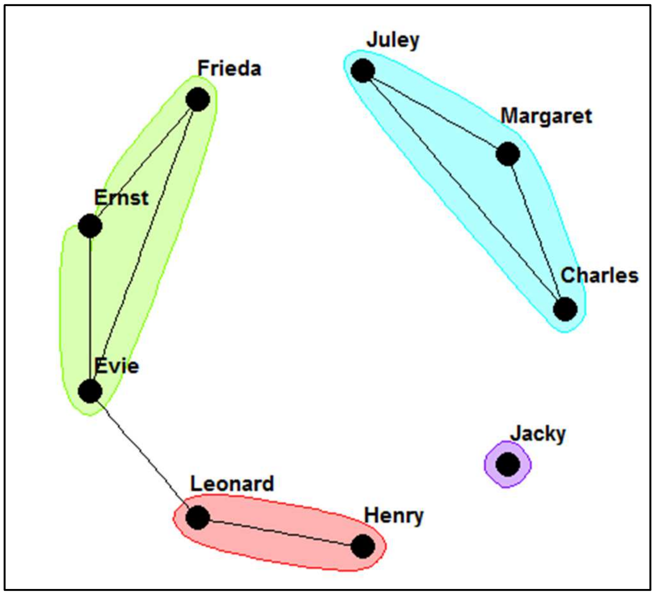


Figure 1—1 Example of Social Network Visualization

In order to understand staff response to these visualizations, I needed a nuanced understanding of the justifications for record aggregates. Therefore, I conducted a literature review on the principle of provenance. This enabled me to identify the fundamental features of

[38] According to David Easley and Jon Kleinberg, a graph "is a way of specifying relationships among a collection of items. A graph consists of a set of objects, called *nodes*, with certain pairs of these objects connected by links called *edges*." Easely and Kleinberg, *Network, Crowds and Markets,* 21.
[39] Newman, *Networks,* 1.

provenance. I applied these insights to the works of archival theorists developing a socially constructed view of the archives.

2 Literature Review

2.1 Introduction

This literature review describes various rationales that have been put forward to justify the principle of provenance. As these rationales do not fully account for the capacity of the archives to function as evidence, I make clear there is a gap that can be filled by this study.

This chapter is divided into six sections. In section 2.2 I explore the notions of provenance put forward by Muller, Feith, and Fruin; Peter Scott; and Duranti. In section 2.3, I look at how contemporary archivists have conceptualized the archives as a social construct, and the implications of this new thinking on the principle of provenance. I then explore the constructivist perspective in greater depth, showing how a socially-constructed archives is able to function as evidence (section 2.4), even while being shaped by the archivist (section 2.5). Section 2.6 looks at studies involving social network analysis and records.

2.2 How Provenance Works

As noted in section 1.1, provenance is a two-part principle comprising both *respect des fonds* and respect for original order. One of the earliest articulations of this principle can be found in Muller, Feith, and Fruin's 1898 *Manual for Arrangement and Description*. In section 16 of the *Manual*, the authors assert that the records from a single organization will naturally subdivide by department, and that respecting these natural subdivisions is the best method of organization for archival material.[40] This description of respect for original order implies *respect des fonds,* hence the statement as a whole is understood as an argument for the principle of provenance. The rationale for respecting this order is that doing so enables one to obtain a holistic understanding of the structure of the organization, much in the way a paleontologist obtains a holistic picture of a biological organism by viewing its fossilized remains.[41]

This holistic perspective of the organization's administrative structure is important because it makes it less likely those viewing the records will misinterpret them. Hilary Jenkinson, a British archivist who read the Dutch *Manual* and disseminated its insights to the English-speaking world in 1922 with a manual of his own, makes this point with the use of an example from his own holdings.[42] He notes that Receipt Rolls of the Exchequer are often interpreted by students as representing the total moneys paid to the British government. Had the students a better understanding of the administrative structure, they would have been more aware that the Rolls were used at a lower administrative level, and thus did not function as a general ledger.[43]

In putting forward the metaphor of the archives as an organism, Muller, Feith, and Fruin acknowledged that like an organism, the archives "changed its state again and again"[44] over the course of its life. They believe that these changes stem from secretaries who act with disregard for the most appropriate arrangement of the records, which is, by administrative structure. [45]

[40] Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 55-56.
[41] Ibid., 71.
[42] Jenkinson, *Manual of Archive Administration,* 12.
[43] Ibid.
[44] Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 71.
[45] Ibid.

They further believed that such arrangements could be easily corrected by bringing together the modern papers with their relevant historical aggregate.[46] Fifty years later, the assumption that documents clearly correspond to a single department was challenged by modern bureaucracies, where departments are subject to drastic merges, splits, and eliminations.[47] These fluctuations make it difficult for a static archives to perform as a representation of the dynamic organization.

The recognition that series were being dismembered in the wake of government restructuring is largely attributed to Australian government archivist Peter Scott.[48] More specifically, he recognized that responsibilities were transferred from one department to the next such that documents relating to any given responsibility were faced with three fates: one, they were kept with the new department, such that the name of their old department – important contextual information – was obscured; two, the documents were split between the departments, meaning they no longer served as a mutual context for one another; three, the archivist created a group based on function which encompassed the entire series.[49] Scott did not think that these "nebulous and fictitious"[50] functions were an appropriate substitute for the names of the departments, as they did not relate the records back to their administrative reality.[51] To solve this problem, Scott proposed a recordkeeping system that kept the series intact while listing the various departments to which it belonged.[52]

[46] Ibid.
[47] Horsman, 12
[48] Scott, as cited in Hurley "Parallel Provenance (2)," 59.
[49] Scott, "Record Group Concept," 81.
[50] Ibid.
[51] Ibid.
[52] Ibid., 83.

Scott's emphasis on the interconnections between the records was predated by Italian archivist Giorgio Cencetti, writing in 1939. To make clear the importance of the connection between records, Cencetti developed a theoretical concept known as the archival bond.[53] As the archival bond is the network of relations connecting an aggregate of records,[54] the bond is a tool which has been used to explore the fundamental question of why some records are interconnected, while other are not. Duranti, for example, argued that the archival bond is made manifest when groups of records are distinct,[55] with distinct groups forming as members of the organization attempt to keep records organized.[56] In some cases, these groups are made explicit with a classification code or a registration number.[57]

As a recordkeeping system generally brings together records which are mutually relevant, it could be said that aggregates bound by the archival bond are comprised of mutually relevant records. More specifically, it could be said that Duranti makes two claims: one, records in distinct groups manifest the archival bond; and two, groups formed by the archival bond are mutually relevant. Problematically, these two claims suggest that the distinct groups of records one finds in a workplace are always mutually relevant. This is unlikely as employees may be

[53] Cencetti, as cited in Duranti, "Archival Bond," 216. It should be noted that Muller, Feith, and Fruin's "natural relation" – arguably another term for the archival bond – predates Cencetti's work. See Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 50.
[54] Ibid.
[55] Ibid.
[56] The perceptive reader will notice that the two traditional notions of provenance that I touch on in this paper have been neatly dichotomized by Horsman. Horsman calls Muller, Feith, and Fruin's notion of an original order that corresponds to departments a *conceptual order*. By contrast, an order reflecting the natural accumulation of records, as described by Duranti, is termed *physical order*. See Horsman, "Last Dance of the Phoenix," 9.
[57] Ibid., 216.

rushed when they place a record in a shared space, or they may be simply unaware that a more relevant folder for their record had been created by a colleague.

For this reason, it is helpful that Duranti puts forward another attribute of the archival bond, this being that it is "determined."[58] In this context, the word *determined* means that records in the same aggregate participate in the same activity, or function.[59] This claim makes sense: records participating in the same activity likely refer to the same events, concepts, or procedures – hence, they are mutually relevant. Problematically, the concept of a function is widely recognized as ambiguous by archival scholars. Eastwood notes that archivists are "far from understanding what we mean by function in archival science, and how function governs creation of records."[60] Fiorella Foscarini's research revealed that we lack methods for determining the scope of a function, and we additionally lack methods for building a comprehensive set of functions representing an organization.[61] The ambiguous nature of the function is perhaps best expressed by Peter Scott, who refers to functions as "nebulous and fictitious."[62] In the absence of a clear definition, it is difficult for theorists to build on the notion of function.

The notion of the archival bond makes clear that two criteria must be satisfied if an aggregate is to count as an instance of the principle of provenance. One, records must be mutually relevant and two, this mutual relevance stems from the reality that gave rise to the records. Groups of records identified by social network analysis seem to offer a means of satisfying these two

[58] Ibid.
[59] Ibid.
[60] Eastwood, "What is Archival Theory," 128.
[61] Foscarini, "Understanding Functions," 21.
[62] Scott, "Record Group Concept," 81.

criteria. That is, records brought together by closely-interacting individuals have a shared origin, and this shared origin, presumably, makes them mutually relevant. However, the question may still arise as to why these two criteria are relevant to the principle of provenance. In the next section, I explore an answer to this question by studying the archives through the lens of social constructivism.

2.3 'Perspectives about Records'

In this section, I make a case that the fundamental nature of the archives is one comprised of perspectives. To do this, I draw on the work of Heather MacNeil, who in turn draws on the work of Joseph Grigely. In *Textualterity,* Grigely notes that the discipline of textual criticism explores the different texts of a single work.[63] Traditionally, textual critics have focused on a methodology for discerning the origins that gave rise to these differences. Grigely proposes a new direction for the field which involves developing a theory for how a text transmits over time. A crucial first step, in his view, is to acknowledge that texts are social constructs. That is, a text is not a physical object, nor is it the author's intended meaning; instead, it is the space between an author and a reader.[64] Grigely illustrates his concept by relating an experience articulated by the poet Rilke, who locates himself in the middle of a gallery hung with Cezanne's paintings, and comments that he feels the paintings "drawing together in a colossal reality."[65] As this experience only transpires when Rilke situates himself in the centre of the room, it is

[63] Grigely, *Textualterity*, 8.
[64] Ibid., 4.
[65] Ibid., 121.

suggested that the experience is not located in the paintings. Instead, the experience is an interplay between Rilke's interpretation and Cezanne's expression.

MacNeil argues that the archives is similarly an interplay between author and reader, and uses the Bakunin Family fonds to illustrate her point.[66] This fonds was possessed by a series of custodians, including a scholar named Kornilov, who annotated the letters in preparation for writing a history of Russia.[67] According to MacNeil, these annotations are evidence of an interplay between Kornilov, the reader, and the Bakunins, the authors. In her view, this particular social construct is worth studying; it tells us, for example, about the political climate in which Kornilov operated.[68] For organizational records, I argue that a social construct emerges within the time frame that the records are being created due to the interaction of employees, who are simultaneously the readers and the authors.[69] To further explore the notion of organizational records as a social construct, I needed a study that moved beyond the notion of a single author-reader relationship, and considered the implications of multiple authors interacting with multiple readers.

The work of oral historian Alessandro Portelli sheds light on the way a community gives rise to a social construct.[70] Portelli interviewed the inhabitants of a town called Terni, asking them to

[66] MacNeil, *Archivalterity*, 9.

[67] Ibid., 16.

[68] Ibid., 24.

[69] Without a doubt, my early exposure to the appraisal strategy described by Terry Cook in 'Mind over Matter,' likely sensitized me to the way records express relationships existing in the same time frame. Cook argues that in conducting appraisal, it is the records which reference the relationship between the citizen, the programme and the agency which should be saved, as this captures "the essential dialectic" and thus the sharpest image of the originating circumstances. See Cook, "Mind over Matter," 57.

[70] I was exposed to the works of Portelli during a student presentation in ARST 517 - *History of Recordkeeping.* Many thanks to Jarin Schexider for her thoughtful presentation.

reconstruct a series of clashes between workers and police which had taken place 30 years previous. Their accounts often referenced the death of a factory worker called Luigi Trastulli, who had been shot by the police. Portelli notes that interviewees often placed the death of Trastulli in 1953,[71] but in fact, several written sources and some oral sources confirm Trastulli's death as taking place in 1949.[72] This suggests that the information represented by a social construct does not represent a hard and fast truth, but is rather a contested truth representing different perspectives.

Arguably, archivists have always been concerned with perspectives in records. Before Muller strong-armed the archival world into believing that archival arrangement should mirror the organization's departments and subfunctions,[73] his contemporary Theodoor Van Riemsdijk described an alternative approach to arrangement which emphasized an open-ended study of the organization. According to Eric Ketelaar, Van Riemsdijk specifically sought to understand "how and why records were created and used by their users"[74] – a statement which indicates a concern with perspectives.

Conceptualizing the archives as the social construct of a community is useful because it suggests a community is distinct from its environment. The idea of a distinct community is important in archival studies, as archivists have traditionally used the distinctness of a community to justify that *respect des fonds* is indeed an act of representation. Jenny Bunn explored several archival definitions for an autonomous community, including Eastwood's

[71] Portelli, *Death of Luigi Trastulli,* 15.
[72] Ibid., 2.
[73] For details on Muller's strong-arming tactics see Ketelaar*, Archival Theory and the Dutch Manual,* 60.
[74] Ketelaar, "Archival Theory and the Dutch Manual," 58.

assertion that a department is said to be a distinct entity once its competence or mandate has been decreed by its parent body.[75] Bunn argued that this explanation sidesteps theoretical issues in the notion of autonomy. As a result, it does not explain all instances of autonomy in the field of archival studies. For example, it does not explain how it is that parent bodies themselves obtain autonomy.[76]

In searching for a better way to define the community that effects the existence of a fonds, Bunn builds on the work of Humberto Maturana and Francisco Varela and proposes "autopoiesis," the notion that the system's "being and doing" makes the system distinct from its environment.[77] This concept was initially puzzling to me. What does it mean to say that a community engages in a collective act of being? Likewise, what does it mean when a community engages in a collective act of doing? Despite my confusion, I had enough clarity to recognize Bunn's "being and doing" in the results of the study, discussed in section 5.1.

2.4 Evidence in the Archives

As noted in section 1.1, the archives is traditionally conceived as an entity comprised of two major features: boundary lines, and the capacity to perform as evidence. This section explores how boundary lines implied by the principle of provenance make it possible for the archives to perform as evidence. According to the *Oxford English Dictionary*, evidence is "information indicating whether a belief is true or valid."[78] This definition makes sense in the context of

[75] Eastwood, as cited in Bunn, "Questioning Autonomy," 6.
[76] Ibid.
[77] Ibid., 10.
[78] *Oxford English Dictionary*, 10th ed., s.v. "evidence."

archival studies: the archives is a site of information, and this information is used by scholars to refute or deny their claims about the past. Archives which have not been accessed by scholars can easily be conceptualized as evidence because they have the potential to act as evidence. In her article “The Archival Bond,” Duranti appears to generally concur with this definition.[79] Duranti extends this definition by arguing that the archival aggregate as whole is necessary for it to function as evidence,[80] and that the meaning of the whole is damaged when one removes member records.[81]

Jenkinson, an archivist working for the British government in the early 20th century, also makes the connection between the aggregated nature of the archives and its capacity for evidence. Jenkinson argues that the viewer is less likely to misinterpret records when he or she contextualizes them in an understanding of the holistic structure of the originating administrative offices.[82] Arguably, the viewer may also use the neighbouring records in a fonds or a series to understand what a particular record means. A viewer who engages in this kind of contextualization reads the content of one record, registers it; moves to the next record, registers *its* content, and so on. This type of context-building, made possible when records of a shared origin are kept together, enables a scholar to assess the evidence presented by the archives. According to this view, the archival bond is located in the interpretive act of reading

[79] Duranti, “Archival Bond,” 214. Duranti appears to agree with this notion of evidence except in one regard: she notes that archives are not evidence until a scholar makes a claim in relation to that archives. This seems like one of those conundrums involving trees falling in forests, so I leave it aside.
[80] Duranti, “Concept of Appraisal,” 335.
[81] Duranti, “Archival Bond,” 217.
[82] Jenkinson, *Manual of Archive Administration,* 12.

the archives, rather than in the process of records creation. This characterization of the archival bond accounts for the capacity of the archives to serve as evidence.

When I wrote to archival theorist Giovanni Michetti for clarification as to whether the archival bond exists as a result of the reader's interpretation, he suggested I consider a legal case where someone who commits a crime is guilty.[83] In such a situation, it may be that the judges and lawyers have an idea of what has happened. However, the fact of being guilty or not guilty can be said to exist apart from these ideas. In the ensuing discussion, I realized that if I claimed the archival bond exists as an act of interpretation, I was effectively opening Pandora's box. That is, different people can connect the pieces of information in an archives in different ways, and in doing so create a completely different version of the event to which the archives refers. Undoubtedly, this conflicts with the concept of the archives as a vehicle of evidence.

To think through this issue, I reread the articles which first set me on the path to thinking that the archival bond is an interpretation. One of these articles was Michetti's "Archives are not Trees," where he shows that the ubiquity of the hierarchy as a tool for structuring information has given rise to a belief that hierarchies have a separate existence from the information they structure. A closer examination reveals that a hierarchy requires content in order to express its meaning.[84] In acknowledging this fact, it is clear there are two kinds of relationships associated with hierarchies: in one version of a hierarchical relationship, the subordinate element is a component of a whole; in a second version of a hierarchical relationship, a subordinate element is a variant of a more general concept. Given that it is common practice to represent a fonds

[83] Giovanni Michetti, email to the author, November 4, 2015.
[84] Michetti, "Archives are not Trees," 1008.

with a hierarchy, it follows that archivists have always specified relationships as part of their work. Michetti suggests that this practice of specifying relationships should become explicit, and could expand to include other kinds of relationships. This would result in a web-like representation of an archives.[85]

The notion that the archives may be better represented as a web called to mind the writings of Willard Van Orman Quine, a 20th century logician whose work "Two Dogmas of Empiricism" posits that human knowledge can be understood as a web. This article begins with Immanuel Kant's distinction between analytic and synthetic statements. According to Kant, analytic statements have a special existence, as they can be determined as true or false without consulting one's senses.[86] For example, one can determine that [a = a] is true without looking at the external world. By contrast, synthetic statements require that one views the world to evaluate if the statement is true or false. The claim 'this computer is black,' is a synthetic statement because one needs to view the computer in question to ascertain if the statement is true.

