May 2016

The Interdependence of Scientists in the Era of Team Science: An Exploratory Study Using Temporal Network Analysis

Mark R. Costa
Syracuse University

Abstract

How is the rise in team science and the emergence of the research group as the fundamental unit of organization of science affecting scientists' opportunities to collaborate? Are the majority of scientists becoming dependent on a select subset of their peers to organize the intergroup collaborations that are becoming the norm in science? This dissertation set out to explore the evolving nature of scientists' interdependence in team-based research environments. The research was motivated by the desire to reconcile emerging views on the organization of scientific collaboration with the theoretical and methodological tendencies to think about and study scientists as autonomous actors who negotiate collaboration in a dyadic manner. Complex Adaptive Social Systems served as the framework for understanding the dynamics involved in the formation of collaborative relationships. Temporal network analysis at the mesoscopic level was used to study the collaboration dynamics of a specific research community, in this case the genomic research community emerging around GenBank, the international nucleotide sequence databank. The investigation into the dynamics of the mesoscopic layer of a scientific collaboration networked revealed the following—(1) there is a prominent half-life to collaborative relationships; (2) the half-life can be used to construct weighted decay networks for extracting the group structure influencing collaboration; (3) scientists across all levels of status are becoming increasingly interdependent, with the qualification that interdependence is highly asymmetrical, and (4) the group structure is increasingly influential on the collaborative interactions of scientists. The results from this study advance theoretical and empirical understanding of scientific collaboration in team-based research environments and methodological approaches to studying temporal networks at the mesoscopic level. The findings

also have implications for policy researchers interested in the career cycles of scientists and the maintenance and building of scientific capacity in research areas of national interest.

The Interdependence of Scientists in the Era of Team Science: An Exploratory Study Using Temporal Network Analysis

by

Mark R. Costa
B.A., University at Buffalo, State University of New York, 2002
M.L.S., University at Buffalo, State University of New York, 2003

Dissertation

Submitted in partial fulfillment of the requirements for the degree of doctor of philosophy of Information Science and Technology

Syracuse University
May 2016

Acknowledgements

I wanted to take this opportunity to say thank you to my advisor for giving me the latitude to explore my interests and encouragement when I needed to move forward. I also wanted to say thank you to members of my committee for providing feedback throughout this process and helping me turn a rough idea into a finished product. I know they invested a significant amount of time to give me constructive and helpful feedback, and I definitely appreciate it.

I would also like to acknowledge the support I received from the National Science Foundation's Science of Science Policy program (grant number 1262535). Without their support, this research and my current career trajectory would not have been possible.

Finally, I would like to thank my family for all of their support over the years. To my parents: thank you for putting just enough stubborn in me to see this through and giving me that first computer – I told you it would pay off. To my wife, Nicole: thank you for all of your support and tolerance – I know this has been a long process. To my daughter, Aria: thank you for keeping me grounded and helping me put this work in perspective. I am sure there are a number of people I should be thanking, please forgive me if I have not included your name, even if

you’ve helped me quite a bit. You’ll have to settle for an acknowledgement offline, as this dissertation is heading out the door and I am out of time.

Table of Contents

List of Tables

List of figures

1 Introduction

1.1 Background and research questions

In his 2001 paper "Reflections on scientific collaboration," Donald Beaver noted that "The MODE of coauthorship was 2 [in 1978]. (It still is today, especially if one counts *laboratories* instead of individual coauthors)." What Beaver was referring to was the fact that collaboration teams are now more likely to be assemblages of two or more research groups, and that the research groups can be viewed as their own entities. Research teams are getting larger (by looking at the number of coauthors per paper); the mean number of authors per paper has increased for almost all scientific disciplines, while the proportion of papers that are solo authored continues to fall, along with the impact of those papers (Wuchty, Jones, & Uzzi, 2007). This trend has progressed to the point where we are now tracking hyper-authored papers, trying to make sense of very large scale collaborative efforts and what they mean for individual contributions to scientific knowledge (King, 2012).

Over fifty years ago, De Solla Price (1963) noted the decline in the tendency for scientists to function independently, although he still seemed to view scientists as independent agents who collaborate with one another (as dyads). What is being observed today is a move from viewing the scientist as the fundamental unit of organization in the sciences to the research group (Ziman, 1994). Henk Moed, an impact analysis expert, makes a similar argument, suggesting that the research group is the fundamental unit of business, and it is the research groups that should be assessed on their contributions to the field, not individual scientists. (Moed, 2006). These observations contradict some of our beliefs about the scientist as an independent creative worker, as well as about the motivations of scientists themselves. Beaver recognizes this, noting that "It

is an open question whether and how such an organizational style can long continue, given individual's self-interest in obtaining recognition of their own creativity."

Although there has been a strong move toward team science, the move has not been complete. Some scientists still have the desire to establish themselves as independent researchers, and to that end may work on smaller projects that are either authorized or unauthorized by group leaders. Individual scientists transition between research groups, bringing their knowledge, skills, and experience with them, while methodological specialists travel between groups, filling a specialized niche within certain communities (Velden, Haque, & Lagoze, 2010). There clearly are cases where scientists act independently of their primary research group; the extent to which this happens most likely varies by discipline, field, and national setting (Whitley, 2000).

A number of factors are contributing to the increasing trend of team-based research. Initially, collaboration emerged along with the professionalization of science (Beaver & Rosen, 1978). After professionalization came specialization; increasing task complexity means that the basic work of research requires more scientists with specialized expertise (Hara, Solomon, Kim, & Sonnenwald, 2003). The emergence of information and communication technologies (ICT) in general, and cyberinfrastructure in particular, enables, or at least facilitates coordination of large-scale team efforts (Szalay & Blakeley, 2009). The problems tackled by modern scientists are more complex, and modern funding institutions are increasingly interested in bringing interdisciplinary teams together to tackle those problems. In addition to the factors mentioned above, scientists who participate in team-based research projects may also experience increases in productivity, fewer errors, greater resilience against failures (Beaver, 2001), and greater citation impact (Uzzi, Mukherjee, Stringer, & Jones, 2013).

Although there are strong incentives to participate in team-based research, there are disadvantages as well. Teamwork requires greater effort invested in coordination and more resources, which often results in principal investigators (PIs) shifting their focus from bench skills to fundraising and administrative skills. Larger teams also render their lesser known participants invisible: “Most participants are invisible, in a formal sense, to the larger research community. They are just ‘names’ on a paper, ‘fractional’ scientists, essentially anonymous” (anonymous researcher, from Beaver, 2001).

Being rendered invisible is a significant problem for scientists because the accumulation of reputation is both a goal and a reward in the social system of science (Merton, 1973). Scientists make contributions to the collective body of knowledge in exchange for reputation. That reputation can then be used to gain more opportunities to conduct research, secure a paid position as a researcher, and secure grants. From this perspective, reputation is a form of capital, which is why the metaphor of social capital has proven to be a useful lens for studying the production of knowledge (Gonzalez-Brambila, Veloso, & Krackhardt, 2008; Lin, 1999). What Beaver (2001) alludes to is that it is more difficult for scientists to establish a reputation outside of their immediate group of collaborators because their names are buried in long lists of coauthors.

At a more basic level, reputation matters, both in terms of professional competence and interpersonal compatibility (Hara et al., 2003; Melin, 2000); research involves risk, and all participating parties want to know that their partners are competent enough to conduct the research, are committed to seeing it through, and are able to work with others in a demanding environment. Having a past reputation as a successful researcher who is good to work with can certainly facilitate the formation of relationships. Trust and comfort with one another are

frequently cited as criteria scientists use to screen potential collaborators. Reputation regarding one's ability to work in a team environment may spread by word of mouth, but the publication record may be the best advertisement of professional scientific competence. Large teams make it difficult for junior researchers to advertise their expertise through the publication record because their names are one among many. The fact that so many scientists are listed as coauthors on papers, and any given project requires the integration of multiple skill sets motivated Moed (2006) to argue that is extremely difficult to assess the impact of any one scientist—only extensive knowledge of each paper and a thorough review of each scientist's contribution to those papers can reveal individual contributions.

Flipping this problem on its head, we can see another problem—if we are to look for a scientist with particular expertise, how would we know which scientist in a group of seven to ten authors has the expertise we are looking for? Even if the problem of reputation development with respect to earning confidence was overcome, the problem of scientists finding one another looms large. Beaver's remarks suggest that, in certain fields of research, collaboration is not organized dyadically. Few scientists are engaged in searchers for potential partners, as some of the prior literature would suggest (Bozeman & Corley, 2004; Melin, 2000). Instead, scientists with their own teams are looking for other scientists with teams that bring the set of skills needed to complete the research.

If the fundamental unit of organization in the sciences is moving toward the research group and research teams are increasingly organized as assemblages of multiple groups, then an argument can be made that scientists are becoming more dependent on their groups for opportunities to participate in research projects. The reason for this is that fewer collaborations involve the integration of multiple independent actors, and instead focus on finding groups that

bring the required expertise together, because it is more efficient to delegate recruitment to the subunits of the project. Think about running a large grant or project with multiple researchers across institutions—it would not be efficient for the PI to hand select all the students and postdocs across all of the institutions; instead, the other researchers would take on the responsibility for selecting their own team members. The question is whether science is becoming more hierarchically organized and whether established scientists are getting caught up in the team assembly process, which leads to the first research question in this dissertation: **How is the increasing prominence of the research group and team-based research impacting scientists' dependence on one another and the research group?**

From Beaver's (2001) interviews and discussions on the composition of groups, it's clear that there are status differentials within groups. Groups are often named after the senior scientist (e.g., http://www.broadinstitute.org/scientific-community/science/core-faculty-labs), and collaborative projects are frequently organized by the group leaders in some fields (e.g., Velden et al., 2010), leaving some ambiguity as to what role other members of the lab play in organizing research projects, identifying which projects they want to participate in, and how frequently they have the opportunity to act independently of their group. Basically, what this line of reasoning is working toward is asking whether scientists' dependence on the group is differentiated. This idea needs further elaboration, which will be done using a framework from Complex Networks that highlights one possible way to identify differential dependencies in the group structure of scientific fields.

The inspiration for thinking about whether scientists' dependence is differentiated in fields characterized by a strong group structure came from both the ambiguity in the literature on scientific collaboration, as well as a specific framework from the field of Complex Networks. In

particular, that framework came about in response to the need to differentiate nodes based on their position within the group structure of complex networks (Guimerà & Amaral, 2005). Before going into more detail on that framework, let's explore the motivation for its creation.

A significant body related to the use of Complex Network Analysis (CNA) frameworks and techniques to study scientific collaboration has emerged over the past 15 years. There are many studies looking at scientists' accumulated sets of collaborative relationships (X. Liu, Bollen, Nelson, & Van de Sompel, 2005; Newman, 2004b) in an attempt to identify major players within a research community, or to identify the relative advantages that structural positions confer within scientific fields (Abbasi, Hossain, Uddin, & Rasmussen, 2011; Bonaccorsi, 2008). Many of those studies focus on the individuals' positions within the larger network, referred to as microscopic network analysis. One of the weaknesses of microscopic network analysis is that it ignores the prominent group structure, or mesoscopic layer, of most real-world networks that arises from and influences interactions in the network (Guimerà & Amaral, 2005). Mesoscopic network analysis focuses on the group structure and actors' positions within that group structure. In mesoscopic network analysis, position is a combination of two variables (versus one variable in microscopic network analysis) that measure the strength and distribution of ties within their group and between groups. The core argument made in (Guimerà & Amaral, 2005; Guimerà, Sales-Pardo, & Amaral, 2007a) is that nodes or actors in a network fall into various roles based on their connections within and between groups. Work by Velden and colleagues (2010) found that there is a good correlation between the group structure of collaboration networks and functional research groups (described later) and that the roles identified by Guimerà facilitate the differentiation of scientists' affiliations with the groups.

The idea of position within the group structure inspired the following question: **What is the relationship between a scientist's distribution of relationships in the group structure and their dependence on other scientists and their group, and how has the relationship between distribution and dependence changed over time?** The rationale underpinning this question is based on the assumption that a scientist's connections reflect the professional relationships that they can tap into for opportunities to participate in research projects, and that scientists with diverse connections to many research groups may have more opportunities to work independently of their group or group leader.

1.2 Theoretical framework

1.2.1 Core concepts

Scientific collaboration—"the system of research activities by several actors related in a functional way and coordinated to attain a research goal corresponding to these actors' interests" (Laudel, 2001). In this dissertation, it is assumed that the actors' shared interest is in publishing research results in order to participate in the reputation based system of science (Whitley, 2000).

Collaboration network—the collection of structural patterns that emerge from the collaborative interactions of scientists within a community. The network, as an object, is the result of graphing the interactions of scientists from trace data as a set of nodes or circles connected via lines or edges. In a collaboration network, the nodes represent the scientists, and lines or edges connecting those nodes represent the presence and intensity of past collaborative interactions (Newman, 2001c).

Dependence—the extent to which one scientist relies on another scientist to either: (a) provide access to research equipment, skill sets, and resources, or to coordinate projects, or (b)

perform the work needed to ensure the successful completion of the research projects(s) he or she is given access to. In an environment where collaborative interactions are brokered by PIs or lab leaders, lab members depend, to varying degrees, on the PI to coordinate research projects, and the PI depends on the lab members to contribute to the projects he or she coordinates.

Scientists may depend on one another for access to technical and financial resources (Stephan, 2012), for their ability to assemble and manage project teams (social capital) (Bozeman, Dietz, & Gaughan, 2001) or for their ability to do the work, in the same way a manager depends on his or her subordinates to do the work assigned to them. In this dissertation, the variations of dependence are aggregated together and viewed as the extent to which a scientist's actions are, or are not, autonomous of other scientists in his or her network and group (see below).

Research groups—Seglen and Aksnes (2000) identified functional research groups as the set of one or more senior scientists, junior researchers, and doctoral students that make up the core of a lab or group. In addition to the core members of the group, there is a set of loosely affiliated researchers who work sporadically over time, or intensely for a short period of time, with the core members of the group. The members of the group may be bound by formal affiliation, but they need not be.

Distribution of relationships—the structural form of social capital. The group structure of a collaboration network refers to the groups that form within the network due to higher rates of interaction between members of the groups in comparison to rates of interactions with scientists external to the group. The patterns of connections between those groups are the group structure (Guimerà & Amaral, 2005): the set of collaborative relationships a scientist has within

their home group, and to other groups, within the research community. Scientists vary in terms of the breadth and depth of their ties to their own research group and external groups, potentially giving them access to different research opportunities and resources. To that end, a scientist's distribution of relationships is a structural form of social capital—the professional relationships a scientist has and the resources available through those relationships (Burt, 2001; Nahapiet & Ghoshal, 1998).

1.2.2 Framework

This dissertation used a subfield of Complexity theory, called Complex Adaptive Social Systems (CASS), to structure the exploratory investigation. CASS argues that social systems are *complex* because local interactions are seemingly random, yet give rise to an ordered structure (Gell-Mann, 2002; Ladyman, Lambert, & Wiesner, 2012a). *Adaptation* refers to both the system's ability to change to external stimuli and the internal agents' response to the social structure that emerges due to their interactions (Ladyman et al., 2012a; Sawyer, 2005; Wagner & Leydesdorff, 2009). Human *social* systems differ because the agents within the system are able to abstract and communicate about the order that emerges through their local interactions (Beckner et al., 2009; Sawyer, 2005).

There are two approaches to thinking about complex systems, and they are deeply intertwined with their methodology. The first approach focuses on questions pertaining to the ability of the system itself to adapt over time, and is associated with research that relies heavily on simulations to explore how combinations of basic interaction rules influence the system's ability to adapt (Holland, 2006). The second framework, called Complex Networks, looks at the network of relationships between actors in the system, analyzing the structural patterns that

emerge and how they influence interaction. The latter approach, which was used in this dissertation, is used to study structures of various networks: brain (Bullmore & Sporns, 2009), social (Girvan & Newman, 2002), metabolic (Guimerà & Nunes Amaral, 2005), and transportation (Colizza, Barrat, Barthélemy, & Vespignani, 2006) networks.

The core thesis of this dissertation is that scientific collaboration, as a system, is comprised of scientists whose relationships form group structures (Guimerà et al., 2007a), and that individual scientists' interactions are influenced by the group structure that has emerged due to their historical interaction, as well as their individual positions within that group structure. What is specifically argued is that the research group, as a sociological construct, reflects the concentration of technical, financial, and human resources needed to conduct research (Ziman, 1994), and that only a select subset of scientists in scientific fields characterized by a strong group presence have the ability and desire to coordinate research activity. As a result, other scientists are dependent on the coordinators for access to research opportunities. However, that dependence should be mitigated by the individual's collection of relationships within and between groups. Furthermore, as the system adapts and the collaborative interaction intensifies in the era of team science (Wuchty et al., 2007), the influence of the group structure grows stronger. As the influence of the group structure grows, the interdependence of scientists, or the description of how two scientists are dependent on each other in a collaborative relationship to varying degrees, evolves as well.

Treating relationship formation in a complex network as a brokered interaction between multiple parties is a departure from traditional complex network analysis, which usually views interactions as a set of dyadically orchestrated ties (Barabási et al., 2002, 2002; H. Jeong, Néda, & Barabási, 2003; Newman, 2001c). There are some exceptions (Estrada & Rodriguez-

Velazquez, 2006; Guillaume & Latapy, 2004; Taramasco, Cointet, & Roth, 2010), but none explores the role that groups and actors within those groups play on tie formation.

1.3 Methodology and context

The research presented in this dissertation was exploratory in nature because there was no strong theoretical guidance regarding the relationships between the core concepts in the research questions (Schutt, 2006). In terms of methodological approach, the research revolved around the temporal analysis of complex networks (Holme & Saramäki, 2012) to model the patterns of **scientific collaboration** over time. **Scientific collaboration** was operationalized as coauthorship on a paper (Glänzel & Schubert, 2005), and **collaboration networks** were reconstructed based on the coauthorship data. The research community emerging around GenBank, the international nucleotide sequencing databank, served as the focus of this study. The portion of the bioinformatics research community that is focused on sequencing and submitting DNA to the repository is an excellent example of a field that is interdisciplinary, organized around the research group, dependent on expensive equipment, and where scientists participate in intergroup team research. Thirty years of publication data on 295,134 articles written by 393,528 authors were used to study the evolving nature of the network and the changing nature of dependence within that network.

Network analysis was conducted at the mesoscopic level (Guimerà & Amaral, 2005) using community detection algorithms to extract the **group structure** (Rosvall, Axelsson, & Bergstrom, 2009). The node role framework developed in (Guimerà et al., 2007a) served as the framework for classifying scientists based on their **distribution of relationships** within that mesoscopic structure. In terms of constructing the network, an experiment was conducted in this

dissertation, where different approaches to constructing networks were compared. More specifically, two major network types—bipartite and unimodal—were analyzed for their ability to help predict future interactions of scientists. The primary motivation behind comparing these two types of networks is that unimodal networks are far more commonly used but model the underlying phenomenon as a series of dyadic interactions, which is exactly the viewpoint this dissertation is trying to move away from. In contrast, the bipartite type of network models the underlying phenomenon as a series of multi-actor interactions, which is what scientific collaboration is in a team environment. The major limitation of using bipartite networks is that there is far less literature to tie the results of bipartite network analysis back to. The final part of the investigation involved analyzing scientists' future dependencies on one another based on their positions derived from historical interactions.

The statement in the following section is a summary of the contextual and methodological problems that motivated this research.

1.4 Problem statement

Scientific collaboration has evolved from a dyadically arranged affair to a team-based activity organized around the combination of research groups, yet approaches to modeling the collaborative interactions of scientists from the network analytic perspective have not reached the point where they reflect that reality. Because dyadic modeling of collaboration remains the dominant approach, we often underestimate the extent to which scientists are dependent on one another to participate in, and conduct, research projects.

There are some caveats to the problem statement, but in this dissertation, I argue that the statement holds true for at least one scientific field – the bioinformatics community (and probably many more).

1.5 Contributions and impact

For many scientific fields, scientists must collaborate in order to produce meaningful research and be considered participating members of the community. If we assume that the goal of science policy is to foster scientific capacity in the various scientific fields, and part of scientific capacity is the human capital contributing to those fields, then tracking scientists' ability to participate in research projects and develop their skill sets is an important consideration for policy research (Bozeman et al., 2001). In particular, understanding how scientists form collaborative relationships is an important component of any framework that analyzes their participation in the research space of their field.

This study contributes to our understanding of scientific collaboration in three ways: (1) the results demonstrate that scientists are more dependent on their research group for opportunities to participate in research and to publish; (2) it provides evidence for the argument that scientists are increasingly interdependent, but; (3) that dependence is highly differentiated by the roles scientists play in connecting the group structure of the collaboration network together. Finally, the dissertation presents an empirically tested refinement of the method for studying scientific collaboration networks that are characterized by a strong group structure.

There are two major limitations to this study, the first being that it is restricted to one research community, so the generalizability of the findings is limited. The second limitation stems from the fact that the core concept of dependence was measured using publication data

only, thus missing out on other types of opportunities researchers have to collaborate on research, including failed attempts to publish results and collaboration on the generation of reusable datasets.

1.6 Organization of this dissertation

The dissertation proceeds as follows: First, a review of the literature on scientific collaboration and complex systems is provided, with emphasis on the following areas: antecedents of scientific collaboration, feedback mechanisms in the system of scientific collaboration, and use of complex systems and network analysis to study the phenomenon of scientific collaboration. Chapter 3 contains the theoretical framework used to structure the dissertation research, followed by the methodology in Chapter 4. The methodology chapter explicitly outlines the operationalization of concepts and the general approach to analyzing the data in this exploratory study. Chapter 4 also contains information related to the selection of the data source and a background discussion of the motivation for, and reasoning behind, the work related to improving methodological approaches for temporal network analysis. Chapter 5 contains the analyses, including the research related to the empirical testing of the methodological approach outlined in Chapter 4. Chapter 6 provides the discussion, relating the theoretical framework to the analyses and observations in the literature. Finally, the dissertation ends with the conclusion, including a discussion of the limitations of the research.

2 Literature review

2.1 Overview

The literature review starts with a discussion with an overview of the operationalization, forms of, and units of analysis of scientific collaboration. A discussion of general trends in scientific collaboration follows, highlighting the emergence of team science as one mode of collaboration. An exploration of the antecedents of collaboration follows. What will be shown is that there can be numerous factors influencing the formation of any particular collaborative relationship, which results in the process looking highly random at the local level (Beaver, 2001). The effects of collaborating are discussed next in order to demonstrate that there are feedback mechanisms that encourage scientists to collaborate.

The apparent randomness of collaborative interactions at the local level, the presence of feedback mechanisms in the system, and the influence of the group structure on scientific collaboration all contributed to the motivation for selecting a Complex Systems framework to structure the research in this dissertation. The chapter concludes with a review of the literature on Complex Systems, with particular emphasis on Complex Systems-based approaches to studying scientific collaboration.

2.2 Operationalization, Forms, and Units of Analysis

There are a number of methods used to measure scientific collaboration. The primary method of measuring collaboration is to analyze the patterns of coauthorship contained in a body of scientific literature. Researchers have also mined the acknowledgments sections of papers, realizing the importance of contributions that do not end in the production of a paper. Finally,

researchers have also used surveys and interviews to collect data. This study relies on the unobtrusive mode of analyzing collaboration via coauthorship patterns. Consequently, the majority of the literature review focuses on studies conducted on coauthorship networks.

Coauthorship and collaboration are often used interchangeably in the literature when in fact coauthorship is an operationalization of the concept of collaboration. Researchers have pointed out that using coauthorship as a measure of collaboration raises several content and construct validity concerns. Construct validity concerns stem from the fact that in some disciplines, authors are given honorary coauthorship even when they have not contributed to the paper (Katz & Martin, 1997). Content validity concerns stem from observations that many instances of collaboration do not culminate in a publication. Thus, coauthorship underrepresents scientific collaboration (Laudel, 2002), This is an acknowledged limitation of the operationalization, which can be addressed to some extent by supplementing the data collection with interviews or surveys if necessary. However, coauthorship is still considered to be both a useful and economical tool to measure collaboration (Glänzel & Schubert, 2005).

Three dimensions can be used to characterize studies of scientific collaboration networks –team size, disciplinary integration, and unit of analysis (see Table 2-1). There are two primary considerations with size; first, whether or not collaboration is “better” than solo authorship on a number of dimensions, including productivity (Abramo, D’Angelo, & Solazzi, 2011; Beaver & Rosen, 1979; Braun, Glänzel, & Schubert, 2001; de Solla Price & Beaver, 1966), impact (Abramo et al., 2011; Defazio, Lockett, & Wright, 2009; Glänzel & Schubert, 2001) and visibility (Beaver & Rosen, 1979; Cole & Cole, 1968; Pao, 1992). The second consideration with team size is the degree to which the coordination costs of managing a large team exceeds the productivity gains associated with collaboration. For example, Persson and colleagues (2004)

found that productivity tends to increase as the number of collaborators increases, up until a disciplinary-specific asymptote, beyond which productivity gains invert and begin decreasing.

Size	Disciplinary integration	Unit of analysis
Single authorship	Intra-disciplinary	Individual
Coauthorship (2 authors)	Inter-disciplinary	Intra-departmental/ Research group
Multi-authorship	Trans-disciplinary	Inter-departmental/Intra-institutional
Large-scale	Academic-Industry	Inter-institutional International

Table 2-1: Dimensions of collaboration

Much of the literature on collaboration focuses on intra-disciplinary interactions in the production of scientific knowledge. However, there are a number of research areas that explore the exchange of ideas between disciplines (Qin, Lancaster, & Allen, 1997), and the interactions between Universities and Industry (Rosenberg, 1998; Wong & Singh, 2013).

In addition to the previously mentioned dimensions, there are also a number of ways to aggregate the production of scientific knowledge. Researchers have analyzed collaboration networks between individual scientists (Abramo et al., 2011; Braun et al., 2001; Melin, 2000), research groups, institutions at the domestic and international levels (Ardanuy, 2011; D. H. Lee, Seo, Choe, & Kim, 2012), and countries (Glänzel & Schubert, 2001; Glänzel & Winterhager, 1992; Luukkonen, Tjissen, Persson, & Sivertsen, 1993).

2.3 Trends in scientific collaboration

Early research into the collaboration patterns of scientists revealed the steady increase of coauthorships and multi-authorships (de Solla Price & Beaver, 1966); subsequent studies have found that this trend continues to hold (Abramo et al., 2011; Braun et al., 2001; Luukkonen, Persson, & Sivertsen, 1992; Persson et al., 2004; Wagner & Leydesdorff, 2005). While the

aggregate pattern holds, researchers have found that (international) collaboration varies by country and does not always increase (Glänzel, Leta, & Thijs, 2006). In comparison, (Wagner & Leydesdorff, 2005) found that internationally coauthored papers doubled from 1990-2000. More recently, Chang and Huang (2013) found that less than 10% of astronomy and astrophysics papers were solo authored while over 50% of the papers were authored by international teams.