Quine looked closely at Kant's various definitions for analytic statements and found them all to be deeply flawed.[87] In light of these difficulties, Quine suggested that analytic statements have no special ontological status. Instead, he proposed that all human knowledge can be understood as statements existing in an interrelated web. To be included in the web, a statement must make sense with other statements. On this view, analytic statements have no

[85] Ibid., 1009.
[86] Quine, "Two Dogmas of Empiricism," 31.
[87] Ibid., 32-45.

special existence. Instead, their special nature comes from the fact they are fundamentally rooted in everything we know, such that it is highly unlikely they will ever be untrue.[88]

Quine's conceptualization of knowledge could explain how it is that the archives functions as evidence. An archives, like the web of human knowledge, is comprised of claims that supposedly make sense with one another by virtue of pertaining to a particular time and place. In other words, the archival bond can exist in the head of the person who views the archives, so long as that person cross-references the information appropriately. Given that this is such radical reinterpretation of the archival bond, a new term seems warranted: the *modal* bond, a word that signifies the possibility and contingency associated with connecting pieces of information.[89] However, this new name is not meant to suggest that the modal bond does not fulfil traditional requirements of the archival bond. Clearly, it does: it is fundamental to the archives as it enables it to perform as evidence.

To examine this process of cross-referencing information more closely, I looked at qualitative research methods. Because qualitative researchers do not believe in objective reality, their goal is instead to secure an in-depth understanding of the phenomenon they study.[90] This is achieved by looking at the phenomenon from various perspectives in a process known as triangulation.[91] Triangulation involves observing both when claims are repeated, which reinforces the validity of their content, and when claims conflict. Robert Stake shows that by

[88] Ibid., 51.
[89] More specifically, the definition of modality is: "the classification of logical propositions according to their asserting or denying the possibility, impossibility, contingency, or necessity of their content." *Merriam-Webster Online*, s.v. "modality," accessed April 13, 2016, http://www.merriam-webster.com/dictionary/modality.
[90] Denzin and Lincoln, introduction to *Sage Handbook of Qualitative Research,* 5.
[91] Stake, *Qualitative Case Studies,* 454.

restricting the scope of an inquiry to a single case, the researcher is able to thoroughly triangulate a variety of perspectives.[92] It could be argued that a fonds too enables a thorough triangulation for restricting its scope to the documentary output of a single person, organization, or community. Stake shows that conflicting claims are valuable precisely because they show the case through different eyes.[93]

Portelli's interviews with the citizens of Terni has elements of a case study, and his exploration into inconsistencies among the narratives illuminate the value of such inconsistencies. When the majority of his interviewees identified the death of Trastulli as taking place in 1953, but some identified the death as taking place in 1949, Portelli looked closely at the data and realized that being unable to avenge the death of a friend for a period of three years was a source of humiliation for the workers.[94] In other words, Portelli looked to make sense of conflicting claims, implicitly operating on the assumption that they existed in a coherent web of knowledge. In doing so, he obtained a clearer picture of what happened and arguably, this is good evidence.

So far, I have argued that the archives is a social construct comprised of the perspectives of those who participated in the creation of the records. These perspectives take the form of distinct claims. A later reader (such as a historian) is able to compare these claims to one another using the modal bond, and in this way evaluate the veracity of the information in the

[92] Ibid., 450.
[93] Ibid., 454.
[94] Portelli, *Death of Luigi Trastulli,* 26.

archives. On the assumption that claims are mutually relevant, similar claims are taken as true, while dissimilar claims represent a contested truth.

2.5 Corrupting the Evidence?

In his *Manual for Archives Administration*, Jenkinson made an example of archivists who ordered state papers according to an artificial classification scheme,[95] describing their actions as "unfortunate," and "unreasonable."[96] Muller, Feith, and Fruin concur, arguing that artificial arrangements are "inadequate" and "superficial"[97] and that archivists encountering such arrangements have permission to undo them. Archivists have understood these passages to mean that they should avoid rearranging the archives when possible. Indeed, these classical theorists give them two options: either leave the order as they found it, or attempt to recreate original order by mapping records to departments. Consequently, archival work appears to be either very passive, or very mechanical.

However, it seems doubtful that either the three Dutch archivists or Jenkinson thought archival work was passive or mechanical. After all, as archivists themselves they were familiar with the complexity of their holdings. Muller, for example, was awed by Van Riemsdijk's ability to study the records and discern what happened in the originating organization.[98] Jenkinson saw the records as a complex entity, evolving out of miscellaneous aggregates that split, disappear and

[95] Examples of artificial schemes are chronological and alphabetical. For a discussion of these ways of organizing see Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 49.
[96] Jenkinson, *Manual of Archive Administration,* 32.
[97] Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 59.
[98] Ketelaar, "Archival Theory and the Dutch Manual," 57.

merge for reasons as varied as a political shift or a paper sized too big for its neighbours.[99] These passages suggest that classical theorists viewed the archives as a complex entity, and the management of these entities as complex work.

So why are contemporary theorists working so hard to battle a rhetoric that archival work is passive and mechanistic?[100] In my view, there are two origins to this rhetoric. One, the classical theorists wrote their statements very strongly. The Dutch *Manual* is littered with "musts" and "shoulds" and has been described as "autocratic" in tone.[101] Arguably, these strong statements reflect Muller's goal to achieve widespread adoption of the *Manual* – his intent was to bowl over his opponents, and a display of strength served this end.[102] Jenkinson, perhaps influenced by the *Manual*, similarly *asserts* the definition of an archives in a way that suggests no room for alternative opinions. For classical theorists, acknowledging an ambiguity in the nature of an archives would have required a completely different tone. A change of tone being too difficult, they simply pretended that the contradictions inherent to an archives did not exist.

Terry Cook identifies a second origin to the rhetoric that archival work is passive and mechanical. Cook argues that it is historians who have perpetrated the myth of passive archival work, and their reason for doing so is to convince themselves that they are the first to access an archives. In support of his point, Cook presents an extensive set of quotes from 19th century

[99] Jenkinson, *Manual for Archive Administration,* 23-28.

[100] Theorists challenging this rhetoric include, but are not limited to: Terry Cook, Jennifer Douglas, Wendy Duff, Verne Harris, and Heather MacNeil.

[101] Barritt, introduction to *Manual of Arrangement and Description,* xxxix.

[102] As further evidence of his strong approach to implementation of the *Manual*, Muller attempted to make the Manual's dictates "binding by ministerial decree," Ketelaar*, Archival Theory and the Dutch Manual,* 60.

historians describing the thrill of being the first to enter an archives.[103] To maintain this illusion of primacy of access, it was necessary to erase archivists from the process by characterizing archival work as passive and mechanical. However, if one looks closely at historian statements, a contradiction emerges. Historian Geoffrey Rudolph Elton says that archivists should conduct a "ground-clearing" operation, the product of which "adheres to the state of the material as they found it."[104] Ground-clearing, to me, suggests a big impact, while adhering to the original state suggests no impact at all. In other words, archivists must influence the records, and simultaneously *not* influence the records.

These conflicting requirements are partially what makes archival work complex. To make the records accessible, while maintaining their authenticity, archivists must consider a wide variety of factors. This became apparent to me in a project I did for ARST 545 - *Advanced Arrangement and Description,* where I subjected a fonds that included a large percentage of correspondence to a social network analysis, and compared the social-network generated groups to the series identified by the archivist. [105] There was almost no overlap between the two ways of subdividing the fonds. Giovanni Michetti commented that the social network analysis only takes one factor into consideration when grouping records – the intensity of interaction. By contrast, the archivist takes into consideration a variety of factors, weighs them against one another, plays with alternatives, and generally deploys a complex human process in doing so.[106]

[103] Cook, "Archive(s) is a Foreign Country," 506.

[104] As cited by Hurley, "Personal Papers and the Treatment of Archival Principles," 156.

[105] Giovanni Michetti, email to the author, January 7, 2015. Credit to Giovanni Michetti for suggesting this project.

[106] Giovanni Michetti, email to the author, April 11, 2015.

Part of the reason arrangement and description is a complex process is because the fonds represents a complex thing – a confluence of perspectives. Given this, is it even possible for the archivist to achieve the traditional aim of staying true to what is represented by the records? Case study researcher Robert Stake suggests the answer is both yes and no. On the one hand, the case tells a story, but this story "exceeds anyone's knowing and anyone's telling."[107] Therefore, the researcher must select a subset of the case when writing up the final report. This sentiment is echoed by archivist Adolf Brenneke who rebelled against the rules in the *Manual* and argued that archival work involved selecting the subset of relationships the archivist deemed most important.[108]

On the other hand, Stake notes that despite the researcher's influence, it is expected that a case should be described in such a way that the reader "can experience [the case's] happenings vicariously and draw their own conclusions."[109] This statement shows a clear convergence between the goals of the archivist and the goals of the case study researcher, which is to present information pertaining to a particular time and place. This makes clear that it is possible for the archivist to facilitate the role of the records as evidence. In doing so however, they cannot avoid imposing their influence on the records.

[107] Stake, "Qualitative Case Studies," 456.

[108] Horsman, "Last Dance of the Phoenix," 20. While it isn't exactly clear what Brenneke meant by the term "relationship," it is clear that he is discerning the subset of an interrelated thing, which makes this relevant for the discussion.

[109] Stake, "Qualitative Case Studies," 450.

2.6 Records and Social Network Analysis

So far, I have used this literature review to contextualise the project in a larger debate, and to describe the conceptual lens by which I will explore the findings of the study. A secondary purpose of the literature review is to present similar projects as a means to ensure that the proposed study is original and worthy of exploration. There is limited research in the area of social network analysis and records, with the exception of two notable projects.

Maria Esteva's dissertation research explored the application of a text mining tool to the fonds of an organization. The fonds was comprised of aggregates which had each been maintained by a single employee, enabling Esteva to compare the extent to which employees used a similar vocabulary in the content of their records.[110] As part of her analysis, she produced a graph showing employees as nodes, with the thickness of edges varying in accordance with the similarity of vocabulary.[111] Employees with similar vocabularies to many others were placed in the centre of the graph.[112] Esteva suggests that the information within the graph can be used to discern the "context and structure"[113] represented by a large aggregate of records. She argues that in discerning this information, archivists are better positioned to preserve the aggregate.

[110] Esteva, *Aleph in the Archive*, viii.

[111] To determine if documents are similar, Esteva uses a method called Term Frequency-Inverse Document Frequency Approach (Tf-idf) to measure the similarity of documents. "Tf-idf considers the length of the document in which a word appears, whether the word is rare or common in relation to the document, and whether it is rare or common in relation to all the documents involved in the set." Esteva, *Aleph in the Archive*, 109.

[112] Ibid., 129.

[113] Ibid., 3.

Esteva describes the kind of information made available by her network visualizations. In one visualization, she observes that "at the centre of the network is the director,"[114] with the centrality of his node position explained by the work required of him as the manager of two important programs within the organization, and his responsibility in editing the organization's official documentation.[115] Additionally, Esteva believes the thick edges connecting the director to the area managers reflect the "strong ties"[116] they share in real life.

In attempting to discern the holistic structure of a large aggregate of records, Esteva's work is arguably the forerunner of the present study. Esteva's work is also a forerunner of the present study for using time to structure the information represented by the records. That is, she created a graph for each year of records and in doing so was able to correlate changes in vocabulary-based relationships to the changes in the organization. For example, she notes that the node representing the receptionist, previously located on the periphery on the graph, became more central during 2001, the year the records relating to the primary responsibility of the receptionist switched from being paper-based to being electronic.[117] Interestingly, Esteva called these explanations for time-based developments "stories."[118] This is consistent with the views of narrative inquiry researchers, who note that in Western cultures, narratives tend to involve a temporal element.[119]

[114] Ibid., 142.
[115] Ibid. Apparently, being in charge of programs and editing official documentation means his documents are similar to others.
[116] Ibid.
[117] Ibid., 259.
[118] Ibid., 152.
[119] Patterson, "Narratives of Events," 31.

Jana Diesner, Terrill Frantz, and Kathleen Carley also conducted a project involving records and social network analysis. They extracted the sender and the receiver metadata from the workplace emails of the Enron Corporation, and used this information to visualize the *components* in the social network – with a component defined as a set of people who sent an email to at least one other person in the group.[120] They discovered that Enron's financial crisis led to the formation of fewer and larger components, indicating that previously disconnected employees began to communicate. They additionally discovered that these new communication lines formed between people of different rank.[121]

To analyze this data, they deployed charts showing the passing of time on the x-axis.[122] This temporal information is augmented by three vertical lines which mark what the researchers believe are turning points in the Enron crisis: Jeffrey Skilling succeeding Kenneth Lay as CEO in December 2000; Skilling's resignation in August 2001; and Enron's motion to file for bankruptcy, which took place in December 2001.[123] As with Esteva, it could be argued that Diesner, Frantz, and Carley use narrative as a strategy for making sense of information in their extensive dataset.

2.7 Summary

In this literature review, I described two ways that provenance has been conceived by traditional theorists. One way of conceptualizing provenance, articulated by Muller, Feith, and

[120] Ibid., 213.
[121] Ibid., 224.
[122] Ibid., 212.
[123] Ibid., 213.

Fruin, involves organizing records by department and their subfunctions. The rationale put forward to justify this notion of provenance is that records produced by the same department are mutually relevant for pertaining to the same context. A second way of conceptualizing provenance, espoused by theorists such as Duranti, asserts that archivists should respect the physical[124] accumulation of records, on the view that these accumulations represent a shared activity. Like the Dutch trinity, Duranti implicitly values mutual relevance in an aggregate, with mutual relevance meaning that the constituent records refer to the same events, concepts, or processes.

To explore the notion of mutual relevance in greater detail, I used the ideas of Grigely, MacNeil, and Portelli, to show that an archives may be understood as the space between tellers and listeners. This socially-constructed archives brings together the perspectives of these tellers and listeners who are part of the same community. Using the ideas of Michetti, Quine, and Stake, I justified this model of the archives by showing that bringing together these perspectives enables the archives to function as evidence.

In the next chapters, I explore the notion of mutual relevance in greater detail by using social network analysis to bring together records pertaining to the same working groups in an organization. This process is described in the next section.

[124] The term “physical” usually refers to paper-based documents in close proximity to one another. I use it also to refer to electronic documents which are perceived as being close to one another for existing in the same folder.

3 Methods

3.1 Research Questions

As stated in the literature review, traditional notions of provenance imply that records in archival aggregates are mutually relevant. To better understand the notion of mutual relevance, I applied a social network analysis to the records produced by a non-academic department at the University of British Columbia. I believed that doing so would identify groups of people who work together, and that their records would be mutually relevant. To be clear, my idea of mutual relevance is that the records refer to the same events, concepts or processes. With this in mind, the questions that drive this research are:

1. Are groups of records formed by social network analysis mutually relevant?
2. Do communities formed by social network analysis represent the working groups in an organization?
3. Assuming that employees will explain the social network visualization using narrative, to what extent do these individual narratives converge with one another?
4. What does the application of social network analysis to records tell us about the principle of provenance?

3.2 Overview of Research Design

The procedure for this project is divided into two distinct stages. The first stage involved a quantitative analysis of records metadata. More specifically, it involved extracting the name of the creator, the name of the modifier, and the date from each record stored in the online platform used by the department under study. This information was subject to several community detection algorithms, with the results of each algorithm visualized as a time-based series of graphs. The second stage involved presenting the social network visualizations to the staff who created the records. The presentation of the visualization took place in the context of a qualitative interview, where staff were asked what social network algorithms they believed best fit with their experience of the organization. Because this project involves both quantitative and qualitative methods, it qualifies as a *sequential mixed methods approach*[125] to research.

3.3 Ethics Approval

This project was approved by the Ethics Review Board at the University of British Columbia on 23 October 2015. Interviewees signed a consent form permitting the researcher to access their workplace files and run the social network analysis. They signed an additional form prior to the interview consenting to be audiotaped. After interviews were transcribed, participants reviewed a condensed version of the transcript, and signed a third form permitting the inclusion of their comments in this thesis. I changed their names in the final report to protect their identities.

3.4 Quantitative Analysis

3.4.1 Dataset Description

This project focuses on a set of records created by a single department within the University of British Columbia. The department consisted of 10 full-time staff, including the manager.

[125] Hewson, *Sage Dictionary of Social Research Methods,* 180.

Additional details regarding the department have been withheld to avoid jeopardizing anonymity.

After staff signed an initial consent form, I distributed a survey asking them to identify the SharePoint sites to which they had access. Some had access to as many as 18 SharePoint sites, while others accessed only three. I identified the five SharePoint sites to which most department members had access, and sent this information to the staff at University of British Columbia Information Technology (UBC-IT). UBC-IT extracted the records from these SharePoint sites, which totalled 1064 records. Next, UBC-IT staff extracted the metadata from each record. They sent me the extracted metadata in the form of a spreadsheet, where each row represented an instance of document modification. The first column indicated the SharePoint site where the document under modification originated. Subsequent columns indicated document name, the date it was created, who created it, who modified it, and the date of modification. The first record was modified in February 2014, and the last in December 2015.