From a network analysis perspective, there is evidence that the level of connectivity increases over time (Wagner & Leydesdorff, 2005). Despite this general trend researchers have found that some research areas can be characterized as sparse networks with most collaboration being intra-institutional (Abbasi et al., 2011). Thus, while the general trend in science moves toward increased collaboration local areas of research may not follow this trend. The implications of this observation are not known, although based on Crane's (1972) observations the lack of collaboration may lead to the decline of a field. More specifically, lack of collaboration results in maintenance of what some would refer to as social capital or general esprit-de-corps of the community. Scientists self-select out of the profession or explore other research areas when opportunities within a specific area decline.

2.4 Antecedents of collaboration

Factors influencing collaboration are divided into three categories (Table 2-2)—social, economic, and cognitive (Luukkonen et al., 1992). The first category encompasses the social factors internal to science, as well as a few factors external to science that influence collaboration patterns. The second factor deals with economic incentives and limitations affecting collaboration choices. Finally, the third category deals with the factors related to the knowledge required to produce scientific research.

Economic	Cognitive	Social
Access to resources; access to equipment; Geographical proximity; Grant driven	Access to specific knowledge/ capabilities	Homophily (institutional, ethnic, and status driven); Interpersonal (friendships); Hierarchical (guided/directed); Political ties

Table 2-2: Economic, cognitive, and social factors influencing collaboration

2.4.1 Economic factors

Economic factors largely deal with the effects of resource constraints on collaboration. Starting at the national level, collaboration is found to be inversely proportional to the volume of scientific output in an area (Luukkonen et al., 1992). The accepted explanation for this phenomenon is that richer countries invest more in R&D infrastructure, thus are not as likely to need to collaborate to fill equipment limitations. In a study on the motivations behind scientific collaboration, gaining access to special data or equipment was ranked the second most important reason (20% of responses) for collaborating (Melin, 2000).

Physical distance has functioned as a good predictor for the probability of collaboration occurring for many years. Results from some studies indicate that physical proximity has the greatest effect on collaboration (Kraut & Egido, 1988). Further research found that collaboration rates tend to decrease exponentially with distance (Katz, 1994). The main reasons distance influences collaboration are - 1) distance reduces serendipitous encounters that lead to collaborative projects, and 2) the costs of supporting travel to maintain coordinating activities is relatively high.

Somewhat counter to this notion, researchers have found that financial support of research encourages collaboration (de Solla Price & Beaver, 1966; Pao, 1992). Along a similar line, researchers who have larger grants are more likely to have larger collaboration networks

(Bozeman & Corley, 2004). Additionally, international collaboration increases over time due to investments in e-science cyberinfrastructure (Gorraiz, Reimann, & Gumpenberger, 2011). IT-enabled collaboration has reduced some of the effects of distance, but physical proximity still has a significant effect on collaboration.

2.4.2 Cognitive factors

One of the earliest theories on why collaboration increases in science was based on the idea that increasing professionalization of science led to increased collaboration (Beaver & Rosen, 1979). This has largely disappeared as a factor simply because science is rarely, if ever, practiced by amateurs. A more relevant factor in contemporary science is the increasing specialization of science. Macro-level influences encourage inter-disciplinary and large-scale research projects, which often require assembling scientists with complementary skill sets.

Of all the factors influencing decisions to engage in a collaborative project, access to another scientist's specialized knowledge and skills is the most significant. Forty percent of scientists reported that access to another's knowledge was the most important reason, a rate roughly twice as high as the next most frequent reason provided (Melin, 2000).

An additional factor that influences the collaborative activities of a research area is the nature of the research itself. International collaboration is more common in basic research areas; conversely, international collaboration occurs at a much lower rate in applied areas of research (Frame & Carpenter, 1979; S. Jeong, Choi, & Kim, 2011; Luukkonen et al., 1992). The desire to retain intellectual property rights is considered to be the driving force behind this phenomenon.

2.4.3 Social factors

The effects of collaboration on social status were addressed in an earlier section. However, social status is also one of the primary influencers of the collaboration activities of scientists. The bi-directional nature of this relationship drives a dynamic system and is present at all levels of aggregation.

There is a strong tendency for highly successful researchers to collaborate with other highly successful researchers. A process of self-selection and recruitment influences this trend (Crane, 1972); for example, Nobel Laureates often collaborate with other laureates (Zuckerman, 1967). This trend continues for slightly less prominent researchers. For example, more experienced (van Rijnsoever & Hessels, 2011) and higher ranked (Vafeas, 2010) scientists tend to participate in collaborative research more often than their counterparts. Additionally, researchers affiliated with more prestigious departments collaborate more often than their peers (Piette & Ross, 1992).

These highly productive scientists, referred to as globals, are more likely to engage in formal collaboration outside of their main group (Pao, 1992), while locals are less productive and tend to have more limited formal collaboration networks. Taken from a slightly different perspective, continuants are highly productive researchers who stay working in an area over an extended period of time, collaborating with scientists entering or passing through the area (Braun et al., 2001). The data indicate that continuants tend to collaborate extensively with less stable and less productive actors in the network as a way to boost their research productivity.

These patterns tend to coalesce over time, creating invisible colleges in research areas. The invisible college is considered to be the in-group in a research area; its members are more

likely to share information and engage in informal and formal collaboration (Crane, 1972; de Solla Price & Beaver, 1966). This additional interaction positively affects the productivity and impact of the members. The reinforcement cycle that ensues is known as accumulated advantage, or the Matthew Effect (Merton, 1968) .

2.5 Effects of collaboration

The subsequent effects of collaboration can be grouped into two broad categories—productivity effects and social status effects. Productivity effects deal largely with the trade-offs between increased coordination costs and benefits of the division of labor. Status effects include the set of relationships between collaboration and the influence and acknowledgment of both the scientist and the scientist's work.

2.5.1 Collaboration and productivity

Research on the relationship between collaboration and productivity indicates that the relationship between the two variables is somewhat complex and not completely linear.

Globally, Persson (2004) and colleagues found that the overall distribution of productivity across all scientists shifted, such that the share of lower productivity authors decreased while the share of high productivity authors increased. From 1980 to 2000, the mean number of papers per scientist increased from 2.48 to 3.02, while the percentage of scientists who only authored one paper decreased from approximately 54% to 51%. At the far end of productivity, the percentage of scientists publishing more than 20 articles approximately doubled from 1% to 2%. This picture changes slightly when productivity is analyzed using normal versus fractionalized counting (S. Lee & Bozeman, 2005), with collaboration having a strong positive

relationship to normal measures of productivity and having little to no relationship to fractionalized counts.

Early research on the effects of collaboration on productivity highlighted a distinction between the hard scientists and humanities, with the productivity of scientists associated with large collaborative groups being higher than peers that had no such association (de Solla Price & Beaver, 1966). Pao (1982) found that collaboration had no significant impact on the productivity of music research scholars. Pravdic and Oluić-Vuković (1986) determined that the selection of collaboration partners affects the productivity rates of researchers. Collaboration with highly productive researchers increases productivity for scientists while collaborating with scientists who are less productive decreases productivity.

There appears to be a cost-benefit trade-off for collaboration, with productivity generally increasing as the number of collaborators increases until a discipline-specific asymptote is reached, after which productivity gains reverse and continue to decline (Braun et al., 2001; Persson et al., 2004). Longitudinally, collaboration does not have an immediate effect on productivity; however, productivity after a funded project increases between collaborators on the project (Defazio et al., 2009).

2.5.2 Status, visibility, and impact

The evidence on whether or not collaboration increases the participating scientists' visibility is once again mixed. For example, through historical research, Beaver and Rosen (1979) found that collaboration, particularly with more prominent researchers, increased the visibility of less experienced researchers. This observation was cautiously validated in some of Merton's research, where collaboration with Nobel Laureates was considered to be a double-

edged sword. To some extent, more of the credit for a collaborative work with a Laureate went to the Laureate, while simultaneously the collaboration often exposes the collaborators' names to a wider audience (Merton, 1968). Cole and Cole (1968) found no significant correlation between collaboration and impact. Socialization appears to be an added benefit of collaboration, increasing the likelihood of a scientist publishing more than once in an area (Beaver & Rosen, 1979).

The evidence is relatively consistent about the patterns of collaboration for eminent scientists. Collaboration was found to be common among 18th Century French scientists who achieved long-term recognition in their fields (Beaver & Rosen, 1979). Zuckerman (1967) found a similar pattern for Nobel Laureates, who are much more likely than the average scientist to collaborate. Furthermore, the formation and impact of collaborations are considered to be directly proportional to the academic excellence of its participants (Jones, Wuchty, & Uzzi, 2008).

Using the nation as a unit of analysis, collaboration is considered to be beneficial for countries with a less prominent stature when they collaborate with a more prominent nation (Glänzel & Winterhager, 1992), a finding built on by Schott (1998), who argued that lagging countries seek collaboration opportunities with leading countries in an attempt to increase their stature. This trend continues as researchers found a strong positive relationship between international collaboration and citation impact for Slovenian authors (Pečlin, Južnič, Blagus, Sajko, & Stare, 2012). Persson and colleagues (2004) found that international collaboration increases citation impact while Glänzel and Lange (2002) argue that the type of effect is most likely field specific.

As bibliometric indicators became more popular in research policy analysis, and concerns about productivity gave way to interest in impact, researchers began exploring the relationship between collaboration and citation or economic impact. The former category is concerned with general academic output while the latter is concerned with the commercialization of scientific labor.

Similar to productivity, the relationship between collaboration and impact is nuanced. For example, in a study of articles published in Ecology (Leimu & Koricheva, 2005), citation rates were found to be generally higher for multi-authored papers. However, self-citation rates increased as well. Choice of collaboration partners also influenced citation impact, with interdisciplinary and inter-institutional collaboration resulting in more citations and intra-institutional collaboration reducing citation impact. With respect to self-citations and increasing impact rates, Van Raan (1998) found that after adjusting for self-citations, the impact amplification effect of collaboration is still present. In a larger study of medical journals, researchers found a statistically significant, yet slightly variable, relationship between the number of authors on a paper and its citation rates (Figg et al., 2006).

In addition to concerns about the effects of R&D investment on academic productivity and impact, policy researchers are curious about the economic impacts of research. Of particular importance is understanding the commercialization opportunities for research, and what role University-Industry collaboration has in commercialization. The USPTO considers such collaboration to be a springboard for economic prosperity (taken from Abbasi et al., 2011) and has been found to trigger new basic research (Rosenberg, 1998). Some have argued that University-Industry collaboration results in higher rates of commercialization because of the more applied nature and targeted outcomes of industry research (Gregorio & Shane, 2003), while

others have argued that increased commercialization rates are possible because university's who collaborate with industry are able to access industry networks (Sætre, Wiggins, Atkinson, & Atkinson, 2009).

It is clear that the relationships between collaboration and status, productivity and impact are neither simple nor universally positive. Nevertheless, collaboration in many forms continues to increase and plays an ever more important role in the production of scientific knowledge. A number of reasons why collaboration continues to grow have been explored, and will be addressed in the next section.

2.6 Complexity, Complex Systems, Complex Networks

As numerous scientists have pointed out, there is no formal definition of complexity or complex systems (Johnson, 2007; Ladyman et al., 2012a). Johnson (2007) argues that complexity is "the study of phenomenon which emerge from a collection of interacting objects." The set of interacting objects is referred to as a system, and complex systems are thought to have several common characteristics, although, the precise set of characteristics differ between scientists (Ladyman et al., 2012a). Although scientists have different formal definitions of complexity and complex systems, there are sufficient commonalities to make the study of complex systems a coherent body of knowledge in that students of the idea are capable of understanding one another and building off of each other's work.

Weaver (1948) identified two forms of complexity—organized and disorganized. The former complexity results in the emergence of order, the latter chaos. The complexity dealt with in this dissertation is of the organized variety, i.e., dealing with the question—how do seemingly random interactions give rise to recognizable order? Simon (1991) referred to this order as

hierarchy; not in the sense that complex systems exhibit command and control order, but layers of order built upon one another. The hierarchy emerges through local interactions of agents in the system, as well as through the interactions of the agents and the order that arises from their interactions, and the interactions of the emergent structures themselves (Cilliers & Spurrett, 1999).

For a system to be complex, there has to be both order and randomness (Gell-Mann, 2002). A fully ordered system is not complex because it takes very little information to summarize the state of the system, and its response to stimuli is linear in the sense that it is predictable (Johnson, 2007). In contrast to a fully ordered system, a chaotic system has no order, and cannot be summarized by anything less than the full description of the system. A system is complex when there are regularities within the system that can be summarized, yet sufficient randomness that the system itself is non-linear in the sense that future states cannot be precisely predicted because random interactions can influence the evolution of the system in unpredictable ways (Ladyman et al., 2012a). Another way of thinking about non-linearity in complex systems is the presence of both delayed and immediate feedback mechanisms that can, but will not necessarily, create large effects (Arthur, 1999; Cilliers & Spurrett, 1999). Put another way, the response is not always proportional to the input.

Complex systems are also adaptive—the system adapts to external events or stimuli through the rearrangement of the relationships between the internal components of the system (Holland, 1992). One of the enduring questions in the area of Complex Systems is—how do systems evolve, and what are the basic mechanisms that promote the constant updating and rearranging of the relationships that give rise to the observed structures in a complex system (Holland, 2006). Because complex systems are in a constant state of adaptation, there is often no

observable equilibrium, as minor changes in the environment result in cascading effects throughout the system (Buldyrev, Parshani, Paul, Stanley, & Havlin, 2010; Goh & Barabási, 2008).

Armed with a description of characteristics of complex systems, we can now explore where studies of complex systems split. In one camp, complex systems are studied by proposing and testing basic building blocks of behavior and how those blocks give rise to ordered patterns of behavior observed in real-world systems (Holland, 2006). In the other camp are researchers who study the emergent structure of complex systems (Albert & Barabási, 2002; Newman, 2003) the basic mechanisms that give rise to that structure (Guimerà et al., 2007a; H. Jeong et al., 2003). One could argue that the increasing focus on temporal networks (Holme & Saramäki, 2012) is the latter group's gradual move toward investigating the dynamics of systems normally addressed by the former group (Mitchell, 2006; Niazi, 2011).

The research presented in this dissertation falls within the complex network camp because the questions focus on how the structure of the system, or more precisely, the actor's position within the structure of that system, influences the actor's behavior. The label "complex adaptive system" is still used because there is the expectation that the system will share the properties of other complex systems; it's just that the emphasis is on structural interpretations and frameworks. The remainder of the literature review explores the use of complex network analysis frameworks to study the system of science, as well as the use of trace data generated by the system of science to study complex networks.

2.6.1 Complex network analysis and scientific collaboration

The use of CNA to study scientific collaboration networks can be traced back to the early 1990's (Logan & Pao, 1990, 1991), although the recent surge in the use of the analytic framework can be attributed to Newman's (Newman, 2001a, 2001b, 2001c) studies analyzing large-scale scientific collaboration networks. We can use the results of his analysis (see Table 2-3) to motivate some of the discussion on how network analysis has been used to interpret the system of scientific collaboration. Measurements such as mean papers per author, mean authors per paper, and collaborators per author are standard measurements and have been employed extensively in the Scientometric literature; other measures, including degree distribution coefficient, clustering coefficient, and size of the giant component are introductions of network analysis.

	MEDLINE	Los Alamos e-Print archive	SPIRE	NCSTRL
Total papers	2,163,923	98,502	66,652	13,169
Total authors	1,520,251	52,909	56,627	11,994
Mean papers per author	6.4	5.1	11.6	2.55
Mean authors per paper	3.75	2.53	8.96	3.59
Collaborators per author	18.1	9.7	173	3.59
Cutoff z_c	5,800	52.9	1,200	10.7
Exponent τ	2.5	1.3	1.03	1.3
Size of giant component (%)	92.6	85.4	88.7	57.2
Mean/Max distance	4.6	5.9	4.0	9.7
Clustering coefficient	24	20	19	31

Table 2-3: Results on the macro-analysis of three scientific collaboration networks [Newman 2001c]

The first thing to note in Table 2-3 is that for each database, the distribution of the number of coauthors per scientist takes on a different value. In each case, the degree distribution

was estimated to follow a power law form with an exponential cut-off due the fact that collaboration is a resource-bounded activity; i.e., one only has so much time to make contributions to research. Furthermore, the power law forms each resulted in $R^2 > .99$ and $P < 10^{-3}$, indicating a good fit for each database analyzed.

Newman noted that there were marked differences in the distributions between scientific disciplines, with these differences reflecting the general social structure of the field. More specifically, he noted that the degree concentration was much higher in the biological sciences, perhaps reflecting the fact that lab managers place their name on every publication coming from the lab, and lab members have fewer opportunities to collaborate outside their group. This observation is empirically reflected in the τ value. A $\tau = 2$ is generally considered to be a cut-off between networks with distinct forms—for $\tau > 2$ the network tends to be dominated by few individuals who have very high degree centrality while networks with $\tau < 2$ tend to be characterized as more egalitarian because more actors participate in collaborative projects.

Another important point to note is that there is some variation across databases with respect to the percentage of scientists who are members of the largest component. In Newman's (2001c) study the size of the giant components range from 52.7% - 92.6%. Other studies have had rates as low as 38% (X. Liu et al., 2005).

There are two ways to interpret these results, and they are not mutually exclusive. The first reason we see lower inclusion rates for the large component is methodological in nature. That is to say, whether the researcher starts by selecting a social focus or a data source, it is possible to end up with a data set that does not contain sufficient data to reconstruct a fully connected network. The other reason why a network may exhibit many fragmented components

is theoretical—some communities may not foster or reward interaction. For example, it is less common for mathematicians to coauthor papers. Therefore, the probability that there are many isolated islands in the mathematics community is much higher than the probability of finding isolated islands in a discipline that incentivizes collaboration, such as high energy particle physics.

The relatively small values for mean distance, or average path length, is another characteristic feature of networks. This is referred to as the small world phenomenon: even relatively sparse networks create opportunities for short paths to emerge. Intuitively this makes sense; if a person knows 100 people, who in turn each knows 100 people that would result in each person being within 2 degrees separation of 10 000 people. We also see that the maximum path length for many networks scales sub-linearly, often at a log(log) rate, as the number of actors in the network grows. It is also important to note that the data source only includes publication data; therefore, it is quite possible that the actual average path length and maximum path lengths are smaller, particularly if we incorporate other social interactions that would facilitate the exchange of information regarding methods, concepts, and knowledge of others' abilities.

We also see that the clustering coefficient varies across the data sources, giving us some insight into whether or not there is a tendency for collaborators of one person to eventually collaborate with each other. As pointed out earlier, the social structure of the biomedical sciences hints that only a small proportion of the population gets the opportunity to forge new relationships, while in other disciplines the opportunities are much greater.

Subsequent research using network analysis to study scientific collaboration networks try to either demonstrate the utility of the approach on new data sets or to address some of the limitations of Newman's (2001c) study. In the former case, scientists use a similar approach as Newman with different research fields serving as the social focus. In the latter case, subsequent research attempts to address some of the limitations of Newman's approach. These limitations include the inability to: fully describe the heterogeneous nature of the network's topology; capture the temporal dynamics of interactions that give rise to the final network state; incorporate non-structural data to better explain the interactions between cognitive and social elements; account for differing intensities in collaboration; and present more nuanced views of actors' positions within the network.

Some of the early work using network analysis involved attempting to identify correlations between network concepts and existing Scientometric indicators and observations, including citation counts, scientific quality and the growth in international collaboration. Wagner and Leydesdorff (2005) hypothesized that the preferential attachment model predicted growth in international linkages. Their results indicate that the preferential attachment model fits reasonably well for only the middle of the distribution. Furthermore, they found that the collaboration clustering coefficients were orders of magnitude higher than what would be expected in a random network, but much lower than the observed values highlighted in Table 2-3. In addition to the clustering coefficient, the observed degree distribution diverged from what was observed in prior research. Wagner and Leydesdorff explained the deviation from prior observed power law distributions by hypothesizing that transients and newcomers occupied the hooked end while continuants occupied the middle of the distribution and hubs occupied the fat tail.

Work by Rigby and Edler (2005) looked at network density as a measurement of collaboration levels and correlated those densities to a normalized citation value of papers produced during a five-year period. Their results indicate that increasing collaboration levels are correlated with decreasing variability of research quality. Yang and Ding (2009) focused on the relationships between various centrality measures and citation counts, with betweenness centrality and PageRank correlations exceeding .52 and .41, respectively.

Another limitation of Newman's work was that it failed to account for differing intensities of interaction between actors in the network. There are at least two non-mutually exclusive ways to conceptualize intensity. The first is the number of times to scientists work together, with the intuition being that the more papers to scientists write together, the more likely they are to know each other well. This approach is used in (Li et al., 2005; Newman, 2001a, 2004b). The second consideration with respect to understanding intensity is to view the strength of the relationship for any collaboration as inversely proportional to the total number of collaborators.

Using a weighted network to represent scientific collaboration patterns can change subsequent centrality measurements significantly. This is particularly true for measurements that rely on simple calculations, like degree centrality. For example, (Newman, 2001a) found that analyzing weighted networks identified scientists who are well connected, not by the number of connections they have, but instead by the quality of their connections. Understanding the intensity of collaboration also makes way for identifying more nuanced relationships between collaborators.

Liu (2005) and colleagues developed a much more sophisticated approach to weight, which not only normalized the strength of a connection based on the number of collaborators, but also by the number of collaborations between authors. Furthermore, their approach enabled the use of a modification of PageRank (called AuthorRank), which is normally restricted to directed networks, to measure prominence within the network. Their work demonstrated strong correlations between degree centrality and PageRank (0.52) and degree centrality and AuthorRank (0.30), with the same author occupying the top spot of all three measures of centrality and AuthorRank. Furthermore, there was a high degree of similarity for top authors in each of the categories.

Similarly, if we try to determine the probability of triadic closure occurring, our intuition would be that the stronger the relationship between actors A and B, and Actors A and C, the more likely B and C would form a connection. Li and colleagues (2005) developed a formula for calculating the weighted clustering coefficient. This concept builds off of earlier work deriving a Weight per degree measurement (Fan et al., 2004) in an attempt to determine the tendency for actors to re-use their previously established connections. Nuanced calculations of weight will impact any subsequent analysis of the network structure, particularly when algorithmic approaches to clustering actors into groups are employed (see Chapter 5).

Even weighted approaches to analyzing scientific networks miss one of their important features—scientific networks are socio-cognitive networks. That is to say; there is a strong interplay between the content of the research and the social connections that result in its production. An initial approach to dealing with this limitation was the TARL (Topics, Aging and Recursive Linking) general process model (Börner, Maru, & Goldstone, 2004). Initial results indicated that the TARL model accounted for significant deviations from power law distribution

models in citation networks by simultaneously growing coauthorship and citation networks. The concept behind the TARL model is to view citations as an expression of cognitive interest and authors as embodiments of the topics cited, then view the evolution of the coauthorship network as being entwined with the evolution of the cognitive interest network. Research interests, via citations, drives collaboration choices, which in turn influence future research interests. The TARL model had the added benefit of looking beyond simple approaches to network evolution driven solely by growth, instead giving recognition to aging as an antagonistic force to preferential attachment (Anthony F. J. van Van Raan, 2000). It is important to note that this is functionally a bipartite graph, although the approach to its analysis is not explicitly consistent with bipartite graph methods.

Ozel (2012b) takes a different approach to looking at the interplay between social and cognitive networks by conceptualizing collaboration networks as a set of 3 related networks—Author-Author (A-A), Author-Knowledge (A-K) and Knowledge-Knowledge (K-K), and then used a meta-network perspective to analyze cascading influences across the three networks. The approaches taken by (Ozel, 2012b) and (Börner et al., 2004) highlight three interrelated weaknesses in the network analysis of scientific collaboration literature—a lack of understanding of how measurements at levels of analysis below the global level match with global measurements (Abbasi et al., 2011; Guimerà et al., 2007a), how different actors contribute to the non-uniform topological properties in networks (Chang & Huang, 2013; Guimerà et al., 2007a; D. H. Lee et al., 2012; Velden et al., 2010), and how those topological features change over time based on actions and/or shifting positions of the actors (Chang & Huang, 2013; Lee et al., 2012).

Using visual methods of analysis, it is quite apparent that many complex networks have non-uniform topological properties. However, early methods of analysis focused on generating

global measurements in an attempt to describe the network without addressing the non-uniform nature of the network. Guimera and colleagues (2007a) argued that, in many cases for biological and technical networks, modules (or groups) form within networks, and that actors' connectivity patterns in comparison to the connectivity patterns of other actors in their module would provide more insight into the role each actor played in the network. To that extent they focused on two measurements—within-module degree (z) and participation coefficient (P). The former measures the extent to which actors connect with other members of their group, while the latter measures the extent to which actors connect to actors outside their module.

Using these two indicators, they were able to identify 7 classes of nodes, with boundaries drawn through sparsely population regions of the zP-plane. The first 4 classes of nodes are considered to be non-hubs with $z < 2.5$, further separated by their relative P values.

		P	z_i
Non-hubs	(R1) Ultra-peripheral nodes	$P \leq 0.05$	< 2.5
	(R2) Peripheral nodes	$0.05 < P \leq 0.62$	< 2.5
	(R3) Satellite connectors	$0.62 < P \leq 0.80$	< 2.5
	(R4) Kinless nodes	$P > 0.80$	< 2.5
Hubs	(R5) Provincial hubs	$P \leq 0.30$	>= 2.5
	(R6) Connector hubs	$0.30 < P \leq 0.75$	>= 2.5
	(R7) Global hubs	$P > 0.75$	> =2.5

Table 2-4: Node role profiles based on participation coefficient (P) and within-module degree (z_i)

Furthermore, there tends to be role-to-role connectivity profiles that differ by network class. Broadly speaking, two main classes were identified based on these connectivity profiles. The first class consists of metabolic and air transportation networks, which are characterized by an overrepresentation of R1-R1 and R5-R6 and protein interactomes and the Internet, which is characterized by an underrepresentation of the two link profiles.

One important question that arose from these findings is whether or not this classification scheme would provide some insight into social networks. This question was addressed in later work investigating the mesoscopic structure and microscopic connection patterns of collaboration networks (Velden et al., 2010). Velden and colleagues argued that focusing on mesoscopic analysis underplays the roles individuals have in the network, while focusing on the microscopic level ignores the fundamentally team-based nature of modern research in many fields.

Results from their analysis reveal several interesting trends that confirm results from previous studies while highlighting some interesting weaknesses in other approaches. First, for their seed group cluster, the centrality measures were extremely high indicating a high level of centralization, while other clusters analyzed had lower centrality values than the seed cluster, but still high enough to be considered hierarchical in nature. The seed group cluster was dominated by a single hub node while other clusters had multiple hub-nodes. These results confirm observations by (Newman, 2001b) that biomedical sciences tend to be hierarchical in nature.