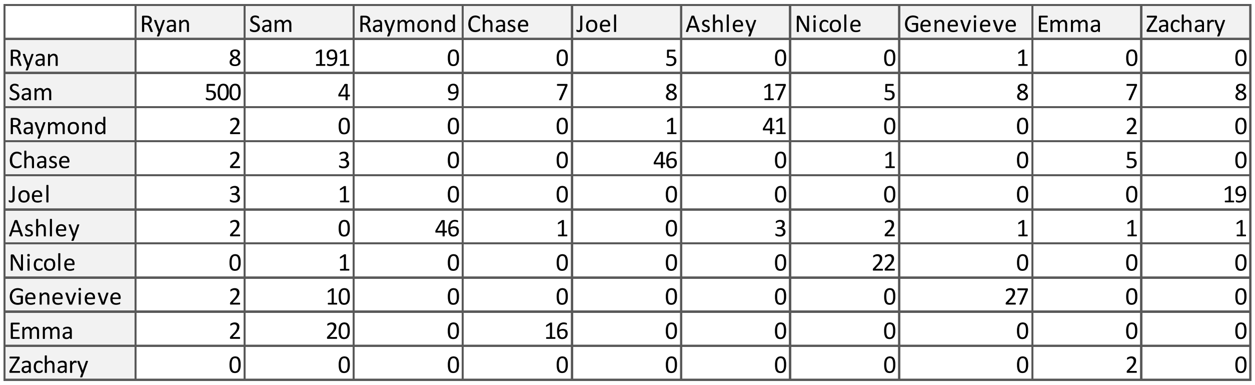
3.4.2 Dataset Preparation

I wanted to show interviewees how groups of intensely-interacting individuals were changing over time, as doing so would enable them to correlate changes in the visualization to changes in the department. To this end, I visualized distinct time periods. To determine the most appropriate number of months per time period, I set a criteria: I wanted to achieve the fewest number of months per snapshot to make the visualization as informative as possible. Additionally, I wanted the visualization to have no empty, or nearly empty, snapshots, as these clutter the visualization and distract the viewer. To determine the most appropriate number of months, I used an R package called 'ndtv,' which stands for Network Temporal Dynamic Visualizations.[126] This package animates the appearance of edges between nodes as they occurred over the 23 month period, and permits the user to adjust the length of the time frames that collectively create the animation. When I set the time frames to a three-month period, there were lulls in which no activity occurred, and lulls in which very little activity occurred. By contrast, setting the frames to represent a four-month period showed activity involving multiple nodes during every frame, suggesting that four months would result in a readable yet informative representation of the data.

To derive an edgelist[127] from the dataset sent to me by UBC-IT, I deleted all but the creator and modifier columns. Next, I converted the edgelists into weighted adjacency matrixes. In an adjacency matrix, each person in the dataset appears twice: once in the first column, and once in the column headers (for an example, see *figure 3—1*). Any given cell in the matrix represents the total document modification between the person listed in the row of the cell, and the person listed in the column of the cell. To create these adjacency matrixes, I used Microsoft Excel to set up the headers and columns of the soon-to-be-filled adjacency matrix, and input a formula into the cells which counted the number of relevant pairs in the edgelist. The formula for this procedure can be found in appendix A.

[126] Bender-deMoll, *ndtv: Network Dynamic Temporal Visualizations* [Computer software].

[127] Newman, *Networks*, 111. An edgelist is defined as a list comprised of pairs of connected nodes.



	Ryan	Sam	Raymond	Chase	Joel	Ashley	Nicole	Genevieve	Emma	Zachary
Ryan	8	191	0	0	5	0	0	1	0	0
Sam	500	4	9	7	8	17	5	8	7	8
Raymond	2	0	0	0	1	41	0	0	2	0
Chase	2	3	0	0	46	0	1	0	5	0
Joel	3	1	0	0	0	0	0	0	0	19
Ashley	2	0	46	1	0	3	2	1	1	1
Nicole	0	1	0	0	0	0	22	0	0	0
Genevieve	2	10	0	0	0	0	0	27	0	0
Emma	2	20	0	16	0	0	0	0	0	0
Zachary	0	0	0	0	0	0	0	0	2	0

Figure 3—1 Weighted Adjacency Matrix

3.4.3 Visualization Software

This section details the evaluation and selection of visualization software. My software options included ndtv,[128] NodeXL,[129] Gephi,[130] and MatLab.[131] I ultimately determined that R[132] would best serve the needs of this project.

As stated in section 3.4.2, ndtv animates the appearance of edges between nodes. Despite the clarity that comes with animation, ndtv was not appropriate for this project as node labels were restricted to numbers, which would likely confuse interviewees. Further, while ndtv permits one to colour nodes manually, it did not have a means of automatically colouring nodes belonging to the same community. Gephi was not suitable for this project as it only offered one community detection algorithm, which limited analysis. NodeXL was also a poor fit for this

[128] Bender-deMoll, ndtv: Network Dynamic Temporal Visualizations [Computer software].
[129] Milic-Frayling et al., *NodeXL* [Computer software].
[130] Bastian, Heymann, and Jacomy, *Gephi* [Computer software].
[131] Moler, Little, and Bangert, *MatLab* [Computer software].
[132] R Core Team, *R* [Computer software].

project, as it does not allow one to specify the position of each node. This meant that nodes representing the same person did not retain their position from one time-based visualization to the next, distracting viewers from the changing communities. According to my research, MatLab would likely give me the ability to determine the position of the nodes, but as proprietary software it was out of my reach.

Ultimately, I selected R, an open-source software environment with the capacity to manipulate the visualization using a programming language. In R, it is possible to download user-contributed packages that add functionality to the basic program. I tried several packages relating to social network analysis, and discovered that the igraph[133] package included community detection algorithms, and was capable of making communities distinct by encircling the member nodes in a coloured shape. It also allowed me to position the nodes in the same place on the graph for each time-based visualization of the dataset.[134] For these reasons I decided to visualize the data with igraph.

3.4.4 Selection of Algorithms

An algorithm is a step-by-step set of operations, often performed by a computer.[135] Taking the information in an adjacency matrix as their input, community detection algorithms execute a

[133] Csárdi and Nepusz. *igraph* [Computer software].

[134] The layout options available in R, such as Fruchterman-Reingold, optimize the position of the nodes so that as few edges are obscured as possible. The optimal layout avoids instances where edges run through a node, as well as instances where edges cross one another. For my project, this layout scheme caused problems: placing nodes in the most readable positon meant that each time-based visualization had a different set of node positions, making it hard to see how communities changed from one frame to the next. In anticipation of a large dataset with many nodes, I wrote a layout algorithm that positions communities separately from one another. I later discovered that the dataset for this study only contained 10 nodes such that it made more sense to bypass layout algorithms altogether and put the nodes into a star formation.

[135] *Oxford English Dictionary*, 10th ed., s.v. “algorithm.”

series of steps that identify groups of people who frequently interact with one another, and infrequently interact with others in the larger network.[136] Community detection algorithms vary in their procedures.[137] I selected Michelle Girvan and Mark Newman's ground-breaking "Edge-betweenness" algorithm because it served as the basis for a number of other algorithms, suggesting an enduring relevance.[138] I also selected Aaron Clauset, Mark Newman, and Cristopher Moore's "Fast Greedy Modularity Optimization" algorithm because I was curious if its procedure would produce useful results.[139] After making these decisions, I received the dataset and realized it was very sparse. As Martin Rosvall and Carl Bergstrom's "InfoMap"[140] algorithm performed well in trials for sparse networks,[141] I included this algorithm in the project as well.[142] In the next paragraphs, I describe the procedures used by each algorithm.

The Fast Greedy algorithm starts from a set of isolated vertices,[143] then pairs the vertices in all possible pairs. For each pair, the modularity is measured, with modularity being a value assigned to a line that separates a group of vertices from the rest of the network. A line with

[136] Girvan and Newman, "Community Structure in Social and Biological Networks," 7821.
[137] Descriptions of these procedures have been articulated by Tamás Nepusz as well as Andrea Lancichinetti and Santo Fortunato. See Lancichinetti and Fortunato, "Community Detection Algorithms," as well as Nepusz, "What are the Differences between Community Detection Algorithms in igraph?" Last modified February 28, 2012. http://stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph/9478989#9478989.
[138] Ibid., 3-4.
[139] Ibid., 3. As noted by Lancichinetti and Fortunato, the primary strength of the Fast Greedy algorithm is its capacity to handle large datasets. As this project involved a small dataset, another choice may have been more appropriate.
[140] Lancichinetti and Fortunato, "Community Detection Algorithms," 4.
[141] It should be noted that Lancichinetti and Fortunato's trials did not include weighted and directed graphs, which is the type of graph used in this project. In other words, I selected InfoMap based on its general performance in Lancichinetti and Fortunato's trials.
[142] The reader may be interested to know that I applied to the data an algorithm very similar to InfoMap known as "Cluster Walktrap." Results were nearly identical. For more information on Cluster Walktrap see Pons and Latapy, "Computing Communities in Large Networks using Random Walks."
[143] To be clear, vertice is another term for node. Newman, *Networks,* 1.

high modularity has many edges within the space it defines, and few edges crossing the line.[144] The program establishes the pair with highest modularity as a group, then repeats this process: vertices and established groups are again paired with one another, and the network divisions are again measured for modularity. The process of combining vertices and testing their modularity continues until all vertices form a single group. Then, the computer looks at each group of nodes that was created in the process, and identifies the groups with the highest modularity.[145]

Girvan and Newman's "Edge-betweenness" algorithm measures all possible edges in the network for "betweenness," with betweenness defined as the number of distinct sequences of nodes and edges that can possibly run through the edge.[146] Once the edge with the highest "betweenness" is identified, it is removed from the dataset and the algorithm is run again to determine the edge of greatest "betweenness" in the remaining edges. As edges are progressively removed from the graph, islands of connected nodes emerge.[147] The process of removing edges continues, splitting the islands until a single edge remains.[148] This process can be represented with a hierarchical tree known as a dendogram,[149] which makes clear the groups of nodes that remain with the removal of each edge. Newman asserts that the user should select the hierarchical division which is most appropriate for their purpose.[150] I used an

[144] Clauset, Newman, and Moore, "Finding community Structure in very Large Networks," 1.

[145] Newman, *Networks*, 382.

[146] Wasserman and Faust, *Social Network Analysis,* 107. Newman notes that to shorten the processing time, it is possible to restrict the length of the path. More specifically, betweenness can be calculated using only geodesics – with geodesics defined as the path between vertices that is shortest. Newman, *Networks*, 382.

[147] Girvan and Newman, "Community Structure in Social and Biological Networks," 7823.

[148] Gregory, "An Algorithm to Find Overlapping Structures in Communities," 92.

[149] Gregory defines dendogram as "A binary tree in which the distance of nodes from the root shows the order in which clusters were split." Gregory, "An Algorithm to Find Overlapping Structures in Communities," 92.

[150] Newman, *Networks,* 384.

option in the igraph program which determined the dendogram aggregates with the greatest modularity.[151]

Rosvall and Bergstrom's "InfoMap" algorithm works by initiating a random walk, defined as a process which starts at some node, then randomly traverses an edge connected to that node. Arriving at the new node, an edge is again selected at random (it could be one leading back to the first node). The process repeats until all nodes have been included on this walk.[152] The algorithm capitalizes on the way that nodes in a community are closely connected such that they frequently redirect the walk to one another. Various sequences of nodes are hypothesized as a community and their entropy is measured, with entropy being defined as the difference between the nodes connected by the walk's movement and the nodes which are the result of a random selection.[153] This "entropy of movements within modules"[154] is weighed against the "entropy of the movement between modules,"[155] with low and high entropy signifying an appropriate network division. For complex graphs, additional measures such as simulated annealing and greedy search are enacted to "check all possible partitions."[156] [157]

[151] A description of the parameters one can choose in using the edge-betweenness algorithm is found here: Csardi, "Community Structure Detection Based on Edge Betweenness," http://igraph.org/r/doc/cluster_edge_betweenness.html

[152] Newman, *Networks,* 157.

[153] Pflatz makes clear that entropy has two meanings in the context of social network analysis. I chose the one that made the most sense, which is "the difference between the actual transmission and a purely random signal." Pflatz, "Entropy in Social Networks," 1.

[154] Rosvall and Bergstrom, "Maps of Random Walks on Complex Communities," 1120.

[155] Ibid.

[156] Ibid., 1121.

[157] I think it is worth noting that Rosvall and Bergstrom frame the problem of community detection in an interesting way. That is, they assign unique binary numbers to each node in the walk, and concatenate them, which means the walk can be represented as a string of 0's and 1's. Then, the problem becomes one of compressing the string while retaining the important information. By identifying the segments of the string representing the nodes of a community, they insert a code that indicates when the walk moves into the territory of a new community. It isn't quite clear to me how the computer detects a pattern of the repeated number representing a community, but it doesn't seem entirely unfeasible either. By creating a code which indicates the

3.4.5 Roles Spanning Working Groups

One variant of the betweenness algorithms that might be useful in future projects (but wasn't available in R) is Steve Gregory's Cluster-Overlap Girvan Newman Algorithm (CONGA). This algorithm operates on the assumption that some nodes are equally members of the two communities they border, and in such cases, it is more accurate to represent them as being located on the border than as members of either community. To see if this is the case, the algorithm "splits" the node into two copies of itself. Then, the "betweenness" of the edge connecting the duplicate nodes is calculated. If a higher number of unique sequences of edges and nodes traverse the newly-constructed edge compared to any real edge, it suggests that the node itself is the locus of betweenness. This strikes me as a very useful variant of the edge-betweenness algorithm; managers, for example, who equally participate in two communities can be acknowledged as border elements, resulting in more accurate representations.[158]

Because this algorithm was not available, and because I was concerned that the manager would obscure working groups by participating in all of them, I removed the manager from the dataset and applied both the Fast Greedy and the Edge-betweenness algorithm to the results. Neither algorithm identified working groups, showing instead isolated vertices. By contrast, applying the InfoMap algorithm to this reduced dataset gave rise to distinct working groups. Therefore, to test my concern that the manager was obscuring working groups, I decided to show staff four visualizations: one showing the results of the Fast Greedy algorithm, a second showing the

switch into a new community, one is able to compress the string, as the binary representations of numbers get lengthier as one progresses through the binary numeral system; being able to reuse the first binary numerals in each community represents a significant space savings.

[158] Gregory, "An Algorithm to Find Overlapping Structures in Communities," 92.

results of the Edge-betweenness algorithm, a third showing the InfoMap algorithm, and a fourth showing the InfoMap algorithm run on a reduced version of the dataset with all instances of the manager's creation and modification removed.

3.4.6 Self-modified Files

The dataset included many instances where the same person both created and modified the record. I considered removing self-modified files on the rationale that what remained would represent activities shared in common. However, I discovered that these self-modified files comprised a high percentage of what was already a very small dataset. More specifically, the dataset was 1064 files, and 86% were self-modified. As modularity-based algorithms suffer from a resolution limit[159] it seemed wiser to include the self-modified files.

3.4.7 Visualization Shown to Staff

The visualization shown to employees during interviews can be found in appendix D. This visualization was printed on a large piece of paper to ensure the information would be readable. *Figure 3—2* shows a modified version of this visualization. In this modified version, node names were removed from all graphs but one to reduce clutter on the scaled visualization. This does not remove any important information, as the node representing a particular person retains the same position on all graphs. Algorithm A represents the Fast Greedy algorithm; Algorithm B represents Edge-betweenness algorithm; and Algorithm C

[159] Fortunato and Barthélemy, "Resolution Limit in Community Detection," 36.

represents InfoMap. Black rows indicate the number of instances of document modification that took place during the defined time period.

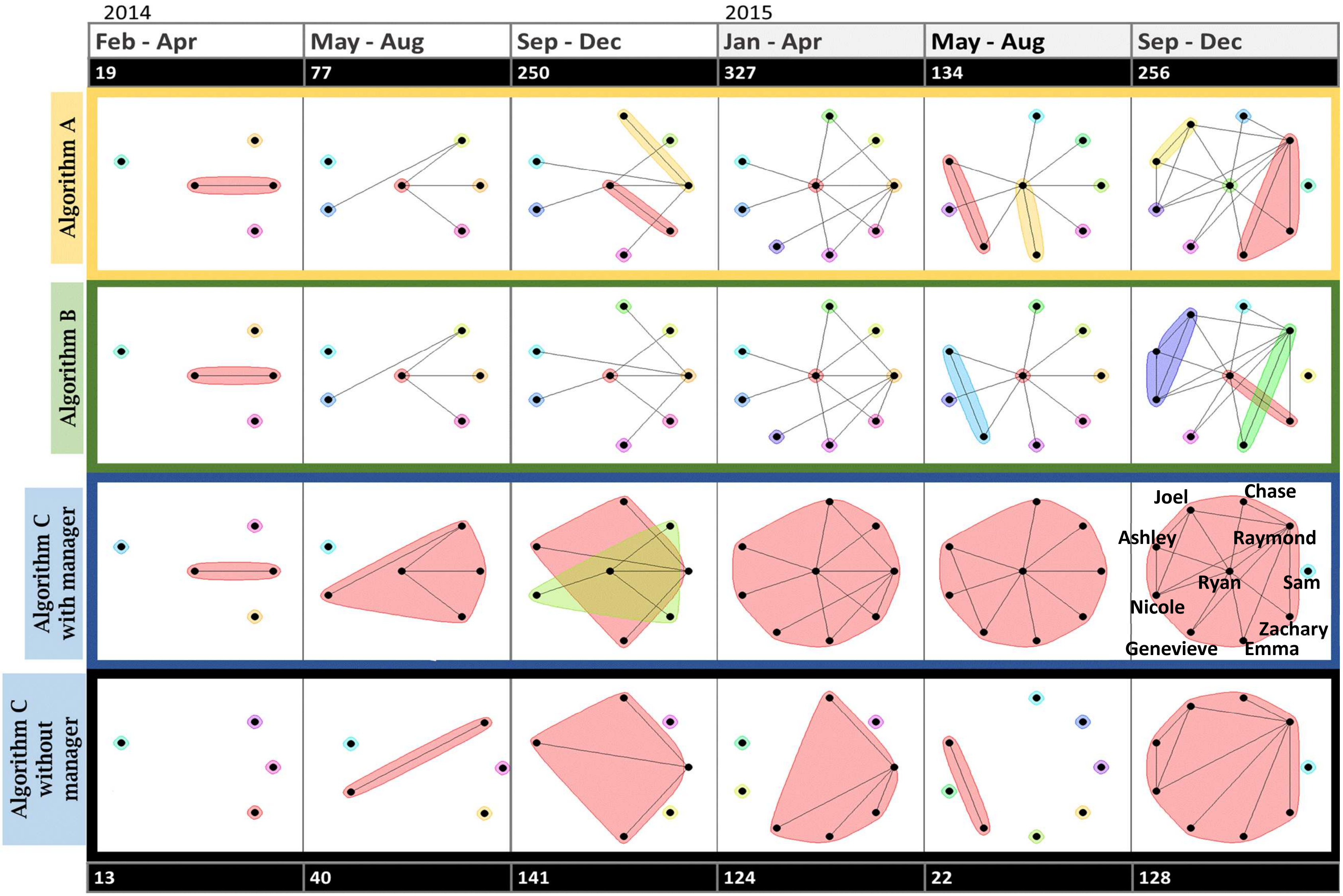


Figure 3—2 Visualization shown to staff during interview

3.4.8 A Mistake in the Coding

Quite late in the project, I realized I had made a mistake when writing the part of the code that converts the matrix to a graph. That is, I wrote that matrixes were "undirected." To be clear, when a matrix is interpreted as undirected, it means the only thing that counts is the number of instances of file modification between two people; the program is agnostic as to who created the file and who modified it. It is for this reason that undirected graphs show a line from one person to the other, while a directed graph signifies the creator and modifier with an arrow pointing from one to the other, as shown in *figure 3—3* and *3—4*.