The results also indicated that previous approaches suffered from a weakness—they could not distinguish between the types of collaboration patterns. For example, career migrations often give the impression that two groups have collaborated when in fact the pattern of interaction is an artifact of one scientist establishing connections to a new set of collaborators and leaving the old collaborators behind. Based on the analysis, three broad connectivity patterns emerged: *1-1*, *1-m*, and *m-m*. In the *1-1* scenario, a single author connects two clusters. The *1-m* scenario is characterized by a scientist from one module connecting to many scientists in another module. Finally, the *m-m* scenario is based on many scientists in one module connecting to many other scientists in another module. A few examples of what gives rise to the three scenarios will

suffice here. *1-1* connections are usually an exclusive cooperation by closely collaborating colleagues or unauthorized collaboration by postdocs while *1-m* collaborations are usually the result of career migrations or one-off services. *M-m* scenarios are the result of much larger projects, where there is a strong emphasis on thematic and methodological cooperation, or in some cases the result of a cooperative agreement between a PI and national institute to bring the PI's research group to the institute. For a more comprehensive list of reasons driving the emergence of the 3 scenarios see (Velden et al., 2010, p. 10).

Research into the connection patterns of Korean research institutes highlights the fact that the categorical structure used in (Guimerà et al., 2007a; Velden et al., 2010) fails to capture differences in connection profiles of groups within the network. Lee and colleagues (2012) analyzed 127 institutions in the Astronomy research community on two dimensions—structural positions (density, efficiency, and betweenness centrality) and relational characteristics of individual nodes (eigenvector and closeness centralities) and compared those results to productivity measures. Institutions with higher densities maintain close, highly productive ties, while institutions with lower densities, higher efficiencies, and higher betweenness centrality serve as intermediaries, fostering or coordinating larger collaborative (and perhaps more innovative) efforts.

One thing to note about (Lee et al., 2012) is that no analysis was done at the micro level, thus making it difficult to fully appreciate the roles of individuals in the network. This has two drawbacks. First, it obscures the reality that scientists' roles within their groups differ. Second, it is difficult to make direct comparisons to the results of (Guimerà et al., 2007a; Velden et al., 2010), thus making it difficult to reconcile the rough classification provided by Lee and colleagues with the classification proposed by Guimera et al.

Conceptually, it is also important to note that, for Lee's study, the concept of density was a meso (group) versus macro level variable, with density being equal to the fraction of all possible edges from each group to every other group. Therefore, the concept of density is, in fact, more closely related to the concept of participation as outlined in (Guimerà et al., 2007a). Therefore, it might be reasonable to conclude that groups with high densities may be roughly equivalent to hubs although this cannot be immediately assumed because of the different levels of analysis used between the two studies.

It can also be noted that scientific collaboration is observed to be fractal in nature (Anthony F. J. van Van Raan, 2000), with patterns repeating across levels of analysis. To the extent that this is true, we can say that groups, just as individuals, can play roles in the network. A question that naturally arises from this view is whether or not individuals within certain types of groups are more likely to fulfill certain roles.

Although both studies give some insight into the non-uniform structure of complex networks, they fail to describe the waxing and waning status of groups and actors within the network over time or how those shifting fortunes affect the macro level properties of the network. To that extent, more recent work has turned to looking at the dynamics of collaboration networks at the mesoscopic level. Chang and Huang (2013) measured the network position of research groups in the fields of Astronomy and Astrophysics over an eight-year period using three measures of centrality (degree, closeness, and betweenness).

There are several interesting results from the Chang and Huang study, particularly as they relate to time dynamics of the network. At the macro level, network density increased from 8% in the first time window to 13% in the final time window, with a cumulative density of 19%.

This observation suggests that, for each time window, the active scientists were more increasingly likely to collaborate with a larger range of individuals, and when looked at cumulatively, scientists were likely to engage in triadic closure over time. The same trend held for degree centralization, starting at 50.63% and rising to 57.37%, with an overall centralization of 63.33%.

At the institutional level, degree centrality tended to increase over time, with the mean number of links to other institutions rising from 43 in the first time period to 53 for the final time period. It is important to note that the distribution of collaborators per organization was highly skewed, with roughly 38.5% collaborating with fewer than 50 institutions, 75% collaborating with fewer than other 150 institutions and 1% collaborating with more than 300 other institutions.

With respect to the relationships between the concepts measured, Chang and Huang found that there was a high correlation between closeness and degree centrality, while institutions with high or moderate closeness and degree centrality had lower betweenness centrality scores. The implication of this is that no institution played a dominant role in bringing different institutions together, although some were much more likely to do so than others. Furthermore, for all institutions except for two, positions changed with respect to one another over time. While two institutions had the highest degree centrality for all 3 time periods, all other institutions can be grouped into one of four categories: continually rising, first rising then falling, first falling then rising, continually falling.

	Ranking movement distributions			
	Diff. < 50	Diff. > 100	Diff. >150	Diff. > 200
Degree centrality				
Continually rising	11.22	5.12	2.15	.99
First rising then falling	15.18	6.60	1.49	.50
First falling then rising	15.18	8.09	2.31	1.16
Continually falling	12.71	4.95	1.49	.66
Total	54.29	24.75	7.43	3.30
Closeness centrality				
Continually rising	12.05	6.27	3.14	1.32
First rising then falling	15.18	6.77	3.80	.50
First falling then rising	16.83	7.10	3.14	.66
Continually falling	14.36	5.94	1.82	.66
Total	58.42	26.07	11.88	3.14
Betweenness centrality				
Continually rising	7.59	2.64	.83	.17
First rising then falling	16.17	6.11	3.14	1.16
First falling then rising	15.35	7.76	4.62	1.65
Continually falling	9.90	2.97	.99	.33
Total	49.07	19.47	9.57	3.30

Table 2-5 breaks down the distribution of how institutions fall into each of the four categories for all three measures of centrality, with the distributions further broken down into categories based on the change in the number of positions. Of the institutions that have moved more than 50 positions, two-thirds only moved substantially during one period. The remaining third exhibited the same mobility for more than one time period.

	Ranking movement distributions			
	Diff. < 50	Diff. > 100	Diff. >150	Diff. > 200
Degree centrality				
Continually rising	11.22	5.12	2.15	.99
First rising then falling	15.18	6.60	1.49	.50
First falling then rising	15.18	8.09	2.31	1.16
Continually falling	12.71	4.95	1.49	.66
Total	54.29	24.75	7.43	3.30
Closeness centrality				
Continually rising	12.05	6.27	3.14	1.32
First rising then falling	15.18	6.77	3.80	.50
First falling then rising	16.83	7.10	3.14	.66
Continually falling	14.36	5.94	1.82	.66
Total	58.42	26.07	11.88	3.14
Betweenness centrality				
Continually rising	7.59	2.64	.83	.17
First rising then falling	16.17	6.11	3.14	1.16
First falling then rising	15.35	7.76	4.62	1.65
Continually falling	9.90	2.97	.99	.33
Total	49.07	19.47	9.57	3.30

Table 2-5: Distribution of the types of changes in centrality rankings; column numbers refer to the position change; taken from [Chang and Huang 2013]

Overall, there was a general trend toward greater connectivity, with the number of peripheral and isolated institutions decreasing over time. Interestingly, 70% of the peripheral institutions remain in the periphery for two successive periods, while the remaining 30% were newcomers. This implies that the ultra-peripheral groups rarely have the chance to move toward the center and that topological dynamics are driven by either newcomers or groups already occupying the center vying for better positions.

Overall, we see that Complex Network Analysis has moved from a framework that is used to generate descriptive interpretations of the structural properties of network and some modeling of basic mechanisms that give rise to the observed topological properties, to being a framework that is used to study the evolution of complex systems from a network perspective,

including the system of scientific collaboration that is responsible for generating formal knowledge products.

2.7 Summary

The literature review covered standard operationalization, forms, and units of analysis of scientific collaboration in research studies, as well as general macro trends in scientific collaboration across all fields of research. In terms of the organization of the review, the emphasis was placed on the factors that influence the formation of collaborative relationships (antecedents) and the outcomes of scientists working together (effects). The antecedents and effects of scientific collaboration will be discussed throughout the next Chapter, and will be used to motivate the selection of Complex Adaptive Systems as a framework for the research conducted in this study. To that end, a review of Complex Systems, and the application of Complex Systems to study scientific collaboration were included in this Chapter.

3 Theoretical development

3.1 Complex adaptive social system

This dissertation adopts the theoretical perspective that science is a dynamic, self-organizing social system with complex, nonlinear patterns of interactions between the actors in the system. Science as a complex adaptive social system (CASS) draws upon several research fields, including organizational studies, complexity theory, network theory, and communications theory, and was recently explored in (Mohrman, Galbraith, & Monge, 2006; Wagner & Leydesdorff, 2009). Viewing the production of scientific knowledge as a CASS provides several advantages related to understanding the relationship between the emergence of team science and the changing nature of scientists' interdependence. First, it provides a useful way of summarizing the motivations for and approaches to forming collaborative relationships. Second, the framework natively supports thinking about the relationships between individual actor's actions and the group structure of the community they both function in and contribute to through their actions. Finally, the framework explicitly acknowledges that the system is sensitive to initial conditions and that the state of the system acts as a constraint and reference point for actor's actions (Sawyer, 2005; Wagner & Leydesdorff, 2009).[1]

The CASS framework explicitly relies on spatial and structural metaphors to help explain the emergence of existing configurations of social relationships, how existing social configurations influence future social configurations, and the actors' opportunities to forge those relationships. As a *system,* CASS is comprised of agents and the relationships between them.

[1] Sawyer refers to the idea as bidirectional causality. Because humans are able to make abstractions and communicate about their social structure, we are both influenced by, and can influence, that social structure.

CASS are considered to be *complex* because they are partially ordered and partially random (Gell-Mann, 2002), yet display ordered structural properties. They often exhibit common properties (to varying degrees), including: a power-law distribution of relationship connections (Barabási & Albert, 1999; Newman, 2001c), small-world social distances (Watts & Strogatz, 1998), and clustering (Klemm & Eguíluz, 2002). These properties are thought to arise from basic forces such as preferential attachment (H. Jeong et al., 2003; Merton, 1968, 1988), assortative mixing (Freeman & Huang, 2014; Jones et al., 2008; Newman, 2002), and triadic closure (Easley & Kleinberg, 2010). CASS are *adaptive* because they change over time and respond to internal and external stimuli and conditions (Beckner et al., 2009; Holland, 1992, 2006). Finally, *social* implies a special type of complex system in which the agents are capable of communication and abstract reasoning about their relationships and the structural patterns they form, which in turn influences the evolution of the system (Sawyer, 2005).

Complex adaptive social systems consist of agents, who interact and build relationships around a shared activity (Holland, 1992). Those interactions give rise to structural patterns, that agents can observe, communicate about, and react to. The fact that agents within a complex adaptive social system can observe, communicate about, and react to the emergent structural patterns is the significant differentiator between CASS and other complex systems (Beckner et al., 2009; Sawyer, 2005). The system is considered to be adaptive and dynamic, with agents responding to both internal and external conditions by modifying their patterns of interactions, which in turn results in changes in the emergent structural patterns. The structural patterns reflect the tendency of agents within the system to form groups or clusters of individuals who are more likely to interact with other members of the group than with members of external groups (Arenas, Danon, Díaz-Guilera, Gleiser, & Guimerá, 2004; Guimerà & Amaral, 2005; Newman,

2004a). Because humans are communicative agents who are capable of developing and sharing abstractions about their environment, they can observe and react to the emergent structural patterns, modifying their behavior based on their internal objectives and assessment of the environment. The ability of people to abstract and communicate about historical interactions makes CASS a special type of complex system with *memory*, in contrast to *memoryless* systems that are comprised of agents and their relationships operating under first-order Markov processes. Systems use their memory of past interactions to anticipate future conditions (Holland, 2006). Past relationships and their resultant structural properties, and the communication about those relationships and processes, influence future interactions (Leydesdorff, 2003). However, memory should fade, allowing more recent interactions to influence interactions more strongly than older relationships.

Agents' internal objectives and assessments of the environment are highly variable and hidden from the observer's view. Consequently, at the local level their behaviors seem highly random, yet give rise to relatively stable patterns of interaction. Gell-Man (2002) refers to this as the "edge" of chaos, where complexity is at its highest. The midpoint between the randomness of individual agents and the describable emergent structures requires the most information (in the technical sense) in order to describe the system. Systems that are purely ordered require little information to describe them while those that are highly chaotic cannot be described because there is no underlying pattern to describe—they are entirely random. Describing complex systems involves modeling or estimating agents' responses to emerging structural patterns (H. Jeong et al., 2003) as well as network mechanisms that influence their patterns of interaction. However, it is not considered possible to determine the causality of the actions of any individual within that system because of the randomness at that level.

Finally, every agent has a *position* with the structure of the system, where structure refers to the patterns of relationships formed between the interacting agents. An agent's structural position due to historical interactions is thought to influence that agent's future interactions because (a) the system and its agents have memory, and (b) they use that memory to guide their future interactions. Agents are capable of drawing on their direct and indirect networks (Wagner & Leydesdorff, 2009) to locate partners to interact with. It is possible to explore the structure of the system at different levels of analysis, including the macroscopic, mesoscopic, and microscopic levels. This dissertation focuses on (a) the mesoscopic level of the networks because it maps directly to the organizational structure of scientific fields (Ziman, 1994) and (b) the agents' positions within that structure because the positional descriptions provide a richer description of the variety of relational configurations actors have within the network (Guimerà & Amaral, 2005).

Concept	Maps to
Agents	Scientists
Shared activity	Knowledge production & publication
Relationships	Collaboration
Structural patterns	Functional research groups (see 3.3.2)
Structural position	Configuration of relationships between groups
Adaptive	Changing over time
Memory, decay of	Half-life

Table 3-1: CASS concepts and their mappings to scientific collaboration

Table 3-1 shows the mappings between the core concepts of complex adaptive social systems and the phenomenon of scientific collaboration. The remainder of this chapter is structured around those mappings, highlighting limitations of the CASS framework and identifying the subsequent research questions and hypotheses in the relevant sections.

3.2 Agents and shared activity

Complex Adaptive Systems (CAS) are comprised of agents and their interactions around a shared activity. Complex Adaptive Social Systems, as a specific subclass of CAS, explicitly assume that the agents are communicative and capable of abstract reasoning about their social order. If scientific knowledge production is the shared activity, then scientists are the agents interacting around the production of that knowledge. This dissertation focuses on the formal interactions that underpin scientific knowledge production, working under the assumption that the formal set of collaborations reflect the informal interactions that also contribute to the broader global system of science. In essence, the formal knowledge production system is a subsystem of the global science system, which has other subsystems that have evolved to support education, outreach, and internal governance.

Looking at scientific collaboration through the lens of CASS involves making certain assumptions about the motivations of the system's participants. First, it is assumed that the norms of science influence the scientists, who in turn are motivated to contribute to the global body of knowledge (Merton, 1973). The second assumption is that scientists are committed to the reputational system of science, and, therefore, are engaged in a search for resources, recognition, and rewards (Whitley, 2000) in an environment characterized by limited resources (Axelrod, 1997). The final assumption is that it is necessary to collaborate to be a part of the system; it is no longer possible to maintain a successful scientific career in most fields without collaborating because the complexity of research requires integration of multiple specialties (Bozeman et al., 2001). Therefore, scientists must be willing to collaborate, which is the fundamental interaction within the system of formal scientific knowledge production.

Scientific collaboration, as an interaction, is the formation of a short-term relationship. Research projects take some time to complete, and, therefore, the relationship is sometimes thought of as a commitment (Hara et al., 2003; Melin, 2000). The formation of a collaborative relationship involves a search process and the application of selection criteria. The next section focuses on the formation of relationships around that shared activity, starting with a basic, dyadic model of relationship formation. It highlights some of the weaknesses of that model in scientific fields dominated by research groups, and then follows up with an explication of how the research group changes the model.

3.3 Agent interaction: Models of scientific collaboration

Two models of scientific collaboration will be outlined and discussed in this section. The first model is a basic model of scientific collaboration assuming dyadic interactions between scientists; the second, a model that incorporates the influence of the group on collaborative interactions. The dyadic model is the simpler of the two models and is implicitly used in the existing literature on scientific collaboration networks. There are several limitations to the dyadic model, as it (a) assumes that the shared activity is organized around dyadic interactions, and (b) does not recognize the influence of the local community on agents' interactions. The first of the two limitations is more significant, as most models and simulations are based on the assumption of dyadic interaction. The core thesis of this dissertation is that the system of scientific knowledge production is not dyadically coordinated, and is instead coordinated by more established agents who can facilitate access to cognitive, economic, technical, and labor resources to get research done. The two limitations are intertwined, with the latter a derivative of the former because the coordinating agents prefer some stability within their team to reduce the burden of coordination. The established agents benefit from a stable group structure that

probabilistically constrains the actions of the less well-established agents. The second model based on the influence of the group structure in collaboration networks, which this dissertation explores, addresses the limitations of the former model by exploring soft constraints on collaboration imposed by the group structure of scientific fields.

Both models share common assumptions regarding the general motivations for collaborating and general trade-offs in the selection of collaborative partners. The first assumption, based on Merton's (1973) norms of science, is that scientists are motivated to make formal contributions to the collective body of knowledge through publications. Scientists also seek acknowledgment for their contributions, and thus prefer to be included as a formal coauthor on papers because science is a reputation-based system, and authorship is one of the most important ways of establishing a reputation (Whitley, 2000). The second assumption is that scientists try to optimize (not maximize) their efforts, balancing effort, risk, and potential impact of the research project. Research projects are inherently risky (Hara et al., 2003) and require significant coordination overhead, particularly when the participants have never worked together before. The additional risk and overhead are balanced against the potential payoff, as the mixing of knowledge through new collaborations has the potential to produce significant innovations (D. H. Lee et al., 2012; Whitfield, 2008). In contrast, the same researchers found that working with established partners tends to increase productivity. The final assumption is that collaboration is necessary to make meaningful research contributions (Bozeman et al., 2001; Parker & Welch, 2013; Wuchty et al., 2007).

3.3.1 Dyadic Model

The basic modelof scientific collaboration, created for this dissertation, involves two scientists, one of whom is the *instigator*, the other of whom is the *target*. The exact nature of who is the instigator and who is the target is not a critical component of the theory used in this research. In some cases scientists will argue that the process is more organic with both parties coming to the conclusion that a collaboration might be useful. However, it is safe to assume that someone has to suggest the collaboration first and that the other steps in the process unfold quickly. Once the instigator has decided to begin looking for collaborators, s/he engages in a series of local and global searches (Wagner & Leydesdorff, 2009) for potential collaborators. Once the instigator identifies potential collaborators, s/he applies a set of *selection criteria* to vet those scientists. Next, the instigator selects and approaches the target scientist, suggesting a potential collaboration. Finally, the target applies *filtering criteria* to assess the instigator and determine whether to accept or reject the offer (Figure 3-1). In some approaches to modeling network dynamics, process is simplified to a node engaging in a ‘unilateral initiative [in proposing the relationship] with reciprocal confirmation [from the target]’ (Bunt & Groenewegen, 2007).

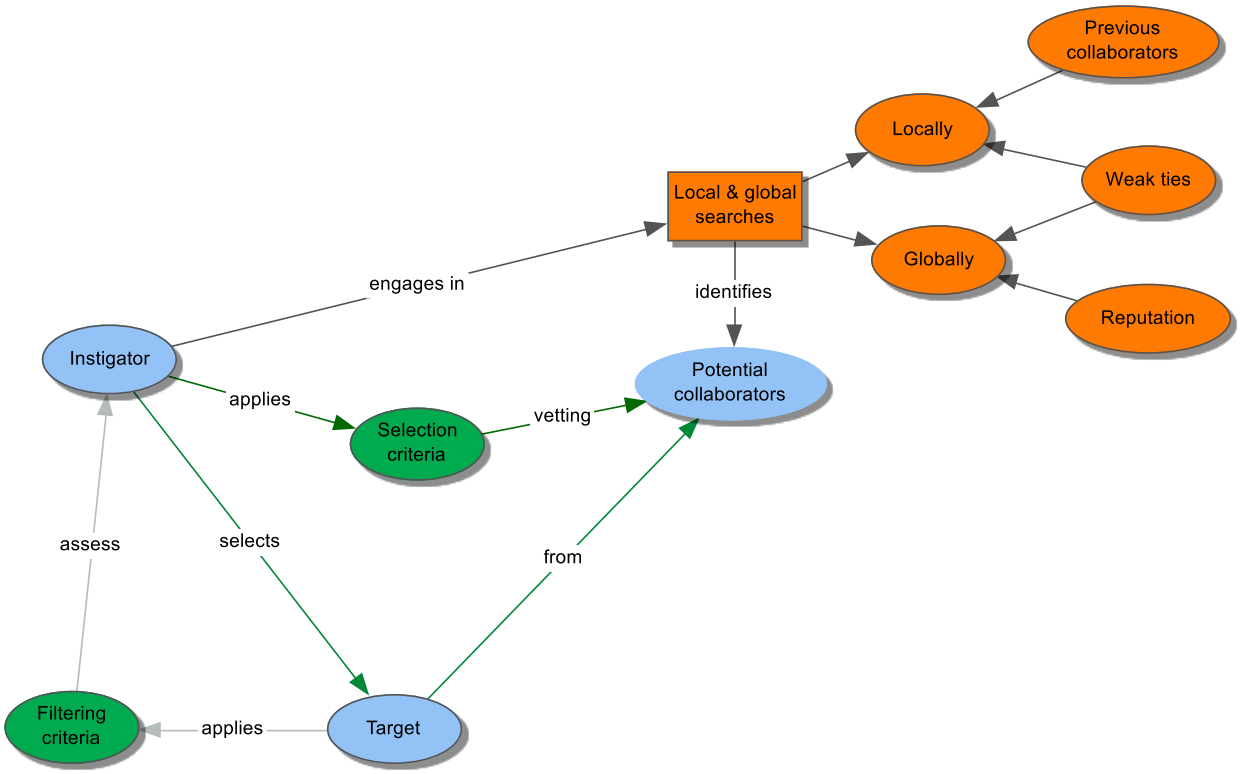


Figure 3-1: The basic model of scientific collaboration

An important point to consider is that the selection criteria may be used to restrict the search process, although it is argued here that the net effect is the same—the selection process is determined by the integration of search and selection. For example, prior research has found that scientists will limit whom they collaborate with to those with whom they've collaborated before (Jansen, Görtz, & Heidler, 2010). Limiting one's search to prior collaborators is a local search from the network perspective in that the connection is pre-existing. The instigator has pre-applied one of the selection criteria. Prior research also demonstrates that scientists apply numerous criteria, sometimes consciously, sometimes subconsciously, to the partner selection process (Bozeman & Corley, 2004; Melin, 2000) and leverage random interactions to generate leads (Jansen et al., 2010). As Beaver (2001) notes, scientific collaboration appears to be

completely random at the individual level while simultaneously exhibiting stable patterns at higher levels of aggregation.

3.3.2 Group model

One of the core concepts in Complex Adaptive Systems is the notion of emergent structures—stable patterns of organization above the level of the individual. The emergence of stable patterns through the interactions of the individual agents distinguishes complex systems from chaotic systems—in chaotic systems there is a tendency toward disorder, and in complex systems, there is a tendency toward order at the edge of chaos (Gell-Mann, 2002). These structural patterns arise from the individual interactions of the agents comprising the cluster, who tend to interact more frequently with one another than with agents outside of the cluster. The decision to interact within the group, or between groups, need not be conscious, but can arise from simple rules of interaction. Furthermore, the emergence of clusters is as much a function of the interaction of the constituent agents as it is of the external agents, who by their actions and inactions help form the boundaries of the cluster.

Complex systems may exhibit multiple levels of order (Heylighen, 1989), where higher levels of order supervene upon lower levels of order and take on similar structures such that they appear to be fractal in nature (Anthony F. J. van Van Raan, 2000). From a network perspective, one would see clusters form in the interactions of agents, and then several of those clusters would be weakly tied to one another through agents who travel back and forth between groups. The extent of interactions between the groups in the larger cluster would be greater than interactions to groups outside of the larger cluster, creating a region with fuzzy boundaries that has two or more levels of order contained within it.

The emergent clusters map to the functional research group in the sciences (Seglen & Aksnes, 2000). Functional research groups consist of a set of core members and a rotating cast of external collaborators and visitors. The core group usually consists of a small number of established scientists, several postdoctoral researchers, and some graduate students. The core members of the group may have formal ties (e.g., employment contracts), or may be informally affiliated yet bound through frequent interaction. In contrast, the rotating cast of researchers comprises visiting researchers and short-term collaborators. In alignment with the concept of cluster or group in complex networks and complex systems, the research group has fuzzy boundaries. Each group has a core set of researchers, as well as scientists at the periphery of the group who have ambiguous status regarding group membership (Calero, Buter, Valdés, & Noyons, 2006; Perianes-Rodríguez, Olmeda-Gómez, & Moya-Anegón, 2010).

As an organizational structure, research groups offer several advantages over scientists acting as independent actors. The advantages are particularly important when the relative complexity of modern research is taken into consideration (Hara et al., 2003). First, research groups facilitate the acquisition and sharing of expensive equipment (Hackett, 2005). Second, research groups also facilitate the development of coordination practices that underpin successful research projects and ongoing knowledge production efforts (D. H. Lee et al., 2012). Third, the stability of personnel associated with groups reduces the costs and uncertainty associated with locating and obtaining access to scientists with specific expertise. At a basic level, this means that scientific research requires a certain amount of effort and attention from the participants, and reducing the uncertainty associated with obtaining that effort and attention makes it easier for researchers to plan projects.

One of the main advantages of the research group is that it reduces the uncertainty when accounting for the efforts and attention of the members. Within this framework, collaboration is assumed because it is not possible to conduct meaningful research as an isolated individual (Bozeman et al., 2001). However, the ways in which the group influences the collaborative patterns of scientists is still unknown. In particular, this dissertation argues that the presence of the research group impacts scientific collaboration in two meaningful ways. First, research projects are more resource intensive, and resource control is centrally managed. Therefore, established scientists with the capacity to bring together the social and technical resources serve as gatekeepers. A secondary impact of the rise of the gatekeeper in scientific research is that the search process for potential collaborators changes to reduce the burden on the gatekeeper.

The role of the group can be explored from both the perspective of the gatekeeper and the junior scientist trying to get on a project. The PI or group leader is the gatekeeper in this scenario. The PI's goal is to produce meaningful research, and to that end wants to identify scientists with the relevant expertise as well as junior researchers who can perform the research tasks under the direction of the experienced scientists on the project. The PI will want to rely on local labor (e.g., doctoral students) to perform the guided labor because their efforts can be accounted for, and searching broadly for talent will provide little additional benefit. Instead, the quality of the student's work is more likely to be influenced by the ability of the scientist guiding the work than the student's knowledge and expertise. Group leaders will also want to bring relevant knowledge and expertise in-house to be a core part of the group's research portfolio. If the PI or group leader is running multiple projects, the junior scientists and students will be expected to focus on one project, while more established scientists in the group with specific skill sets will work on several of the PI's projects, depending on where their skill set is needed.