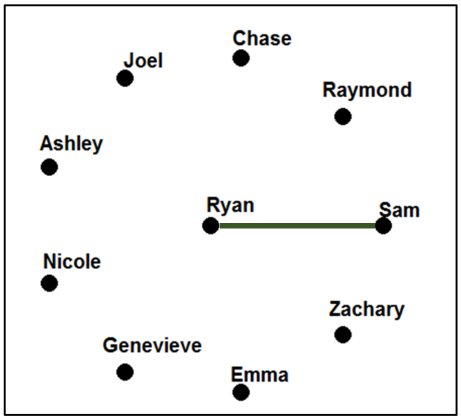


Figure 3—3 Undirected Graph

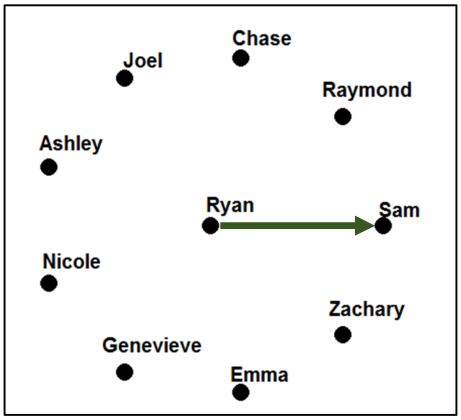


Figure 3—4 Directed Graph

When the computer is told to read the matrix as undirected, it assumes that only half of the cells in the matrix are relevant, as the other half merely switches the position of the creator and modifier from one axis to the other. On realizing I had conducted interviews with visualizations reflecting only half the information in the adjacency matrix, I ran each visualization again with the "directed" parameter. Of the 24 visualizations, only two would have been different if the coding mistake had not happened (*figure 3—5*), and as it turns out, the new results better fit with the perspectives of the interviewees.

The code for uploading the adjacency matrix to *R* can be found in appendix B.

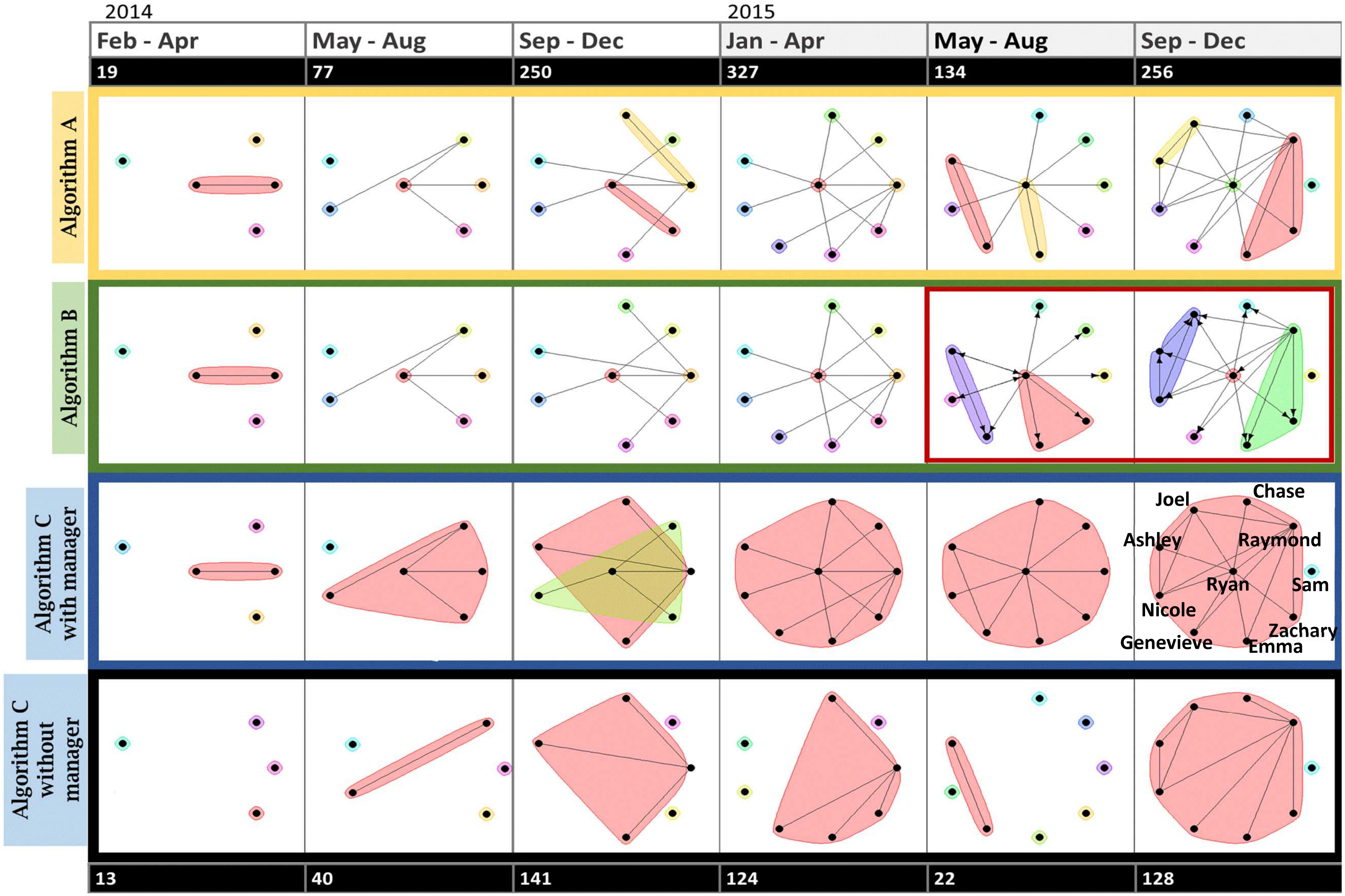


Figure 3—5 Four Algorithms with Corrected Visualization

3.5 Qualitative Analysis

This section describes the second of the two major stages in the procedure – the qualitative interviews with record creators.

3.5.1 The Boundary of a Case

According to Bent Flyvbjerg, the "decisive factor"[160] in determining a study as a case study is whether it can be defined within a boundary. As he notes, "a case study is not so much a methodology as a choice of what is to be studied."[161] From Flyvbjerg's perspective, this thesis qualifies as a case study because the focus of the research is a department with a clearly defined administrative boundary, made manifest by the clear articulation that the department's staff worked together. This articulation was voiced by both the manager, who identified the members of his staff, and by staff members themselves, who confirmed that the visualization included core members of the department.

The case study boundary does not have an objective existence; instead, it is a social construct. That is to say, the line representing the case study boundary could be drawn differently depending on the perspective of the one who views it. To illustrate, one staff member was in the process of transitioning into a sister department which had a similar mandate but a slightly different focus. Additionally, interviewees identified two people who seemed to be doing work within the scope of the department, but were not identified as staff by the manager. Possibly,

[160] Flyvbjerg, "Case Study," 301.
[161] Ibid.

they were either full-time staff who declined to participate in the project, or they had an unusual full-time status – perhaps reporting to multiple departments. The department boundary was also blurred by the manager's recently expanded portfolio, which included several new departments from across the university, and by the presence of Co-op students, who were part of the department but were not included in the study as they did not access the SharePoint platform.

Similar to Flyvbjerg, case study researcher Robert Stake asserts that a case is a choice of what is to be studied.[162] Stake further notes that the restricted scope of a case means that case study data pertains to the same set of circumstances. This is an important feature of a case study as it enables the researcher to "thoroughly triangulat[e]"[163] the data, which contributes to the credibility of the findings. It could be argued that this study is a case study because it involves obtaining information pertaining to a limited context – which is, the department under study. Further, this study involves examining the information for consistencies and inconsistencies.

However, it could be argued that despite fulfilling the basic qualification of a case study, this thesis does not quite achieve case study status. This is because, as Stake notes, the qualitative case study is also characterized by a small number of research questions which are focused on "complex, situated, problematic relationships."[164] In this study, the research questions were complex and problematic, but are not so much situated in the case as they are situated in archival theory.

[162] Stake, "Qualitative Case Studies," 443.
[163] Ibid.
[164] Ibid., 448.

3.5.2 Narrative Inquiry

Case study research is focused on the experience of those within the case.[165] To better understand these experiences I drew on the insights of a methodology known as narrative inquiry. This methodology was particularly relevant in the context of this study, as interviewees responded to the visualization by offering a contextualizing narrative.

Theorists in the field of narrative inquiry debate the meaning of the term "narrative." Phillida Salmon, for example, argued that narrative is an act that involves "imposing a meaningful pattern" [166] by drawing connections between pieces of information. In drawing these connections, it is clear that some kind of ordering is taking place. As Catherine Kohler Riessman notes, this ordering is essential as it enables a listener "to make sense of another's words."[167] Riessman's emphasis on the listener is echoed by Salmon, who notes that narratives are

> ..in a fundamental sense, co-constructed. The audience, whether physically present or not, exerts a crucial influence on what can and cannot be said, how things should be expressed, what can be taken for granted, what needs explaining and so on.[168]

Corinne Squire, Molly Andrews, and Maria Tamboukou describe a typology of narrative research. One type of narrative research requires that the researcher inquires into events that have happened to the storyteller.[169] A second type involves the researcher inquiring into the storyteller's experiences.[170] This project adopts a lesser-known third type which is concerned

[165] Ibid.
[166] Salmon, "Looking Back on Narrative Research," 78.
[167] Riessman, "Looking back on Narrative Research," 81. In Western cultures, this ordering generally takes the form of a temporal sequence, but alternatives are possible.
[168] Salmon, "Looking back on Narrative Research," 80.
[169] Squire, Andrews, and Tamboukou, introduction to *Doing Narrative Research,* 5.
[170] Ibid.

with co-constructed narratives, and the social patterns they reveal.[171] More specifically, this project seeks to understand a collective narrative which is specific to the department.

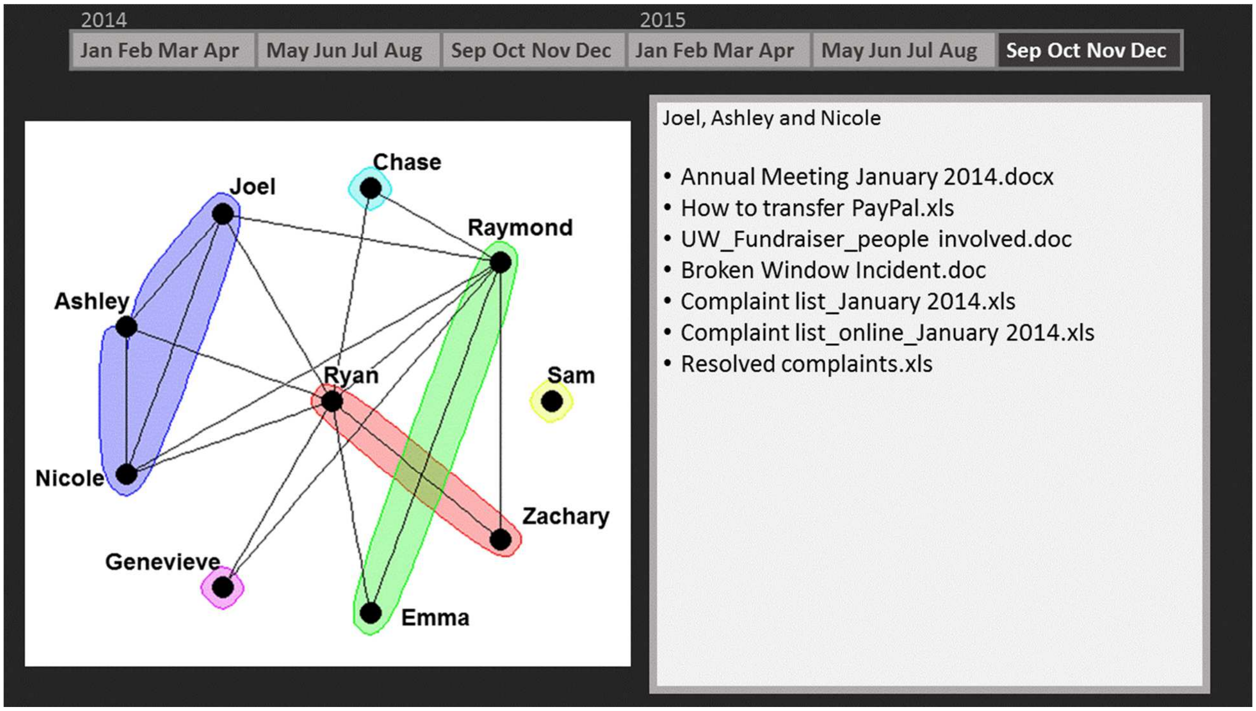
3.5.3 Interview Description

During interviews, a hardcopy of the visualizations created in section 3.4.7 was presented to employees. Employees were asked to choose the algorithm that best represented the pattern of file modification that involved them. I made clear that the last visualization was a repeat of Algorithm C (InfoMap), but with the manager removed, along with my reason for doing so. That is, I shared my hypothesis that managers modify so many files they make it hard for the algorithm to identify groups of frequently-interacting employees. By presenting a visualization that excluded the manager, I could test this hypothesis. Once employees identified a representative algorithm, a particular working group within the visualization was identified, and they were asked if they remembered participating in that working group. This question was repeated with different working groups; when relevant, interviewees were also asked about instances where they worked alone.

At this point, I showed them an interactive version of the visualization on my laptop. When the working groups on the screen were clicked, a list appeared showing the name of files produced by the members of the working groups (*figure 3—6*). Interviewees were asked if they thought the files belonged together. The interview always concluded with an open-ended question, in which interviewees were asked if they thought that organizing the records by social network

[171] Ibid.

analysis would help people make sense of the records 100 years in the future. The complete interview guide can be found in the appendix C.



Joel, Ashley and Nicole

- Annual Meeting January 2014.docx
- How to transfer PayPal.xls
- UW_Fundraiser_people involved.doc
- Broken Window Incident.doc
- Complaint list_January 2014.xls
- Complaint list_online_January 2014.xls
- Resolved complaints.xls

Figure 3—6 Screenshot of Interactive Visualization

3.5.4 Conducting Effective Interviews

I learned how to conduct an interview by reading Irving Seidman's "Interviewing as Qualitative Research." While Seidman does not identify as a narrative inquiry researcher, his approach is clearly informed by the idea of narratives. For example, he argues that "we interview in order to come to know the experience of the participants through their stories."[172] Seidman notes that listening is the most important skill in interviewing. As I am sometimes inclined to interrupt when something does not make sense to me, it was helpful to learn that these

[172] Seidman, *Interviewing as Qualitative Research*, 122.

questions can be brought up at a much later point in the interview.[173] Seidman's description of the "public voice"[174] also proved to be a very useful concept. The public voice is often used when the speaker is highly aware of their audience. To identify the "public voice," I found myself searching the facial expression and gestures of the interviewees. In the end, I realized that the process of looking for the "public voice" was primarily beneficial because it helped me pay attention. Seidman's suggestion to follow through on hunches was also illuminating. This made me realize that it is perfectly fine, and indeed beneficial, to use one's instincts to explore meanings during an interview.[175]

3.5.5 Selection of Interviewees

In this study, 10 people contributed to the records in five SharePoint sites. Ultimately, I conducted interviews with six people. In deciding who to interview, I took several factors into consideration. My first criterion was that the interviewee must have access to all five SharePoint sites which were visualized. In doing so, I respected the access restrictions set by the department. My second criterion was that the interviewee must have participated in the modification of records for at least six months of the 23-month dataset. I reasoned that these interviewees were more likely to be knowledgeable about the department and its records.

[173] Richardson et al., as cited in Seidman, *Interviewing as Qualitative Research,* 88.
[174] Which is based on Steiner's description of the "outer voice." Seidman, *interviewing as Qualitative Research,* 81.
[175] Seidman, *Interviewing as Qualitative Research*, 85.

3.5.6 Interview Analysis

The interviews were recorded, transcribed, and analyzed using Seidman's guide. In the first stage of the analysis, Seidman suggests reading the transcript multiple times and marking off what it is meaningful, erring on the side of inclusion.[176] He cautions against overthinking at this stage in the process.[177] Once this is completed, Seidman suggests that the researcher become familiar with the data by making thematic connections and creating condensed summaries.[178] I made thematic connections by copying marked-off transcript passages to a post-it note, with a different colour for each interviewee. I rearranged these post-its so that similar comments would be located in proximity to one another. I also created condensed summaries of the interview, reducing the transcript to approximately half of its original size. This had the effect of making me scrutinize the words very closely to ensure that I did not lose their meaning. These condensed summaries were sent to interviewees, along with a Renewal of Consent form. All summaries were confirmed as accurate.

Seidman notes that once summaries and thematic analyses are completed, the researcher's mindset can shift from becoming familiar with the data, to explicitly interpreting the data. To facilitate this stage in the research process, Seidman recommends that researchers ask themselves a series of questions, beginning with the most general: what did I learn in the process of conducting and analyzing interviews?[179] For me, this question became relevant after

[176] Ibid., 120.
[177] Ibid., 121.
[178] Ibid., 121-127.
[179] Ibid., 130.