When the PI looks for external collaborators to solve problems, he or she is looking for a scientist who has a team that can solve those problems. Once again, this is based on the argument that the research group is the fundamental unit of organization in the sciences, and each group has an area of expertise that can be combined with other groups' expertise to generate knowledge (Moed, 2006). The implication of this approach to generating knowledge is that junior scientists are less likely to be searched for as independent entities, but instead are recruited into projects based on the reputation of the group and the group leader. Beaver (2001) alluded to this pattern of activity when he quoted the scientist who remarked that junior scientists are more likely to be unknown to the broader research community, and instead are fractional authors on papers associated with the group leader whose name is on the paper.

From the junior scientist's perspective, their own search efforts are functionally limited to the local group because they are expected to focus on their group's projects. No one outside of the group will seek them out because they have no reputation, so he or she is more dependent on the group and group leader for opportunities to participate in research projects. It may be possible for the junior scientist to build up a network of connections within and between research groups if their home group actively collaborates with other groups. Those connections help the scientist establish a professional identity and exposes them to more opportunities to participate in research projects as their reputation for having certain expertise spreads. Scientists with an established reputation are more likely to be a target in the search process, or, at least, easier to find.

Up until this point, the argument is that the research group has become the fundamental unit of organization in the sciences and that scientific collaboration is more about assembling teams from multiple groups to work on specific research problems or projects. Under the team-

based regime, research is more likely to be coordinated by established scientists who have the capacity to bring together the people and resources to conduct the research. Furthermore, the established scientists are more likely to favor some level of organizational stability to make the process of coordinating projects easier. The implication of this organizational structure is that junior scientists are dependent on established scientists to provide access to research opportunities because the established scientists manage the technical and human resources needed to conduct research. However, it was also noted that more established scientists may be dependent on the junior scientists to perform the work, as their presence reduces the burden associated with ensuring there is sufficient human capital to perform the research.

Within the framework outlined above, *dependence* is the extent to which one scientist relies on another scientist to either (a) coordinate and provide access to research projects, or (b) perform the work needed to ensure the successful completion of the research project(s) he or she is given access to. This dissertation focuses on scientists' dependence on one another, exploring the question—**How is the increasing prominence of the research group and team-based research impacting scientists' dependence on one another and the research group?** It was noted earlier that a scientist's position within the group structure might influence their opportunity to participate in research projects, either because those connections are a reflection of the scientist's being established in the community or because having more connections to other groups improves the scientist's findability. There is an expectation that a scientist's position within the group structure should influence their dependence on other scientists for opportunities to participate in research projects because that position is a reflection of their social capital, or resources available through the relationships they've built (Burt, 2001; Nahapiet & Ghoshal, 1998). The expectation that scientists who have connections both within and between

research groups have more opportunities to participate in research projects leads to the second research question—**What is the relationship between a scientist's distribution of relationships within the group structure and their dependence on other scientists, and how has the relationship between distribution and dependence changed over time?**

Fundamentally, the framework alters the search process in the dyadic model of scientific collaboration. If the instigator is a PI, his or her primary concern is assembling a team of researchers who can do the work needed to make the project successful. That includes the "worker bees": lower-skilled undergraduate and graduate students, moderately skilled postdocs and early career professionals, and other teams that bring the requisite resources to the project. The PI will limit his or her search for the lower skilled labor to his or her lab, and will likely draw on the postdocs or junior faculty in the lab if they are available. PIs attempt to cultivate a local labor force that is dependable in the sense that they can execute the tasks assigned to them and they are available to do the work; basically, it takes less cognitive effort for the PI to arrange to put lower skilled bodies on a project, freeing up time to search more widely for expertise that is harder to find. When the PI searches broadly for the needed expertise, he/she may follow the dyadic model—looking for prominent scientists who fulfill the needed requirements. However, if the scientist finds a potential collaborator, that collaborator may bring his or her research group into the collaboration, so that they can assist with the research.

What we see is that the dyadic model still holds in the sense that it accurately depicts the way the prominent scientists search for collaborators. However, it breaks down when we see that other, less well-known scientists are brought into the collaboration to provide the labor needed to do the bench work (Beaver, 2001; Stephan, 2012). For the less well-established scientists, their search process is limited to the group because they lack the social standing and expertise to

barter for access to projects in other groups. That, and the PI they work for may use their employment status as a lever to direct their work efforts. No one outside of the research group is going to search for them because they do not have the reputation for specialized skills that are worth the effort of tracking down. As Beaver (2001) said, they are unknown to the community. The scientists with little established reputation are dependent on the PI to provide access to research projects that he or she organizes, or is invited to participate in. That access is made in exchange for a commitment to do the work. One way of summarizing the exchange is: "I [the PI] will let you [the junior researcher] participate on this project if I can depend on you to do the work." The question is: Can we estimate scientists' dependence on one another based on the relationships they have within and between the research groups? Will looking at dependence through the lens of distribution of relationships in the group structure give us a way to tease out differences in dependence, and maybe in future studies, provide a framework for teasing out qualitatively different types of dependencies?

3.4 Summary

The complex adaptive social system framework is a useful lens to explore the dynamic nature of scientists' collaborative interactions at the community level. Within the framework, scientists are treated as autonomous agents who interact around a shared activity—in this case the production of scientific knowledge. The framework explicitly acknowledges that the actions at the individual level are seemingly random (Beaver, 2001; Gell-Mann, 2002), yet give rise to relatively stable patterns of organization. In turn, the stable patterns of organization are thought to influence the interactions that they are built from (Ladyman et al., 2012a; Wagner & Leydesdorff, 2009). The pattern of interaction at the individual level and the individuals' responses to the emergent structural patterns is a form of bidirectional causality (Sawyer, 2005).

A basic model of the formation of scientific collaboration, built around the dyadic interactions of scientists was outlined. The formation of relationships within that model is dependent upon the interactions between, and outcomes of, a search process (Wagner & Leydesdorff, 2009) and the application of screening criteria. An argument was made that, from a sociological perspective, the basic unit of organization within the sciences is the research group. Furthermore, the nature of scientific research has been shifting toward a team science environment (Seglen & Aksnes, 2000; Velden et al., 2010; Wuchty et al., 2007), where scientists from multiple groups often get together to form teams that work on projects. The way in which these team projects are organized act as a functional constraint on scientists' opportunities to collaborate. An argument was made that scientists face different concerns regarding their participation in the knowledge creation process, and those concerns influence their dependence on other scientists. Less well-established scientists seek access to opportunities to conduct research, while more established scientists either try to leverage their reputation to gain more opportunities, or focus on ensuring the success of research projects and their group's general capacity to maintain productivity.

All of this activity takes place in an evolving system, where the general trend is toward more team-based research in an environment dominated by research groups. Several related questions emerged from this line of reasoning. The first question was—How is the increasing prominence of the research group and team-based research impacting scientists' dependence on one another? The expectation is that the nature of dependence might vary based on the scientists' distribution of relationships within the group structure, which reflect both their accumulation of connections that facilitate the search process as well as their access to the resources embedded in those relationships (Lin, 1999). The second question—*What is the relationship between a*

scientist's distribution of relationships within the group structure and their dependence on other scientists, and how has the relationship between distribution of relationships and dependence changed over time? focuses on the changing nature of science while also exploring whether scientists' dependence can be differentiated based on the idea that position within the group structure reflects both opportunities and concerns.

4 Methodology

4.1 Overview

This dissertation is an exploratory study, which is an appropriate choice given the fact that there are few examples in the literature to guide the research, and the theory did not provide sufficient guidance on the expected relationships between concepts to generate a testable hypothesis (Schutt, 2006). Specifically, the concept of dependence has not been tested within a complex systems framework, which typically treats the ability to form relationships as a dyadic interaction and not a brokered interaction between multiple parties. Although this study was exploratory in nature, the complex systems framework did provide two potential methodological approaches: simulations and complex network analysis. Simulations are traditionally associated with complex adaptive systems, where the researcher focuses on identifying and testing simple rules of interaction at the individual level that will produce observed aggregate behaviors (Holland, 2006). In contrast, complex network analysis (CNA) focuses on traces of interactions between agents in the system, and describing, analyzing, and modeling the emergence of the structural properties of those traces at different levels of aggregation (Barabási et al., 2002; H. Jeong et al., 2003; Newman, 2001a).

This dissertation used the latter approach to studying complex systems because it provides a rich set of concepts, models, and techniques to support exploratory analysis. Researchers use CNA to study actors' positions within the structure emerging from their interactions and how those positions influence their future interactions (Abbasi et al., 2011; Hill, 2008; Larivière, Gingras, & Archambault, 2013). Temporal network analysis, which was used in this dissertation, is an extension of CNA that focuses on the dynamic and evolving nature of

networks (Holme & Saramäki, 2012). The specifics of that process are described in the next section.

4.2 Operationalization of concepts

This section outlines the operationalization of core concepts, as well as a discussion of the rationale for the choices made. It starts with the concepts that are central to the network analytic framework before discussing the methods used to study the temporal dynamics of networks. The section concludes with a discussion of the limitations of current approaches to studying temporal networks, which is then used as a motivation for the proposed experimental test of the different approaches.

4.2.1 Scientific collaboration

Scientists can collaborate in many ways; De Haan (1997) identified six ways scientists can collaborate—coediting a publication, sharing supervision of Ph.D. projects, coauthoring a proposal or publication, participating in formal research projects, and organizing conferences (from Mali, Kronegger, Doreian, & Ferligoj, 2012). In this dissertation, scientific collaboration was operationalized as coauthorship of a research article under the assumption that if two scientists coauthor a paper together, they have collaborated on the related research. Several researchers are critical of operationalizing collaboration as coauthorship (Laudel, 2002; Melin, 2000), as it both undercounts and overcounts instances of collaboration. Not only does it miss five forms of collaboration identified by De Haan, it also misses informal collaboration that does not warrant shared authorship of papers (Cronin, Shaw, & La Barre, 2003). Coauthorship sometimes also overstates collaboration, particularly when honorary coauthorship is given.

Although operationalizing collaboration as coauthorship has its disadvantages, it still is one of the most effective approaches to studying scientific collaboration on a large scale (Glänzel & Schubert, 2005) because other methods of identifying instances of collaboration (e.g., through surveys) are unreliable. Operationalizing collaboration as coauthorship also results in the measurement of one of the most important activities in science—publication. Scientists, as professionals (Beaver & Rosen, 1978), are expected to be productive and contribute to the shared body of knowledge (Merton, 1973) and are rated on their productivity. Publications serve as markers of expertise are integral to the reputation and reward systems of science (Whitley, 2000).

4.2.2 Collaboration network

Looking at scientific collaboration through a network lens involves identifying entities and the relationships between those entities. From a visual perspective, entities can be depicted as points or circles, which are connected via lines when a relationship is present between two entities. Entities are referred to as *nodes* or *vertices*, and relationships are referred to as *edges*. The entire set of nodes and edges constitutes a *graph*. In a scientific collaboration network, the entities can be individuals or aggregations of individuals (e.g., Chang & Huang, 2013; Newman, 2001c), but the relationship is always a collaboration.

Network analytic approaches are powerful tools for studying large-scale communities, as evidenced by the rise in popularity of the approach for studying scientific collaboration. The measurements and models associated with network theory can be used as lenses for studying the general distribution of relationships (Barabási & Albert, 1999; Ding, 2011; Newman, 2001c), the prominence of actors in the community (X. Liu et al., 2005; Newman, 2004b), the relative

advantages of certain positions within the community (Abbasi, Hossain, & Leydesdorff, 2012; Bonaccorsi, 2008), modeling the growth of communities (H. Jeong et al., 2003), and tracking knowledge diffusion across social networks (when used in conjunction with citation analysis) (Ozel, 2012a, 2012b).

There are two ways to operationalize a collaboration network: either as a unimodal network where all relationships are dyadic in nature (Barabási & Albert, 1999; D. H. Lee et al., 2012; Tomassini & Luthi, 2007), or as a bipartite or affiliation network (Guillaume & Latapy, 2004; Guimerà, Sales-Pardo, & Amaral, 2007b). In the latter operationalization, there must be at least two types of nodes—actors and activities or organizations. In bipartite networks, all relationships exist between the two types of nodes. It is possible to transform bipartite networks to unimodal networks, but the reciprocal is not true as certain information is not encoded in unimodal representations of networks. The majority of this dissertation builds off of a large body of literature that uses unimodal projections of networks; exceptions to this are explained later. There are drawbacks to viewing team science as a unimodal network; specifically, unimodal networks assume relationships are formed between individuals. That assumption is not valid in team-based research, and violations of that assumption have practical considerations. Using bipartite projections to identify the group structure of networks is more effective than unimodal projections (see Chapter 5).

4.2.3 Dependence

Dependence is defined as the extent to which one scientist relies on another scientist to either: (a) provide access to research equipment, skill sets, and resources, or to coordinate projects, or (b) perform the work needed to ensure the successful completion of the research

projects(s) he or she is given access to. By definition, collaborative relationships are considered to be symmetrical at the dyadic level; however, the symmetricity of the relationships masks what each participant offers in the relationship. Some scientists have access to technical, economic, and cognitive resources, either under their direct supervision or through their professional networks. Other scientists may only be able to offer specialized skill sets or labor. Both parties want to produce research; scientists in the former category need people to do the work or fill in skill gaps while scientists in the latter category need people to bring the resources and people together to make complex research projects possible. There is no implied seniority in the concept of dependence—it is possible for a senior scientist to be dependent on a junior scientist to publish in an area because the junior scientist has specific expertise (e.g., computational analysis), or because the senior scientist prefers not to serve as a coordinator.