I had written up the results, and reconsidered them in light of the works discussed in the literature review.

3.5.7 Summary

In this chapter, I outlined the various methods used in the execution of this research project. I described the raw dataset, and detailed the process by which it was rendered into a series of social network visualizations. I also described the interview process, contextualizing it in the theoretical framework of case study and narrative inquiry. In the next section, I will show the outcome of executing these quantitative and qualitative procedures.

4 Presentation of Results

4.1 Overview

This section returns to the four questions that guide this research. These questions are:

1. Are groups of records formed by social network analysis mutually relevant?
2. Do communities formed by social network analysis represent the working groups in an organization?
3. Assuming that employees will explain the social network visualization using narrative, to what extent do these individual narratives converge with one another?
4. What does the application of social network analysis to records tell us about the principle of provenance?

In writing these questions, I made several assumptions, including the assumption that employees would tell contextualizing narratives. Section 4.2 explores whether this assumption was valid. Section 4.3 addresses the first research question, while sections 4.4, 4.5 and 4.6 represent an extended reflection on question two. The third research question is addressed in section 4.7, while section 4.8 highlights some unexpected insights from the interviews. Question four required such in-depth analysis, it was moved to the next chapter, *Discussion of Results.*

4.2 Narratives in the Interview

During the interview, I asked participants to select the visualization which best fit their experience of the organization. I viewed their answers through the lens of narrative inquiry, as outlined in section 3.5.2. According to narrative researcher Phillida Salmon, the process of telling a story involves taking events that seem to be random and disconnected, and connecting them in a way that is understood as meaningful, both for the listener and the teller.[180] Arguably, the social network visualization represents something disconnected: working groups form and dissolve in no apparent pattern. In some cases, interviewees were able to explain these patterns in a meaningful narrative. For example, the node representing Sam was shown as highly connected to other nodes during 2014, less connected in early 2015, and then completely isolated in December 2015. Noticing this, several interviewees explained what was going on, and their narratives are very similar:

> Ryan: So Sam was in a transition period from December 2014 [he was still involved in] onboarding staff. [In] September 2015 he was much more able to step back a bit more. That's not surprising to me that he is sort of on his own little island.
>
> Genevieve: [Looking at Algorithm C with Manager September – December 2015] So Sam technically reports through [sister department], so he is less connected with this department than he previously was. [..] because I had been here for long enough and working more independently, he didn't need to check in as often.
>
> Sam: I think [the visualization] reflects a shift in my role where I don't have as many lines that are connected to me. [This change occurred] mid-March 2015, [where] my role would have had a much more reduced role with the [department name] group.

Arguably, the narrative of Sam's transition to a sister department is a good example of a collectively-constructed narrative, simply because it surfaced so many times during the interviews. This narrative is also a good example of the way participants may assign meaning differently depending on their perspective. For example, Emma, who joined the department mid-2015, did not frame the narrative as one that was about transition, likely because she had not witnessed it. Instead, she simply noted that Sam was disconnected from the core department:

[180] Salmon, "Looking Back on Narrative Research," 78.

> Emma: It makes sense that Sam is on the outside [..] he tends to sit more on the [one] side [of the department]. So it makes sense that he's more of an outsider and the rest of us all sit around the same table.

Having established that interviewees told narratives, and that in some cases, these narratives represent a collective perspective, I next examine employee responses to the aggregates of files.

4.3 Mutually Relevant File Aggregates?

This section aims to answer the first question, which is: *Are groups of records form by social network analysis mutually relevant?* Interviewees viewed the files associated with each working group by viewing an interactive version of the visualization on a laptop. The interactive version displayed a single four-month period on the screen, and was enabled so the coloured shape representing a working group could be clicked to reveal a list of files produced by members of the working group (see *figure 3—6*).

At this point in the interview, participants almost always noted that the files did not belong together:

> Sam: [Regarding the folders associated with May to August 2015] They don’t necessarily all belong together at the very top level. I can see sub-folders for some of these items.
>
> Emma: I would say it represents quite a [..] large grouping of all of our projects.
>
> Ryan: So I would say it makes sense on the [visualization showing working groups], but the files themselves I would not group as all together.
>
> Chase: Now that I understand the information [looking at files] I might change my previous answer. [...] Algorithm C Without Manager for me is very good representation of who I might have interacted with. But when you get into it in further detail, it doesn't necessarily represent just the work that I was doing. So it's a bit general. Once we get into the detail, it almost seems a little bit too general.

In other words, the visualization grouped people in a meaningful way, but the files these people produced did not pertain to a particular project or task. Interviewees may have expected that the file aggregate would contain only the files shared amongst group members, rather than the self-modified files of individuals within the group. Indeed, Genevieve states this expectation quite explicitly:

> Genevieve: [.. but there's a] mix of the specific program names, this [program name] is very specific to one program area, whereas the rest [of the files] are more overarching. What I would actually expect from this - having everyone in the department in the circle - is to see the overarching files, not so much the singular specific files.

Raymond was the sole interviewee who did think the groups of records were mutually relevant, but he was looking at an aggregate produced by only one person: himself. Therefore, there was less potential for a mix of projects. As he comments:

> Raymond: [Looking at visualization of Algorithm B (Edge-betweenness) from May 2014 to August 2015] Looking at the files it's the waivers [..] So yes, it's right, and it is how I remember it.

Raymond also noted that there would be peaks and valleys in SharePoint usage, and believed there would be spikes in April and December corresponding to group retreats. I created a bar chart showing SharePoint usage over time (*figure 4—1* and *4—2*). One bar chart visualized the complete dataset, while the second is based on a dataset with self-modified files removed. In the second chart, the April and December spikes are more prominent. This suggests that Raymond perceived the dataset informing the social network visualization as one that excluded self-modified files.

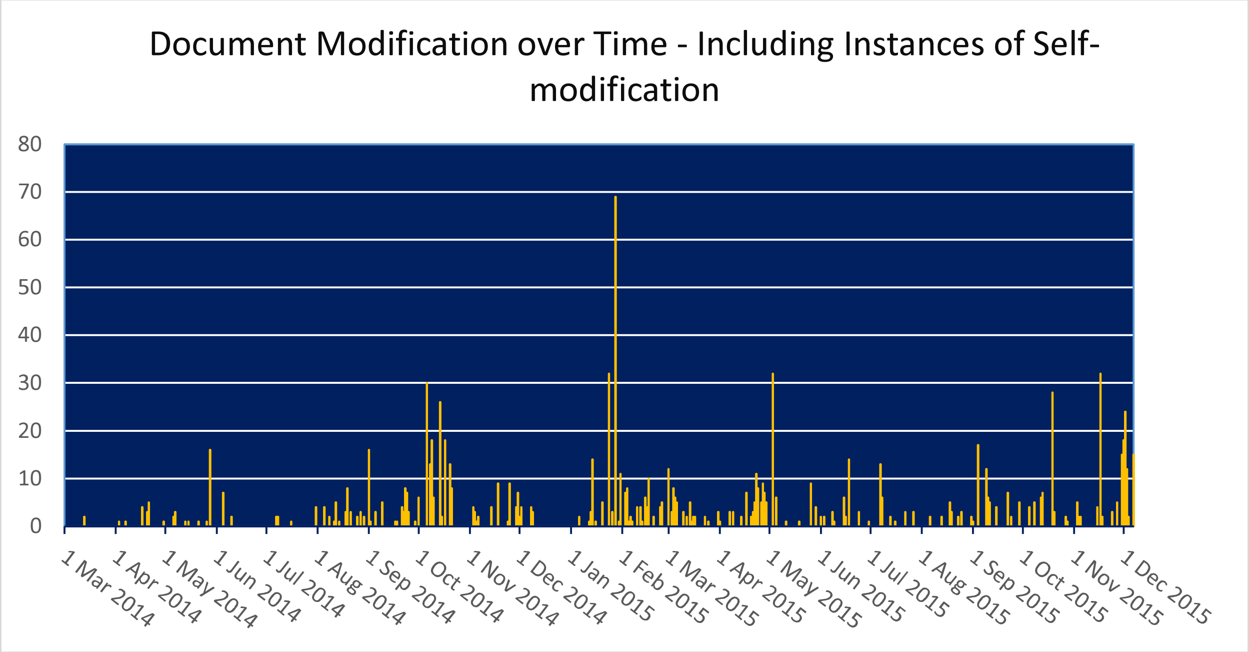


Figure 4—1 Document Modification over Time – Including Instances of Self-modification

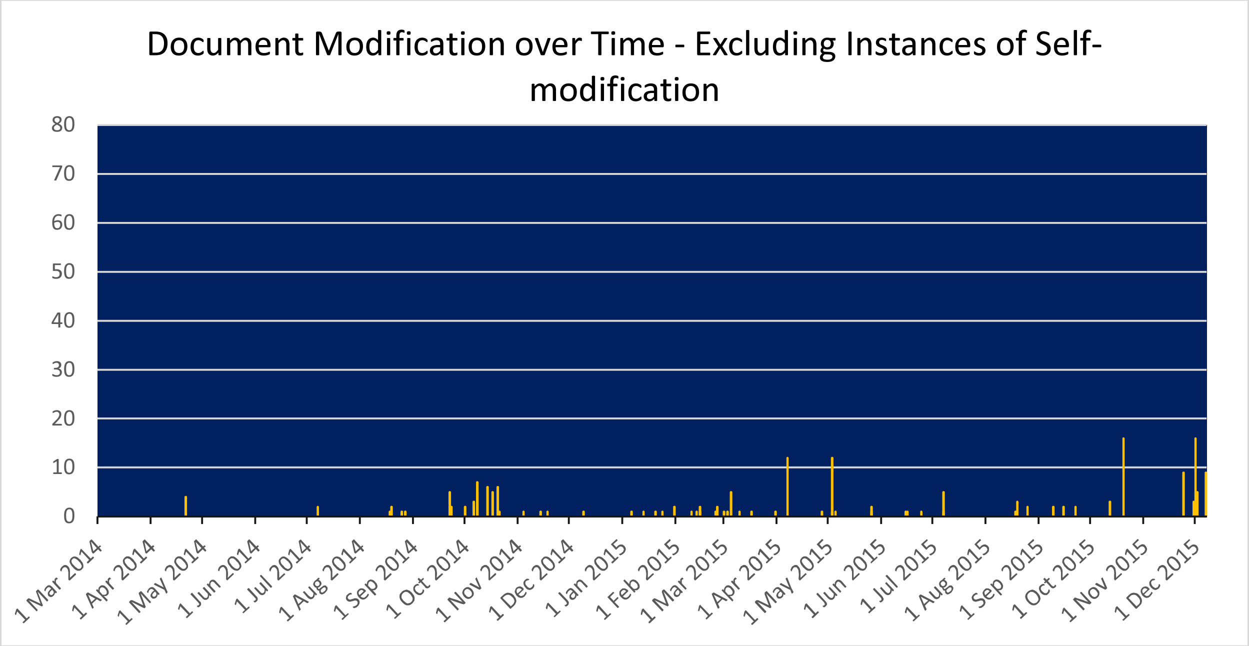


Figure 4—2 Document modification over time – Excluding Instances of Self-modification

4.4 Persistent Relationships versus Activity

This section aims to answer the second question, which is: *Do communities formed by social network analysis represent the working groups in an organization?* To remind the reader, I presented four sets of visualizations (*figure 3—2*) to staff and asked them to pick the one that best represented the pattern of record modification that involved them. For the manager, I modified the question slightly and asked which visualization best represented the department, because I believed he had more information than the others with regards to the overall patterns of file modification in the department. Results are shown in *figure 4—3*.

Algorithm	Tally
A	-
B	II
C (with manager)	III
C (without manager)	I

Figure 4—3 Tally of Algorithm Selections

Noticeably, the department splits along two lines, with approximately one third picking Algorithm B (Edge-betweenness), and two-thirds picking Algorithm C (InfoMap). What could account for this difference? The two people picking Algorithm B (Edge-betweenness) seemed to be focused on a representation of the department where the emphasis was on the persistent working relationships. The four who picked Algorithm C (InfoMap) emphasized the activity that brought together the working group.

Admittedly, the difference in perception may have arisen due to an ambiguity in the question. Those who read my interview guide (see Appendix C) will see that I asked staff to identify the visualization which was most representative with regards to its pattern of document modification, which may have suggested that I was not particularly concerned with working groups based on persistent relationships. However, this question was immediately preceded by a short explanation for the inclusion of Algorithm C (without manager), which alluded to the concept of working groups based on persistent relationships. I said:

> You'll notice that the last visualization is a repeat of Algorithm C, but with the manager removed. I did this because one of my research questions is about managers. I hypothesize that managers modify so many files they make it hard for the algorithm to identify groups of frequently-interacting employees.
>
> **In your opinion, which of these four visualizations do you think best reflects the pattern of file modifications that involved you?**

In other words, I simultaneously implied that the visualization represents both working groups as well as patterns of document modification. I believe this caused interviewees to try and think about the visualization in both ways:

> Raymond: I would say it's Algorithm B [..] I'm looking at this from my perspective, and I'm trying to think of when I use SharePoint, and how I use SharePoint, and I certainly wouldn't consider myself in such a close proximity to Joel, or Chase, or Genevieve.

Raymond prefaced his answer by saying that he was trying to think when and how he used SharePoint (a reference to an activity in the organization), but his ultimate justification for picking Algorithm B (Edge-betweenness) was its representation of working groups. For the manager, this question was especially ambiguous, as I deviated from the script, and asked him which visualization best represents the organization (as opposed to asking about patterns of document modification). "Represents" is an ambiguous term, as a department can be represented in a variety of ways. This is his response:

> Ryan: It's interesting because the challenge with these [Algorithm C and D] it says everyone is together, and I don't think Algorithm C is reflective of the way that we

> collaborate. It doesn't actually illustrate too much. To me, A and B are a little bit more reflective.

Noticeably, “everyone is together” indicates a working group, but “the way we collaborate” suggests activity. At a later point in the interview, the manager notes that Algorithm B (Edge-betweenness) is most representative because it showed an increased use of SharePoint after January 2015, a justification related to the activity taking place in the department. Interviewees who picked Algorithm C (InfoMap) similarly revealed two perspectives: they strongly justified their choice in terms of patterns of file modification, but later in the interview felt compelled to identify persistent working groups. The identification of these working groups is described in the next section.

4.5 Groups based on Persistent Working Relationships

As the interview guide makes clear, I did not ask interviewees directly for the names of people who persistently work together. Rather, in the process of setting me straight on the four-month visualizations, interviewees mentioned groups that generally worked together. One working group was so distinct that it had its own name.

There was clear consensus on the named working group:

> Emma: I would say [there is an] [name of group] - so that would be Raymond, Zachary, myself - we tend to work on things a lot more intensely.
>
> Ryan: Raymond, Zachary, Emma and another person - they aren't involved in the study – they are the [group name] team.
>
> Genevieve: Emma, Zachary, Raymond and Tim, are the [name of group], so they are working together. I wasn't surprised to see these guys [on a particular visualization].

> Raymond: So Zachary, for example is day-to-day with me, every day, same with Emma and her counterpart Tim.

There was also consensus on another group, although this one did not have a name:

> Emma: Joel and Ashley tend to work quite closely together as well, and sometimes you can add Nicole in there as well.
>
> Ryan: So to me Ashley, Nicole, Joel, and Ryan would be a group that I would not be surprised with [generally – not specific to any time].

A third working group also surfaced:

> Emma: Sam and Genevieve work very closely together so that surprises me that they don't share a connection there - because Genevieve reports to Sam structural-wise.
>
> Interviewer: So you said you work with Sam a lot?
> Genevieve: Yes, he's my manager.
>
> Sam: Genevieve is now starting to take over my previous role, and she is leading a lot of the collection of information and data.

I visualized these findings in *figure 4—4*.

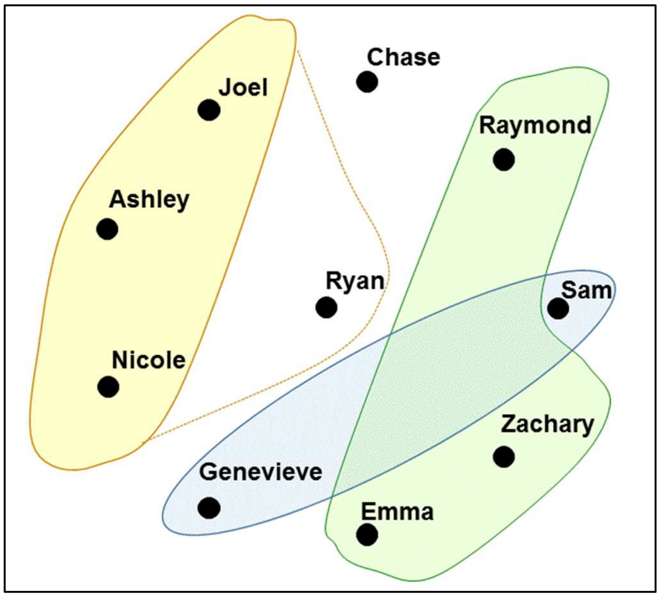


Figure 4—4 Groups Identified by Interviewees

There were also comments that complicated the identification of these working groups, namely, that between Genevieve, Ashley, and Ryan.

> Raymond: Genevieve and Ashley working in May and August does not surprise me because they would be working on a publication [..]. [They also work together to create documents] in training their staff, getting protocols, opening, closing, that sort of thing. [These documents] would be shared on SharePoint.
>
> Genevieve: So any projects that I would work on that, would connect to promoting the [program] would connect with Ashley and also with Ryan.
>
> Raymond: Ryan his line would go to Chase and his working relationships would be with Nicole, and with Genevieve, and with Ashley. But you won't see it with Zachary, Emma. [There would be] a little bit with me, and a little bit with Sam.

I visualized these alternate working groups in *figure 4—5*.

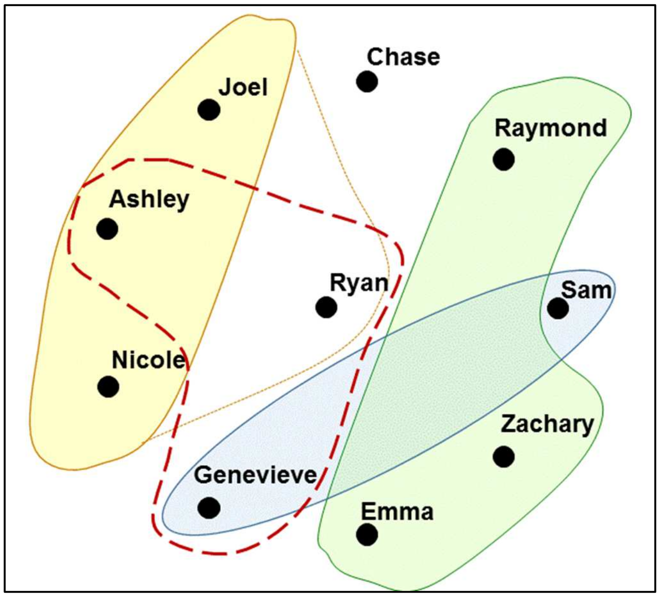


Figure 4—5 Additional Working Groups Identified in Interviews

Algorithm B (Edge-betweenness), applied to the data in the 2015 September-December timeframe (*figure 4—6*), is fairly similar to the interviewee-identified working groups (*figure 4—4*). Sam and Genevieve, noticeably, do not appear as a community in *figure 4—4*. There may be several reasons for this. Genevieve indicated that she prefers to communicate to programmers by email, rather than SharePoint, suggesting she may also be communicating to her manager via email.[181] Indeed, she mentioned that most of the files she worked on mutually with Sam were stored on her computer. She also mentioned that she was working more independently during 2015, which might explain why there are no lines between her and Sam during the September to December 2015 timeframe.

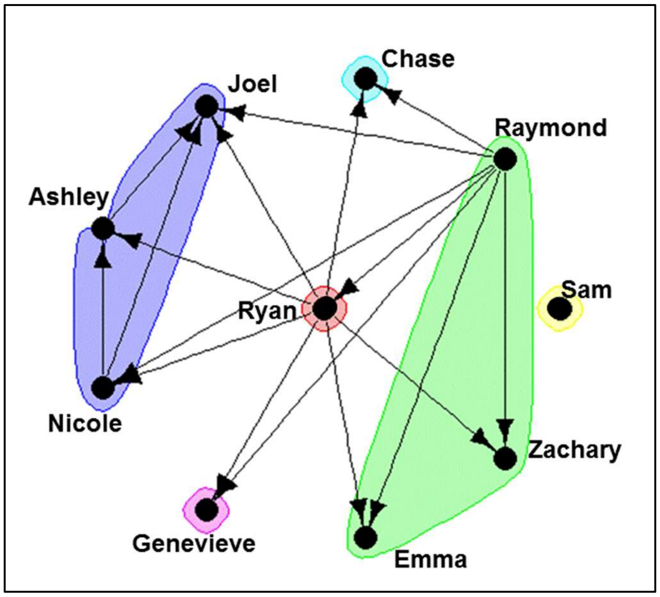


Figure 4—6 Edge-betweenness Algorithm September-December 2015

Given this, Edge-betweenness seems to be the best algorithm for identifying working groups as perceived by the participants in this study. In the next section, I discuss Algorithm C (InfoMap), which elicited narratives about groups based on projects.

[181] A discussion on the extent to which the SharePoint records represents "the whole of records" can be found in section 5.2.

4.6 Groups based on Activity

Generally, people who picked Algorithm C (InfoMap) were very clear in their choice and their rationale. That is, they thought that Algorithm C (InfoMap) represented the widespread sharing of documents on SharePoint:

> Emma: I use SharePoint to share documents. I use it to share with everyone else, so everyone can modify it. Very rarely do I ever upload something that would be just mine. So I would say [this visualization fits because] it's one document we're all editing at some point.
>
> Genevieve: [That's] how I use SharePoint. I only put files on there that I need to share with someone else. So any of my individual files I would just keep on our server off SharePoint or on my computer.

People also arrived at a strong consensus with regards to the particularities of the document-sharing activities. For example, there was a product - referred to alternatively as the guide, publication, and brochure - that all participants identified as the reason that people were brought together in the visualization of Algorithm C (InfoMap):

> Chase: [Regarding the visualization from September to December 2014] We were all creating a huge brochure that gets printed so everyone is communicating [..] their programs or their events, whatever the case might be. Everyone is communicating with Sam 'this is my program, this is everything that I need to include.' It's all within one document, and we were constantly adding more and more and more to it.
>
> Ryan: I know exactly what project we are talking about there [indicating Algorithm A (Fast Greedy) September to December 2014]: [for Chase and Sam] it is likely the guide.
>
> Sam: [Regarding the visualization from September to December in 2014.] - So the big projects in this time period would be our Spring-Summer publications. So those Spring Summer guides I was mentioning, so they [the programmers] all needed to submit content. We use SharePoint to collect that content.

This brochure is also created in the spring time period:

> Genevieve: [Looking at Algorithm C with Manager January to August 2015, indicates that it makes sense] It probably would have been when we do the Fall Guide production in May/June.

In sum, people often used both the general activity of the organization and specific activities to make sense of the visualizations.

4.7 Narrative Convergence?

The previous three sections (4.4, 4.5, and 4.6) explored question three, which is: *Assuming that employees will explain the social network visualization using narrative, to what extent do these individual narratives converge with one another?* On viewing the social network visualization, staff made sense of the data in one of two ways. In one way, the visualization was explained using groups with persistent relationships. In a second way, the visualization was explained using groups in pursuit of a joint activity. Apart from this split, narratives were remarkably consistent with one another. That is, staff were in strong agreement with regards to the constituents of the persistent working relationships; they were also in strong agreement with regards to the activities performed by the activity-based groups.

4.8 Folder Structure as Narrative

During interviews, comments emerged that shed an unexpected light on the concepts of narratives and records. This section discusses these unexpected discoveries.

When I approached this project, I believed that the social network analysis would present to staff disparate pieces of information, from which staff would be able to derive a meaningful narrative. However, I could not quite imagine how a narrative was expressed in the records, despite claims by Horsman[182] and Hurley[183] that this is the case. Therefore, it was very interesting to me when a participant explicitly asserted that the aggregate of records told a story:

> Chase: I always think that money speaks for itself; the movement of money speaks for itself. It tells you what was happening. It tells you what purchases were made, what was brought in. [..] So I think the general ledger for any financial activity and how it's currently organized - and probably the university keeps their financial information - tells its own story. [..]

Interestingly, the interviewee seems to suggest that the storytelling capacity of a group of records is especially true of financial files. It made sense to me to discriminate and to say that some records tell stories while others do not. My conviction on this point stems in part from my past experience doing archival work, where records were so disordered that no clear narrative emerged. Riessman argues that incoherence, often associated with narratives about trauma or illness, do not disqualify such narratives as stories.[184] Similarly, it could be said that seriously disordered archives are still narratives despite their incoherence. The disorderliness of the SharePoint sites in the department under study was noted during interview:

> Interviewer: If someone new came into the organization, do you think that with a little bit of study they could figure things out?
> Sam: In this SharePoint, no, it's a mess.

Speaking to this disorder, Sam also suggested that a holistic narrative does not emerge because there are several conflicting narratives. The construction of a folder structure is a narrative

[182] Peter Horsman notes that "If any principle should govern archival theory, it is not the fonds, but rather the visualisation through description of functional structures, both internal and external: archival *narratives* about those multiple relationships of creation and use so that researchers may truly understand records from the past." (Italics mine). Horsman, *Last Dance of the Phoenix,* 22-23.
[183] Hurley, "Parallel Provenance (1)," 110.
[184] Riessman, "Looking Back on Narrative Research," 82.

because it involves connecting various pieces of information in the form of files. People undertake this task differently because they have different perspectives. In Sam’s words:

> Sam: [The current files] are grouped by [..] the creator, and how they process information, and how they want to see it organized. [In SharePoint there wasn’t] a pre-set file structure, or even segregated by areas. It was just each individual person who's going in there and creating [saying] ‘this makes sense to me, [this is] how I access information.’ [They] may not necessarily make sense to the other users or modifiers.

In other words, records which are allowed to grow organically represent a variety of narratives, informed by different perspectives. This point was reinforced by another interviewee, Chase:

> Chase: When I started the files didn't follow any organization system within [name of program], they were just very, they were all over the place, there was a file that said like ‘[name of folder] files,’ and you would think, well what is in [name of folder] files, that doesn't tell me what is in there.

What Chase makes clear is that someone named the files in a way that made sense to him or her, but did not necessarily make sense to others, indicating a difference of perspective. It should be noted of course, that some names and folder structures can be more readily understood by a variety of people. But even in such cases, it is inevitable that they express a perspective. Chase emphasized this when he described his efforts to organize the folder structure within the Shared Network Drive used by himself and his Co-op students:

> Chase: It makes sense in my mind how things fall into seasons [..my folder structure] has made it easy for me.

It could be argued that a well-functioning folder structure represents a narrative understood by all members of the department. As discussed in section 2.6, people of Western cultures tend to structure their narratives with an allusion to time.[185] From this, it would follow that time-based narratives are more likely to be understood in a Western context. Indeed, two people referenced well-functioning file systems within the department, and noticeably, both systems had a temporal element. Emma, commenting that it was easy to find records in the hard drive noted:

> Emma: We also have our hard drive files and that's organized by area project and then year.

Likewise, Chase commented on the Shared Network Drives:

> Chase: As a general statement, I think [department name] did a pretty good job with how we organized our files and generally speaking, we group our files based upon three seasons: Fall, Winter and the Spring/Summer. And that's sort of the starting point for pretty much everything, and then subfiles of course fall into that.

4.9 Summary

This chapter presented the results of the qualitative interviews. It showed that employees did not believe aggregates brought together by social network analysis were mutually relevant, which was likely due to the presence of self-modified files. A future study which excludes self-modified files would resolve this uncertainty in the results. Despite the absence of mutual relevance in the files, there were signs that social network analysis could be used as a means of bringing together mutually relevant records, as staff noticed that some groups of people identified by the algorithm were representative of the organization. Significantly, there was a

[185] Scholars of narrative inquiry note that alternatives to temporal sequences are possible: "non-Indo-European stories may be structured so that later actions, states or events precede earlier ones. In addition, some narrative traditions organize stories around place, or around the hierarchy of ranks of the characters or their relationship to the speaker, rather than around time." Patterson, "Narratives of Events," 31.

split in how staff conceived working groups. Some characterized a working group as one comprised of persistent working relationships, while others characterized a working group as one involved in the pursuit of a joint task. In the next chapter, the results of the qualitative interviews will be analyzed in greater detail.

5 Discussion of Results

5.1 Boundaries, Autonomy, and the Fonds

In "Questioning Autonomy," Jenny Bunn argues that a community is autonomous when it is distinct from its environment, and that this separation occurs when the community engages in an act of "being and doing."[186] Initially, I was unclear what it means to say that a community engages in "being" and "doing." I realized, however, that when results were split between Algorithm B (Edge-betweenness) and Algorithm C (InfoMap), that there are two different ways to conceptualize a group of people: either they can "be" together, in the sense that they are widely regarded as having persistent working relationships, or, they can "do" together, which means they are focused on a particular project or task. In this case, it could be argued that there were no autonomous subcommunities represented by the records in this study, because groups recognized as persistent working groups ("being"), and groups recognized as activity-based groups ("doing") did not overlap.

However, this study did reveal an autonomous community in terms of the department as a whole.[187] Sam was previously both a member of the department, and a contributor to the department's activity. When he moved to another department, he effectively breached the boundary of the department. Interviewees were clear about what was happening with regards to Sam's transition, but less clear with regards to the working groups within the organization.

[186] Bunn, "Questioning Autonomy," 10.

[187] As noted in section 2.3, Bunn's understanding of autonomy involves a community making itself distinct from its environment. Bunn contrasts her understanding with the views put forward by Eastwood, who defines the locus of autonomy in the department-creating power of the larger organization. The department under study was clearly formed by a higher power in the university, but this is not what makes it autonomous, according to Bunn.

This relatively strong consensus with regards to Sam's transition indicates that the department itself is distinct from its environment, suggesting an autonomous community.

It could be argued that traditional theorists have grappled with the concept of "being and doing," but with different terms. I argue that "function" is a reference to doing, and "administrative office" is a reference to being. These terms are closely associated with one another, and with the idea of an autonomous community. For example, Muller, Feith, and Fruin believe that *function* represents a specific activity, such as auditing,[188] and that a host of functions are associated with an administrative office.[189] Further, such offices are viewed as "independent branches,"[190] phrasing that evokes autonomy. Similarly, Schellenberg considered *function* to be an umbrella term covering the activities enacted by a department.[191] He notes that it is only possible to group records by department when the organization is stable and the functions are well-defined,[192] suggesting he was aware that autonomous communities are necessary for records management, and that such communities are defined by their "being and doing." In other words, Bunn's "being and doing" has deep roots in the archival discipline, even if theorists did not use these particular terms.

Interviewee responses are consistent with the nature of the algorithms. Algorithm B (Edge-betweenness) operates on the assumption that the communities are the product of a static set of relationships, which is why it focuses on finding the node pair that connects one community to the other. By contrast, Algorithm C (InfoMap) operates on the assumption that the

[188] Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 25.
[189] Ibid., 19.
[190] Ibid., 58.
[191] Schellenberg, *Modern Archives,* 53.
[192] Ibid., 59.

community under representation is the product of dynamic interactions amongst constituent nodes. It works by sending a signal which is redirected from node to node, with communities being sets of nodes that frequently redirect the signal to one another (see section 3.4.4). Therefore, it is not surprising to me that staff thought Algorithm B (Edge-betweenness) captured persistent working groups, while Algorithm C (InfoMap) better represented document modification activity.

Arguably, distinct communities produce distinct groups of records, and distinct groups of records can be used to represent the organization. As noted in the previous paragraph, Algorithm B (Edge-betweenness) appears to produce groups reflecting persistent working relationships, while Algorithm C (InfoMap) recognizes groups of people who jointly pursue an activity. If both algorithms identify the same group of people, it is likely a sign that the group represents an autonomous community. In short, records managers in search of such communities would do well to apply both Algorithm B (Edge-betweenness) and Algorithm C (InfoMap) to an aggregate of records.

5.2 Returning to Horsman: Mutually Relevant Records

In his article "The Last Dance of the Phoenix," Peter Horsman suggests that archivists, wanting tidy and well-defined aggregates, have perpetuated the concept of *respect des fonds* to better manage their holdings. He challenges the notion that *respect des fonds* protects the provenance of the records, with provenance defined as a "conceptual whole based on the functioning of business processes."[193] In his view, this conceptual whole rarely exists in

[193] Horsman, "Last Dance of the Phoenix," 22.

practice: records from different regional offices are brought together only in the archives, and up to 98% of the records in a fonds may be destroyed by the archivist in the act of appraisal. Given this, Horsman argues that archivists need not group records, and that their task concerns the description of records. In his view, description affords archivists the flexibility to represent the records accurately.[194]

I agree with Horsman that collocating records into a whole that does not exist during the active life of the records can distort the representation of the records. However, in making his point he gives examples of records that could belong to an autonomous community. The example of records from different regional offices could very well represent a community: in an era of easy communication there could be working relationships and shared activity that transcend geographically dispersed workplaces. Similarly, the partial fonds that remains after significant destruction may still belong together by virtue of being the product of an autonomous community.

It may be that the crucial criterion by which records can be said to represent a community, in Horsman's view, is that they need to be complete in some sense; that is, they need to be a "conceptual whole" of business processes. This notion of a whole is a poor criterion for a fonds, a realization that came to me when one of the interviewees attributed the poor representation of the social network visualization to low level of SharePoint usage. He clarified that the department used a variety of platforms to store and share documents, including work computers, Shared Network Drives, SharePoint and email. In reflecting on Horsman's claim

[194] Ibid.

that the records represent a conceptual whole, I realized that the business processes of the organization, in addition to being dispersed across various platforms, are often conducted verbally; records, therefore, are always a partial representation. With this in mind, it does not make sense to posit "completeness" as a criterion for an archival fonds.

By contrast, it makes more sense to argue that records should be kept together because they facilitate the mission of those who search for evidence. One might achieve this objective by keeping together the records of an autonomous community, effectively keeping together various narratives pertaining to the same events. When these narratives align with one another, they reinforce any given claim that is made about the archives. The idea that a claim becomes "true" when it makes sense with other claims in the fonds is a well-regarded strategy for justifying general knowledge claims, and is rooted in the philosophy of W. V. Quine. When these narratives conflict, they enable a researcher to access the multiple perspectives represented by the fonds. Researchers appreciate having access to these multiple perspectives: in her explorations through the documentary evidence of the Soweto uprising, Helena Pohlandt-McCormick shows that exploring multiple realities gives one a nuanced understanding of a historical event.[195]

Horsman's belief that arrangement is a poor representation of a fonds makes sense in some cases, but not so much in others. His belief makes sense with regards to the records of a department that does not contain subcommunities; in these cases, there are no community-based aggregates in need of representation. However, Horsman's claim that *respect des fonds*

[195] Pohlandt-McCormick, "In Good Hands," 313.

has no theoretical value is questionable. Communities reflect real relationships and activities. Records associated with these communities are mutually relevant due to their shared context. To intermix the records of various communities both obscures the representation of something that actually exists, and hinders research by making it difficult to cross-reference relevant information.

Part of the reason it makes so little sense for the archivist to impose groups within a fonds is because employees may have already attempted to construct a folder structure. As noted in section 4.8, this meaning-making activity can be understood as a subjective narrative, which is worth preserving for the light it sheds on staff perspective. At the same time, I acknowledge that this folder structure may be difficult to understand. The expression of multiple narratives may result in recordkeeping chaos, as each employee connects the same set of records in different ways.

When such seriously disordered records arrive at the archives, archivists are instructed to impose their own folder structure. But if imposing a folder structure is a subjective meaning-making activity, it means that the folder structure produced by the archivist also reflects a subjective perspective – that is, the viewpoint of the archivist. Initially, it was puzzling to me that archivists are instructed to write their finding aids in an "objective," manner given the inherent subjectivity of their work. How to make sense of this? Hans Booms argued that archivists who adopt the mindset of the people who lived during the time and place when the records were created will produce finding aids which are more representative of the records.[196]

[196] Booms, "Uberlieferungsbildung: Keeping Archives as a Social and Political Activity," 28.

By extension, archivists arranging current records will achieve better results if they strive to be consistent with the mindset of the current era.

I suspect that archivists operating under these Boomsian assumptions have labeled their efforts in this regard a striving for "objectivity." I came to this realization during an assignment to arrange and describe the personal papers of a Vancouver politician who participated extensively in civic organizations, which involved attending meetings on a frequent basis. Personally, I avoid large meetings because I have a severe-profound hearing loss. When writing up the finding aid of this local politician, I was tempted to start the biographical sketch by saying that she was a person with full hearing, as this seemed relevant to the life she had chosen. Of course, I did not, but I realized I was not being objective in omitting this true fact; I was being mainstream.[197]

At the time of the assignment, I decided it made sense to adopt the assumptions of the mainstream if one's aim is to communicate the finding aid to the public. The majority have this mainstream perspective, and minorities are well familiar with it, so often does it erase their own. Since writing the politician's finding aid in 2013, I have come to see that there can be an alternative to Booms's view that the skill of the archivist is to adopt the mainstream perspective. In this alternative view, the archivist instead brings out the various voices of the creators of the records.[198] To me, this approach is consistent with the spirit – if not the letter –

[197] The word mainstream means "normal or conventional ideas, attitudes or activities." *Oxford English Dictionary*, 10th ed., s.v. "mainstream."

[198] It was only writing the final draft of this thesis that I realized the extent to which I had been influenced by the ideas of Terry Cook. Cook believes that in conducting appraisal the archivist needs to be sensitive for the voice of the marginalized, much in the way I have argued for a description practice that recognizes the voice of the marginalized. See Cook, "Mind over Matter," 57.

of traditional archival theory, which aims to achieve an accurate representation of the circumstances that gave rise to the archives.[199] Speaking personally, I would prefer to be the archivist who disentangles various perspectives from an archival fonds rather than the archivist who obscures these perspectives with the voice of the mainstream. The work of the former seems more challenging, and more just.[200] [201]

[199] In this paper, I focus specifically on the way that various members of an organization impact the records by imposing a folder structure. Jennifer Douglas explores the impact of individuals on the records more broadly, noting that creators, custodians, and archivists each affect the records. She specifically notes that glossing over this impact introduces distortion in the representation of a fonds. See Douglas, "Honest Description," [forthcoming].

[200] It should be noted that documenting the perspectives within a record aggregate may make it difficult to standardized the description of finding aids. I think this issue is significant and requires more thought, perhaps in another paper. I will point out that the names of creators, the roles they played in the organization, and a short paragraph expressing their perspectives are categories that can be repeated across finding aids.

[201] Additionally, it could be said that this approach demands more work for the archivist, as it is more work to express multiple perspectives rather than a mainstream perspective. Again, this is an interesting point that merits additional discussion. As an initial thought, I note it is difficult to measure the "work" required to put oneself in the shoes of others, and to compare this to the "work" of conforming to the mainstream.

6 Conclusions

6.1 Summary

According to Heather MacNeil, "we need more ways of making sense of the world, not fewer."[202] This thesis explored several ways of making sense of an archives, running the gamut of metaphors from interrelated webs to systems of logic. These attempts to view the archives differently were borne out of a frustration with the gaps and inconsistencies of traditional archival theory. For example, Muller, Feith, and Fruin assert that records should be aggregated by department on the rationale that doing so assists the users.[203] This is contradicted by a later statement on the same page that archival arrangement should not cater to the needs of users.[204] Eastwood attempts to pinpoint the nature of the archives with the claim that its member records reflect a common function. He admits that function is one of those concepts we do not understand.[205]

For this reason, I felt compelled to start from scratch by applying metaphors to the concept of the archives. Ultimately, the metaphor which best accounted for the claims of traditional archival theory was the idea that the archival fonds is akin to a case study. A case study, like an archival fonds, is considered a means of acquiring knowledge. Also like an archival fonds, a case study posits a boundary around the scope of the case, and explicitly justifies this boundary in terms of evidence – that is, by bringing together various narratives pertaining to the same

[202] MacNeil, "Trusting Records in a Postmodern World," 45.
[203] Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 54.
[204] Ibid.
[205] Eastwood, "What is Archival Theory," 128.

events, one is able to obtain a deeper understanding of the circumstances to which they refer. To adopt this metaphor and its clear rationale for provenance means acknowledging that the archives is a social construct comprised of subjective narratives.

This socially constructed view of the archives made it possible to interpret the responses of the interviewees in this study. To remind the reader, interviewees were shown a set of four visualizations (see section 2.5.9), each of which represented the results of a social network analysis algorithm. More specifically, each of the four visualizations were comprised of six snapshots showing the department's groups of closely-interacting individuals during a four-month period. These groups were determined by applying the social network algorithm to the creator and modifier metadata extracted from the documents on the department's SharePoint site.

Interviewee responses were split between Algorithm B (Edge-betweenness) and Algorithm C (InfoMap). Individuals selecting the same algorithm offered similar rationales: those choosing Algorithm B (Edge-betweenness) justified their choice by saying that it represented the persistent working relationships within the department, while selectors of Algorithm C (InfoMap) justified their choice by saying that it represented the document-sharing activity that took place within the organization.

To me, these rationales were strongly reminiscent of a dualism in traditional archival theory. This dualism is the concept of an administrative office (which identifies a set of persistent working relationships) and the concept of a function (which identifies the activity jointly pursued by members of that working group). This dualism also surfaces in the work of Jenny

Bunn, who argued that a community simultaneously engaging in being and doing makes itself distinct from its environment. When archivists make it a rule to keep the records of a distinct community together, what they really mean is that one should keep together those records stemming from a coincidence of group membership and shared activity.

In my view, Muller, Feith, and Fruin's traditional conceptualization of provenance omitted this crucial piece of information. Had they recognized this omission, they may have foreseen that their conceptualization of provenance would be irrelevant in those cases where functions are frequently reassigned from one administrative office to another. It might be unfair to expect 19th century archivists to analyze a problem that only became clear with the advent of modern administrative practices, but it should be noted that departments did exchange functions with one another during Muller's time.[206] Additionally, the *Manual* recognizes that it may be unclear what archivists should do when the function associated with a record aggregate is passed from one administrative body to another.[207] This suggests that an administrative subgroup is complex, and worth exploring in greater detail.

For Muller, there were other signs that an arrangement based on administrative structure might not serve as a complete representation of an organization. Muller highly respected the work of a colleague, Van Riemsdijk, who located provenance in physical accumulations.[208] Van Riemsdijk, sought to explain these aggregates using information from a variety of sources. This information not only included the administrative structure, but also included "the record-

[206] Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 19.
[207] Ibid., 24.
[208] Horsman, *Rise of the Phoenix,* 9.

creating process"[209] – phrasing that clearly suggests activity. According to Ketelaar, Van Riemsdijk specifically means activity in the form of business processes and workflows.[210] Had Muller given serious consideration to Van Riemsdijk's approach, he might have realized that emphasizing "being" in the form of administrative structure was a limited means of representing the organization.

Muller, Feith, and Fruin suggest that within the fonds of a department there will be "numerous dossiers,"[211] and that the best approach to arrangement is to keep these dossiers intact. They further suggest that dossiers are mutually relevant, even when this doesn't appear to be the case.[212] The modern equivalent of a dossier is a folder, and as this study showed, it is not necessarily the case that folders are mutually relevant. As one interviewee explained, disorderliness results when multiple people group records in a way that is meaningful to them, but not necessarily to anyone else.

Another interviewee noted that records tell a narrative. As narratives are inherently subjective, and involve bringing together discrete pieces of information, this strongly suggests that a collaboratively-constructed folder structure may represents a confluence of subjective narratives. The archivist who receives this aggregate and attempts to rectify disorder by

[209] Ketelaar, "Archival Theory and the Dutch Manual," 58.
[210] Ibid.
[211] Muller, Feith, and Fruin, *Manual for Arrangement and Description,* 50. To be clear, a dossier is "the aggregation of all the records that participate in the same affair or relate to the same event, person, place, project, or other subject." *The InterPARES 2 Project Terminology Database*, s.v. "Dossier," accessed April 13, 2016, http://interpares.org/ip2/ip2_terminology_db.cfm.
[212] Ibid.

imposing a more coherent folder structure effectively imposes his or her own subjective process of meaning-making on the records.

Noting that some archivists hold the belief they tell the story of the records "objectively," I argued their claim is not valid. Instead, this claim likely reflects an attempt to adopt a standpoint understood by the general public. In adopting a mainstream perspective, archivists inevitably create a competing subjective narrative which may erase the perspectives of the creators. To me this seems inconsistent with the general archival impetus to *respect* the records, which is the heart of the principle of provenance. Proposing that archival work should instead involve the drawing-out of perspectives, I noted that archivists need a tool which is flexible enough to permit the expression of multiple narratives. As Horsman suggests, that tool is description, not arrangement. Arrangement is limited to expressing the statement that a set of records are mutually relevant; description makes clear why alternative groupings for records are considered mutually relevant by their creators.

6.2 Strengths and Limitations of Research

In this section I discuss the strengths and limitations of this study. This thesis demonstrates the benefits of exploring the nature of the archives using metaphor; for example, by conceptualizing the archives as a brain I was led to hypothesize that archives are comprised of record clusters interlinked via the interaction of their creators. This metaphor led to social network analysis, and subsequently the research results, which shed light on the way an archives represents both an autonomous community and the overlapping perspectives expressed by that community. This metaphor-based approach to research was well-framed by

Iser's soft theory, which prioritizes new ways of seeing the object under study, and celebrates metaphor as a means of doing so. Framing the literature review as a conceptual lens made clear the role it played in the research process.

This study was also limited in the number of ways. One problem arose when Algorithm B (Edge-betweenness) was not visualized correctly during interviews. This meant that I had to extract mentions of persistent working relationships from the interviews, and assess the extent to which the corrected visualization represented these working groups. Without a doubt, this assessment would be more reliable had it been undertaken by a member of the organization. A second limitation was that the study involved a large number of self-modified files. As the large number of self-modified files may have been the reason that members of the organization perceived the files as representing a diversity of programs, this detracts from my claim that the mix of files is owing to the absence of autonomous communities.

This study focused on a single, small department, and was further focused on a subset of files within the department. The findings should be understood as context-specific findings, and are not generalizable. By repeating the process in a variety of different organization, and with a variety of record aggregates, generalization may be achieved. As this was an exploratory study, it could be argued that it did not set out to uncover general truths about archives, and is instead an exercise in thinking carefully about the nature of archives. In other words, the primary value of this study is the information it offers to those in pursuit of the same theoretical goals, and specifically, the light it sheds on the notion of provenance.

6.3 Future Research

As with many exploratory studies, this one suggests several possibilities for future research. For example, replicating the study in various contexts may offer new insights with regards to the applicability of social network analysis as a method for creating record aggregates. The results may differ when the record aggregate includes fewer self-modified files, or when the record aggregate more completely approximates the total records used in the organization.

Additionally, future projects might further investigate the idea that the archives is comprised of narratives; questions that might be considered include: when employees collaboratively construct a set of folders, how do they overcome differences in perspective? To what extent do classification strategies, such as classifying by subject or function, help staff arrive at consensus? Likewise, does the social network analysis, with its allusion to time, activity, and working groups, help staff construct a common narrative? If yes, how can this common narrative be translated into a folder structure? The method developed in this work, of visualizing social networks over time alongside the records linking the network may prove useful in future research that seeks to elicit feedback from records creators or records managers.

Significantly new ways of thinking about archives might be possible if we acknowledge that archives are narratives, as it suggests that memory-keeping based on oral traditions might have more in common with paper-based archives than has been previously recognized. Provided that a partnership between members of oral cultures and members of archival institutions is possible, what can be learned from people who explicitly recognize narrative as a means of memory?

Bibliography

Andrews, Molly, Corinne Squire and Maria Tamboukou. “What is Narrative Research?” Introduction to *Doing Narrative Research,* 1-21. Edited by Molly Andrews, Corinne Squire and Maria Tamboukou. London: Sage Publications, 2008.

Bak, Greg. "Continuous Classification: Capturing Dynamic Relationships Among Information Resources." *Archival Science* 12, no. 3 (2012): 287-318.

Barritt, Majorie Rabe. “Coming to America: Dutch *Archivstiek* and American Archival Practice.” Introduction to *Manual for the Arrangement and Description of Archives,* by Samuel Muller, Johan Adriaan Feith, and Robert Fruin, xxxv-l. Reprint, Chicago: The Society of American Archivists, 2003.

Bastian, Mathieu, Sebastien Heymann and Mathieu Jacomy*. Gephi* [Computer software]. Compiègne: Université de Technologie de Compiègne, 2008.

Bender-deMoll, Skye. *ndtv* (Network Dynamic Temporal Visualizations) [Computer software]. Seattle: University of Washington, 2015. http://CRAN.R-project.org/package=ndtv

Booms, Hans. “Uberlieferungsbildung: Keeping Archives as a Social and Political Activity.” *Archivaria* 33 (1991): 25-33.

Bunn, Jenny. “Multiple Narratives, Multiple Views: Observing Archival Description.” Phd diss., University College London, 2011.

Bunn, Jenny. “Questioning Autonomy: An Alternative Perspective on the Principles which Govern Archival Description.” *Archival Science* 14 (2014): 3-15.

Clauset, Aaron, Mark E.J. Newman, and Cristopher Moore. "Finding Community Structure in very Large Networks." *Physical Review E* 70, no. 6 (2004): 1-6.

Cook, Michael. *The Management of Information from Archives*. 2nd ed. Aldershot: Gower Publishing Ltd, 1999.

Cook, Terry. "Mind Over Matter: Towards a New Theory of Archival Appraisal." In *The Archival Imagination: Essays in Honour of Hugh A. Taylor*, edited by Barbara Lazenby Craig, 38-70. Ottawa: Association of Canadian Archivists, 1992.

Cook, Terry. “The Archive(s) Is a Foreign Country: Historians, Archivists, and the Changing Archival Landscape.” *The Canadian Historical Review* 90, no. 3 (2009): 497-534.

Csárdi, Gábor and Tamás Nepusz. *igraph* [Computer software]. Budapest: Eötvös University, 2005. http://igraph.org

Csárdi, Gábor. “Community Structure Detection Based on Edge Betweenness.” In *R igraph manual pages*. Last modified 2015. http://igraph.org/r/doc/cluster_edge_betweenness.html

Denzin, Norman K. and Yvonna S. Lincoln. Introduction to *The SAGE handbook of Qualitative Research,* 1-32*.* 3rd ed. Edited by Norman K. Denzin, and Yvonna S. Lincoln. Thousand Oaks: Sage Publications, 2005.

Diesner, Jana, Terrill L. Frantz, and Kathleen M. Carley. "Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different"." *Computational & Mathematical Organization Theory* 11, no. 3 (2005): 201-228.

Douglas, Jennfer. "Archiving Authors: Rethinking the Analysis and Representation of Personal Archives." PhD diss., University of Toronto, 2013.

Douglas, Jennifer. "Towards More Honest Description." *American Archivist* 79 (Spring/Summer 2016) [forthcoming].

Douglas, Jennifer. "What We Talk About When We Talk About Original Order in Writers' Archives." *Archivaria* 76, no. 1 (2013): 7-25.

Duranti, Luciana. "The Archival Bond." *Archives and Museum Informatics* 11, no. 3-4 (1997): 213-218.

Duranti, Luciana. "The Concept of Appraisal and Archival Theory." *American Archivist* 57 (Spring 1994): 328- 45.

Duranti, Luciana. *Diplomatics: New Uses for an Old Science.* Maryland: Scarecrow Press, 1998.

Easley, David, and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. New York: Cambridge University Press, 2010.

Eastwood, Terry. "What is Archival Theory and Why is it Important?" *Archivaria* 37 (1994): 122-130.

Edelman, Gerald M., and Giulio Tononi. *Consciousness: How Matter Becomes Imagination*. London: Penguin, 2000.

Esteva, María. "The Aleph in the Archive: Appraisal and Preservation of a Natural Electronic Archive." PhD diss., University of Texas at Austin, 2008.

Flyvbjerg, Bent. "Case Study." In *The SAGE Handbook of Qualitative Research, 4th ed.,* edited by Norman K. Denzin, and Yvonna S. Lincoln, 301-316. Thousand Oaks: Sage Publications, 2011.

Fortunato, Santo and Marc Barthélemy. "Resolution Limit in Community Detection." *Proceedings of the National Academy of Sciences* 104, no. 1 (2007): 36-41.

Foscarini, Fiorella. "Understanding Functions: An Organizational Culture Perspective." *Records Management Journal* 22, no. 1 (2012): 20-36.

Gillean, Dan. "The Consequences of Ignoring Records Management: A Personal Reflection on my time with the Government of British Columbia." (Scholarship application to *ARMA International Education Foundation,* June 2011). Accessed June 2, 2015 from http://www.armaedfoundation.org/pdfs/Paper_Gillean_Dan_2011.pdf

Girvan, Michelle, and Mark E.J. Newman. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99, no. 12 (2002): 7821-7826.

Gregory, Steve. "An Algorithm to Find Overlapping Community Structure in Networks." In *11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, edited by Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin and Andrzej Skowron. Berlin: Springer, 2007.

Grigely, Joseph. *Textualterity: Art, Theory, and Textual Criticism*. Ann Arbor: University of Michigan Press, 1995.

Hewson, Claire. “Mixed Methods Research.” In *Sage Dictionary of Social Research Methods*, edited by Victor Jupp, 179-180. London: Sage Publications, 2006

Horsman, Peter. "The Last Dance of the Phoenix or the De-Discovery Of the Archival Fonds." *Archivaria* 1, no. 54 (2002): 1-23.

Hurley, Chris. "Parallel Provenance Part 1: What, if Anything, is Archival Description?” *Archives and Manuscripts* 33, no. 1 (2005): 110-141.

Hurley, Chris. "Parallel Provenance Part 2: When Something is Not Related to Everything else." *Archives and Manuscripts* 33, no. 2 (2005): 52-89.

Hurley, Chris. "Personal Papers and the Treatment of Archival Principles." *Archives and Manuscripts* 6, no. 8 (1977): 351-365.

Iser, Wolfgang. *How to do Theory*. Malden: Blackwell, 2006.

Jenkinson, Hilary. *A Manual of Archive Administration*. Rev. ed. London: Percy Lund, Humphries and Co., 1937. https://archive.org/stream/manualofarchivea00jenkuoft#page/n5/mode/2up

Ketelaar, Eric. “Archival Theory and the Dutch Manual.” In *The Archival Image: Collected Essays*, edited by Yvonne Bos-Rops, 55-66. Hilversum: Verloren, 1997.

Lancichinetti, Andrea and Santo Fortunato. “Community Detection Algorithms: A Comparative Analysis.” *Physical Review E Statistical, Nonlinear, and Soft Matter Physics* 80, no. 5 (2009): 1-11.

Luckhardt, Grant C., and William Bechtel. *How to do Things with Logic*. Hillsdale: Lawrence Erlbaum Associates, 1994.

MacNeil, Heather. "Archivalterity: Rethinking Original Order." *Archivaria* 66 (2008): 1-28.

MacNeil, Heather. "Trusting Records in a Postmodern World." *Archivaria* 51 (2001): 36-47.

Milic-Frayling, Natasa, Marc Smith, Ben Shneiderman, Derek Hansen, Cody Dunne, Eduarda Mendes Rodrigues, Udayan Khourana, Jure Leskovec, Bernie Hogan, Itai Himelboim, Libby Hemphill, Robert Ackland, Scott Golder, Vladimir Barash, and Brian Keegan. *NodeXL* [Computer software]. San Jose: Social Media Research Foundation, 2014.

Moler, Cleve, Jack Little, and Steve Bangert. *MatLab* [Computer software]. Natick: MathWorks, 2016.

Muller, Samuel, Johan Adriaan Feith, and Robert Fruin. *Manual for the Arrangement and Description of Archives.* Reprint, Chicago: The Society of American Archivists, 2003. First published 1898 by the Netherlands Association of Archivists.

Nepusz, Tamás. “What are the Differences Between Community Detection Algorithms in igraph?” Modified February 28, 2012. http://stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph/9478989#9478989

Newman, Mark. *Networks: An Introduction*. Oxford: Oxford University Press, 2010.

Otte, Evelien and Ronald Rousseau. “Social Network Analysis.” *Journal of Information Science* 28, no. 6 (2002): 441–453.

Patterson, Wendy. “Narratives of Events: Labovian Narrative Analysis and its Limitations.” In *Doing Narrative Research*, edited by Molly Andrews, Corinne Squire, and Maria Tamboukou, 22-40. London: Sage Publications, 2008.

Pflatz, John L. “Entropy in Social Networks.” Paper presented at SocInfo 2012, Lausanne, Switzerland, December 2012.

Plato. *Republic.* Translated by George Maximilian Anthony Grube. Revised by C. D. C. Reeve. Indianapolis: Hackett, 1992.

Pohlandt-McCormick, Helena. “In Good Hands: Researching the 1976 Soweto Uprising in the State Archives of South Africa.” In *Archive Stories: Fact, Fiction, and the Writing of History*, edited by Antoinette Burton, 299-324. Durham: Duke University Press, 2005.

Pons, Pascal, and Matthieu Latapy. "Computing Communities in Large Networks using Random Walks." In *Computer and Information Sciences: 20th International Symposium on Computer and Information Sciences*, edited by PInar Yolum, Tunga Güngör, Fikret Gürgen, Can Özturan, 284-293. Berlin: Springer, 2005.

Portelli, Alessandro. *The Death of Luigi Trastulli and Other Stories: Form and Meaning in Oral History*. Albany: State University of New York Press, 1991.

Quine, Willard Van Ormond. “Two Dogmas of Empiricism.” In *Quintessence: Basic Readings from the Philosophy of W.V. Quine,* edited by Roger F. Gibson, 31-53. Cambridge: The Belknap Press of Harvard University Press, 2004. Originally published in *From a Logical Point of View* (Harvard University Press, 1953).

R Core Team. *R: A Language and Environment for Statistical Computing* [Computer software]. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Riessman, Catherine Kohler. “Looking Back on Narrative Research: An Exchange.” In *Doing Narrative Research*, edited by Molly Andrews, Corinne Squire, and Maria Tamboukou, 78-85. London: Sage Publications, 2008.

Rosvall, Martin, and Carl T. Bergstrom. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences* 105, no. 4 (2008): 1118-1123.

Salmon, Phillida. “Looking Back on Narrative Research: An Exchange.” In *Doing Narrative Research*, edited by Molly Andrews, Corinne Squire, and Maria Tamboukou, 78-85. London: Sage Publications, 2008.

Schellenberg, T.R. *Modern Archives: Principles and Techniques*. Chicago: University of Chicago Press, 1956.

Scott, Peter J. "The Record Group Concept: A Case for Abandonment." In *Debates and Discourses: Selected Australian Writing on Archival Theory 1951-1990*, edited by Peter Biskup, Kathryn Dan, Colleen McEwen, Greg O'Shea, and Graeme Powell, 79-90. Canberra: Australian Society of Archivists, 1995.

Stake, Robert. "Qualitative Case Studies." In *The SAGE Handbook of Qualitative Research, 3rd ed.,* edited by Norman K. Denzin, and Yvonna S. Lincoln, 443-466. Thousand Oaks: Sage Publications, 2005.

Tononi, Giulio. "An Information Integration Theory of Consciousness." *BMC Neuroscience* 5 (2004): 42-64.

Ward, Matthew, Georges Grinstein, and Daniel Kelm. *Interactive Data Visualization*. Natick: A.K. Peters, 2010.

Wasserman, Stanley, and Katherine Faust. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press, 1994.

Appendices

Appendix A – Code for the Construction of Adjacency Matrix

```
#this macro creates a list of unique names from your edgelist, enabling you to use these names
in the adjacency matrix

Sub Macro4()

    Range("A1:A2").Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.Copy
    Range("C1").Select
    ActiveSheet.Paste
    Range("B1:B2").Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.Copy
    Range("C1").Select
    Selection.End(xlDown).Select
    ActiveCell.Offset(1, 0).Select
    ActiveSheet.Paste
    Range("C1").Select
    LR = Range("C" & Rows.Count).End(xlUp).Row
For i = LR To 1 Step -1
    If WorksheetFunction.CountIf(Columns("C"), Range("C" & i).Value) > 1 Then Range("C" &
i).Delete shift:=xlShiftUp
Next i


    Range("C1").Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.Copy
    Range("F2").Select
    ActiveSheet.Paste
    Range("G1").Select
    Selection.PasteSpecial Paste:=xlPasteAll, Operation:=xlNone, SkipBlanks:= _
        False, Transpose:=True

End Sub

#this macro tells you how many rows you have in the list just created
Sub Macro5()

Macro5 Macro
```

```
MsgBox Range("C1").End(xlDown).Row

End Sub

#to construct an adjacency matrix, copy-paste (using transpose specificiation) the list of names
so that it they are arrayed along the top. Then, fill in the spaces of the new matrix with the
following code (adjust range to reflect number of names)
Sub Macro6()
Sub Macro6 ()
Range("G2:P11").Formula = "=SUMPRODUCT(($A$1:$A$803=$F2)*($B$1:$B$803=G$1))"
End Sub
```

Appendix B – Code for Uploading Matrix to R

```
require(igraph)
cat<-read.csv(file="AdjMat.csv",header=T,sep=",")
cat<-as.matrix(cat)
g <- graph.adjacency(cat, weighted=T, mode = "directed")
#convert to undirected for fast greedy
g<-as.undirected(g)
#substitute various algorithms on the next line
fg<-cluster_fast_greedy(g,merges=T,modularity=T,membership=T)
g<-simplify(g)

plot(g,vertex.size=10,layout=layout_as_star, vertex.color="black",vertex.label.color='black',
vertex.label.family="sans",vertex.frame.color=
"black",vertex.label.font=2,vertex.label.dist=1,vertex.label.cex=.95,
edge.width=1,edge.color="black",mark.groups=communities(fg))
```

Alternative layout code to keep clusters distinct:

```
cat<-read.csv(file="AdjMat.csv",header=T,sep=",")
cat<-as.matrix(cat)
g <- graph.adjacency(cat, weighted=T, mode = "undirected")
fg<-cluster_fast_greedy(g,merges=T,modularity=T,membership=T)
V(g)$membership <- fg$membership

one<-as_ids(V(g)[membership==1])
one<-induced.subgraph(g,one)
one<-layout_in_circle(one)
oneone<-as.matrix(V(g)[membership==1])
```

```
one<-cbind(one,oneone)

two<-as_ids(V(g)[membership==2])
two<-induced.subgraph(g,two)
two<-layout_in_circle(two)
twotwo<-as.matrix(V(g)[membership==2])
two<-cbind(two,twotwo)

three<-as_ids(V(g)[membership==3])
three<-induced.subgraph(g,three)
three<-layout_in_circle(three)
threethree<-as.matrix(V(g)[membership==3])
three<-cbind(three,threethree)

one[,2]<-one[,2]+5
one[,1]<-one[,1]+5

two[,2]<-two[,2]+5
three[,1]<-three[,1]+5

black<-rbind(one, two, three)

black<-as.data.frame(black)
black<-black[order(black$V3),]
black <- black[,-3]
black<-as.matrix(black)

V(g)$name <- c(1:51)

plot(g,vertex.size=12,vertex.color="lavenderblush3",layout=black, vertex.label.dist=0,
vertex.frame.color='white', vertex.label.color='black', vertex.label.font=1)
g<-simplify(g)
```

Appendix C – Interview Guide

Hi ---, My name is Kate Chandler. Thank you for taking time out of your day to do an interview. It means a lot to me because the information you provide will help me evaluate the results of this study. By participating in this study, you're helping to develop a tool that is designed to make it easier for organizations to manage files.

This conversation should take less than 30 minutes. Just so you know, I have a hearing loss, so I may ask you to repeat at times.

[for soft voices/people with beards/people who put hands in front of mouth: To help me hear, I was wondering if you would wear this microphone that sends your voice directly to my hearing aids.]

For accuracy, I would like to audiotape this interview so I can have it transcribed. Your comments will not be associated with your name to ensure confidentiality.

[give them time to sign]

[prompt: If you prefer not to be audiotaped, I can take notes instead. As I am taking notes, I was wondering if you would wear this microphone that sends your voice directly to my hearing aids.]

The purpose of this study is to test a new method of organizing workplace files, using an approach called social network analysis. I've applied this social network analysis method to the files in the SharePoint sites used by your department; xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxx*. To be clear, the social network analysis method does not involve accessing file content, and I have not accessed the content of your files. Instead, the method involves extracting information about who is modifying the files.*

This information has been analyzed using three social network algorithms. Each algorithm comes up with different results. What I hope to discuss is the extent to which the results of these algorithms reflect what goes on in your organization. To this end, I'm going to ask you five questions.

1.

To the best of your memory, what are the main projects you have worked on over the past two years?

[prompt: **Please describe what you do at work**]

2.

As part of the analysis, I've shown the results of the algorithms in a visual way. Here is an example of a visualization of a social network analysis algorithm. You can see that there are dots, which represent people; lines, which indicate that one person modified the files of another; and coloured shapes, which indicates groups of people who frequently modify one another's files.

Do you have any questions?

[answer any questions]

3.

As you can see here, [show a paper copy of one visualization over time] *I divided the SharePoint files into four month segments, and made a visualization for each segment.*

[give them a moment to study the visualization]

I then repeated this process for two more algorithms [show a paper copy of the visualization called "Four Visualizations".] You'll notice that the last visualization is a repeat of algorithm C, but with the manager removed. I did this because one of my research questions is about managers. I hypothesize that managers modify so many files they make it hard for the algorithm to identify groups of frequently-interacting employees.

In your opinion, which of these four visualizations do you think best reflects the pattern of file modifications that involved you?

[prompt: which of these four visualizations do you think best represents your organization?]

4.

[Open the Interactive PowerPoint presentation that pertains to their choice of algorithm]

This PowerPoint presentation shows the files associated with each working group in the [Algorithm One/Two Three] visualization. If you click through the slides, you can see that it is the same visualization as the one on the piece of paper.

[allow them to click through it]

[If there are periods of time when the visualization indicates they did not modify files:] ***This visualization indicates that you did not modify files in the first eight months of 2014.*** **Do you remember that to be the case?**

[If there are periods of time when the visualization indicates they modified files with a coworker:] ***This visualization indicates that you and a co-worker modified the same files from August to September 2014.*** **Is that how you remember it?**

5.

For this group of files [pick a group of files they modified with another person], does it seem to you that the files belong together?

[If yes:] **Why do they belong together?**

6.

As you probably know, there are many different ways to group workplace files. For example, you can group files by the month they were created, or by their subject. Archivists are very concerned with how files are grouped. We believe that the way the files are grouped will affect the way future visitors to the archives will interpret the files.

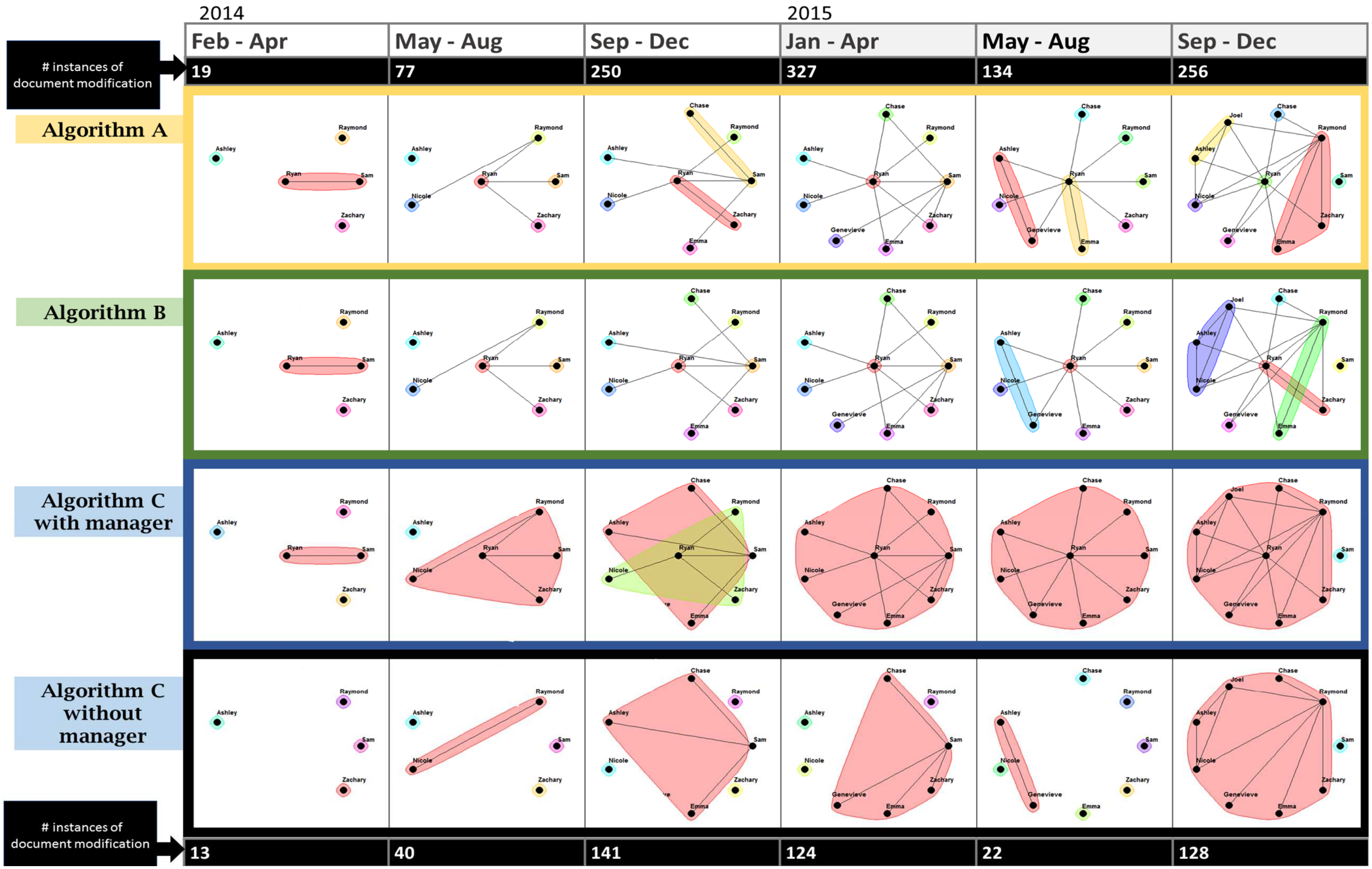
With this in mind, is there anything else you'd like to tell me about the way these files are grouped?

Thank you so much for your time. I am going to have this interview transcribed, and will summarize the interview. I will send this summary to you shortly, and you will be able to revise the comments. I will also ask you to sign a Renewal of Consent form – if you do not like the summary, I will destroy the transcript

and summary. If you do sign, I may use your insights in my thesis - without your name attached."

Appendix D – Scaled Version of Visualization Shown to Staff

Words are difficult to read in this visualization. It is shown here to help readers better understand interview proceedings. Please refer to *Figure 3—2* for a more readable version.


2014
2015
Feb - Apr
May - Aug
Sep - Dec
Jan - Apr
May - Aug
Sep - Dec
# instances of document modification
19
77
250
327
134
256
Algorithm A
Algorithm B
Algorithm C with manager
Algorithm C without manager
# instances of document modification
13
40
141
124
22
128
Ashley
Raymond
Ryan
Sam
Zachary
Nicole
Chase
Emma
Genevieve
Joel