A scientist's dependence on another scientist was operationalized as the portion of the scientist's papers that the other was a coauthor on. Dependence is a continuous variable, both conceptually and operationally. A scientist can depend on another scientist significantly or very little, depending on how often he or she works with the other scientist. To make this more concrete, a publication list of two scientists drawn from the data used in this dissertation (§4.4) for the years 1982–2003 is provided below (Table 5-10). The scientist on the left side of the table was the more senior of the two scientists, the scientist on the right was the senior scientist's postdoc. *Scientist 1* first published in 1982, and had three publications before 1985 while *Scientist 8466* also first published in 1982 and had one publication before 1985. Using the operationalization provided above, *Scientist 1* had a dependence score of 0.25 toward *8466*, while *Scientist 8466* had a dependence score of 1.00 toward *Scientist 1* at the end of 1982.

Author ID	Publication ID	Year	Author ID	Publication ID	Year
1	1	1982	8466	305	1982
1	305	1982	8466	4526	1985
1	790	1982	8466	5030	1985
1	4358	1985	8466	6462	1986
1	4729	1985	8466	6601	1986
1	5523	1986	8466	10474	1987
1	6874	1986	8466	40804	1994
1	7018	1986	8466	71575	1997
1	8453	1987	8466	141185	2003
1	13650	1988			
1	15625	1989			
1	21277	1990			
1	82516	1998			
1	84177	1999			

Table 4-1: Publication list for two authors

Two things to note here—first, all scientists are interdependent to some degree because all relationships go two ways, only the strength differs between directions. Second, scientists' dependence on one another can change over time if they follow different research paths or begin working with different research groups.

4.2.4 Research groups

The concept of research group refers to the functional groups that serve as the foundation of modern science (Seglen & Aksnes, 2000). Functional groups can consist of one or more senior scientists, several junior researchers, and doctoral students. The core of the group may be bound together by formal affiliation. In addition to the core members of the group, many research groups have rotating members, scientists who either collaborate frequently with the group, perform one-off collaborations, or visit for an extended period of time in to conduct research on specialized equipment or provide specialized expertise (Velden et al., 2010).

The concept of the research group is operationalized as a *module* identified with a *community detection algorithm* in a network. In complex network analysis, communities are clusters of actors who are more likely to interact with one another than with actors outside of the cluster (Guimerà & Amaral, 2005; Lancichinetti & Fortunato, 2009; Newman & Girvan, 2004). Each community is referred to as a module; reliably and validly identifying modules is an ongoing area of research. A persistent challenge is determining exactly where the boundaries of a module should be (Danon, Duch, Diaz-Guilera, & Arenas, 2005; Lancichinetti & Fortunato, 2009). In many instances, placing scientists in a module is a relatively straightforward process because the density of the relationships between the members far exceeds the density of connections to other scientists. However, in boundary cases, it can be difficult to determine exactly where a scientist belongs because that scientist's connections are distributed nearly evenly to many groups. Although the process has some margin of error, qualitative follow-up on the use of community detection algorithms suggests that they perform well (Velden et al., 2010).

Infomap (Rosvall et al., 2009) was used for identifying the modular structure of the network in this dissertation. Infomap, which is in turn based on the map equation, is an information theoretic approach to community detection in networks. The Infomap algorithm (Rosvall et al., 2009; Rosvall & Bergstrom, 2007) uses a random walk method to identify scientists on common paths, then proposes solutions by clustering scientists into modules, encoding their location using Huffman codes, and evaluating the solution by assessing its ability to reduce the information needed to encode the location of scientists in the network. This dissertation used the Infomap algorithm versus algorithms designed to maximize modularity (Newman & Girvan, 2004) for two reasons: First, its performance has been demonstrated in prior research (Lancichinetti & Fortunato, 2009; Velden et al., 2010). Second, from a theoretical

perspective, the random walk approach models the search for collaborative partners in that the search process involves asking other scientists for information regarding frameworks and collaborators in the search for answers to research questions; the search process often involves several steps, as scientists are iteratively guided toward the person or information they seek.

Compression of information is possible because information in a network tends to flow through certain nodes more frequently than others, such that many nodes are most frequently and easily reached through the information conduits. Thus, the most effective way of encoding the location of scientists that do not serve as a conduit for information is to nest them under a more conductive scientist, which is what the Huffman codebook does in order to compress the information required to encode the location of scientists in the network. Beaver's (2001) observation that scientists are rendered invisible in team science environments matches with the results of the information-theoretic algorithm—most information flows through the prominent scientist in a module, and most scientists operating in the modules can only be found through the prominent scientist.

4.2.5 Distribution of relationships within the group structure

The distribution of relationships within the group structure is intended to be an estimation of the scientist's social capital (Burt, 2001; Nahapiet & Ghoshal, 1998), where we expect the past relationships of a scientist to reflect access to useful resources in the network. Traditionally, in microscopic network analysis, estimation of actors' social capital is done through centrality measures (e.g., betweenness, eigenvector, closeness, alpha) (Borgatti, Jones, & Everett, 1998). However, looking at the positions of scientists based on centrality measures ignores the modular structure of networks described above (Guimerà & Amaral, 2005). The modular, or group,

structure is important because resources are aggregated at the group level in scientific fields (Stephan, 2012; Ziman, 1994), so connections within and between different groups are more important to track than connections to individuals.

When the modular structure is taken into consideration, scientists have two distinct types of connections—intra-module and inter-module. Intra-module connections are to other scientists within the module, reflecting the scientist's connections to the members of their functional research group. Inter-module connections reflect the breadth of a scientist's connections to other modules within the community. Scientists can be classified into *Roles* based on the distribution of their ties within and between research groups (Guimerà & Amaral, 2005). The extent to which a node is connected to nodes in its own module is called *within-module degree* (z_i). The extent to which a node balances its connections to its own module and connections to external modules is its *participation coefficient* (*P*). The participation coefficient will tend toward zero (0) as the distribution of relationships moves toward being solely intra-module, and will tend toward one (1) as links become more evenly distributed amongst modules. The within-module degree is normalized by the rate at which all other scientists within the scientist's group collaborate with one another and will tend toward zero if the scientist has far less intra-module activity than other scientists in the module.

Scientists are classified into one of two categories and one of seven roles based on their within-module degrees and participation coefficients (see Table 5-10). Scientists with low within-module degrees, classified as non-hubs, fit into one of four roles depending on their participation coefficients. Peripheral and ultra-peripheral scientists have low *P* and z. Satellite and kinless scientists are not as strongly connected to their home modules as peripheral scientists or any of the hubs, but they interact with many external groups. Hubs are intra- and inter-

modularly well connected; connector and global hubs have connections to many modules throughout the network.

		P	z_i
NON-HUBS	**(R1) Ultra-peripheral nodes**	$P \leq 0.05$	< 2.5
	(R2) Peripheral nodes	$0.05 < P \leq 0.62$	< 2.5
	(R3) Satellite connectors	$0.62 < P \leq 0.80$	< 2.5
	(R4) Kinless nodes	$P > 0.80$	< 2.5
HUBS	**(R5) Provincial hubs**	$P \leq 0.30$	>= 2.5
	(R6) Connector hubs	$0.30 < P \leq 0.75$	>= 2.5
	(R7) Global hubs	$P > 0.75$	>= 2.5

Table 4-2: Node role assignment based on the Participation coefficient (P) and within-module degree (z_i)

The module assignments were established by the calculation of the weighted within-module degree (z) and participation (***P***) coefficients for each scientist. The definitions and equations are taken from (Guimerà et al., 2007a) and provided below.

Participation coefficient—measures the extent to which a node connects to other modules outside of its own module. The participation coefficient is equal to the difference between one and the sum of the number of edges (k_s^i) from node *i* to nodes in module (*s*), divided by the total degree of node i (k_i), squared. The participation coefficient will tend toward zero as the proportion of edges within the module increases, and will approach one as its links become uniformly distributed among many modules.

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{k_s^i}{k_i} \right)^2$$

Equation 1: Participation coefficient

Within-module degree—is a *Z*-score, measuring the extent to which node *i* is connected to nodes within its own module, relative to other nodes in its module. The calculation is based on the difference between the number of links ($k_{S_i}^i$) of node *i* to nodes within its module and the

mean number of within-module links for all other nodes within that module ($\langle k_{S_i}^j \rangle_{j \in S_i}$), normalized by the square root of the difference between the mean of the squares of all within-module links ($\langle \langle \left(k_{S_i}^j\right)^2 \rangle_{j \in S_i}$) and the mean of the within module links squared ($\langle k_{S_i}^i \rangle_{j \in S_i}^2$).

$$z_i = \frac{k_{S_i}^i - \langle k_{S_i}^j \rangle_{j \in S_i}}{\sqrt{\langle \left(k_{S_i}^j\right)^2 \rangle_{j \in S_i} - \langle k_{S_i}^i \rangle_{j \in S_i}^2}}$$

Equation 2

4.2.6 Adaptive systems and temporal evolution of the network

Change is a constituent of complex systems; although there are many definitions of complex systems, all of them explicitly include some form of interaction and change (Ladyman, Lambert, & Wiesner, 2012b). Complex systems may differ in what drives that change—for complex adaptive systems, the elements of the system are engaged in a continuous process of adapting to the patterns of the organization they create through their interactions (Arthur, 1999; Leydesdorff, 2003; Sawyer, 2005). Furthermore, complex systems have memory (Goh & Barabási, 2008), where past relationships and interactions influence future decisions and interactions of the agents in the system. Arguing that the production of scientific knowledge is a complex adaptive system means that scientists are engaged in a process of knowledge production which involves a continual process of relationship maintenance based on their perceptions of their existing and historical relationships (Wagner & Leydesdorff, 2009).

The goal, from the methodological perspective, was to track the evolution of the network and identify the group structure that was influencing the collaborative interactions of the scientists in the community. Achieving that goal was particularly challenging because the study

of temporally evolving networks is a relatively nascent field of study with few established methodological procedures (Holme & Saramäki, 2012). The general approach is to use time windows: slicing the network into pieces and analyzing each piece as an independent, static structure (ibid). It is not an exact method, but is considered to be an effective approach for studying slowly evolving networks like collaboration networks (Aggarwal & Subbian, 2014).

Determining the appropriate sampling frequency for studying the dynamics of slowly evolving networks is particularly challenging and outside of the scope of this dissertation. Instead, this dissertation focused on trends over time, which makes polling at larger intervals acceptable. Selecting polling rates, in either case, is difficult and still subject to a researcher's discretion. In terms of building the network, there are two approaches, one based on cumulative networks (Barabási & Albert, 1999; Newman, 2001b) and the other on effective networks (Brunson et al., 2013; Tomassini & Luthi, 2007). Cumulative networks look at all historical relationships until the point the network, or network slice is being analyzed. In comparison, effective networks identify a suitable time window and only use the interactions within the time window to recreate the network slice.

Cumulative networks are easy to implement, will not result in erroneously dropping actors from the network, and are capable of highlighting cumulative advantage obtained over many years of activity. However, cumulative networks give equal weight to all relationships, regardless of their age or period of dormancy. Effective networks are easy to implement as well and will only capture recent relationships, but may drop scientists who are temporally dormant and will provide little insight into which scientists are well-established through years of collaborative interactions. The recency issue was of particular concern to the research in this dissertation because the process of extracting the group structure of the networks relies on the

strength of the connections between scientists to determine whether they are in the same group. The presence of historical links makes it difficult for the community detection algorithms to identify the relationships currently contributing to the group structure of the network.

Additional problems emerge at the methodological level when we try to track the evolving nature of the mesoscopic layer of the network. The core problem, at this point, involves chaining together solutions over time. Identifying groups within networks involves the use of *community detection algorithms*, and all community detection algorithms evaluate their solution against an *objective scoring* or *evaluation function* for only the representation of the network at hand and not against prior or future representations, similar to a memoryless system. So if a community detection algorithm is used to analyze a series of network snapshots, each solution will be based solely on the snapshot it was assigned to evaluate, and not on prior solutions—there's no continuity between solutions (Gauvin, Panisson, & Cattuto, 2014; Kawadia & Sreenivasan, 2012).

Addressing the continuity between partitions in a network is an unsolved problem with several researchers working on it. One example comes by way of Kawadia & Sreenivasan (2012), who proposed an additional optimization criterion called the estrangement confinement method to evaluate the proposed group structure based on their relationship to prior solutions as well as their ability to partition the current network. Rosvall et al. (2014) propose tracking the historical interactions as *nth* order Markov dynamics, using the map equation scoring function outlined in §4.2.4. Another group of researchers proposes assessing the quality of the partitioning process using null models (Bassett et al., 2013). None of the proposed methods of assessing the results of the community detection algorithms over adjacent partitions are based on how the group structure is expected to influence the interactions of the community members. Yet, one

core component of complex adaptive systems is that they are reflexive, in that agents respond to the emergent order their interactions produce (Ladyman et al., 2012b).

A decision was made to refine and test the method used to track the evolution of the mesoscopic layer of the network because there was no clear guidance in the literature on how to do this, taking into consideration the need to account for agents' reflexivity. The approach used to refine and test the method is outlined in the next section.

4.3 Temporal dynamics of scientific collaboration

The methodology used in this dissertation calls for tracking the evolving nature of the mesoscopic layer of a scientific collaboration network. The literature offers two ways to construct network snapshots over time—the cumulative (Holme & Saramäki, 2012) and effective network approaches (Tomassini & Luthi, 2007), and neither has a distinct advantage. Cumulative networks provide no way of discerning between dormant and active relationships, and effective networks proved to be unstable in the pilot test of this dissertation because established actors would suddenly appear in and disappear from the network between successive time slices. Furthermore, there is no clear guidance on whether the relationships in the network should be modeled dyadically, or as a set of affiliations (i.e., a bipartite graph). In reality, most collaborative interactions involve multiple actors working around a project (their affiliation) (Guillaume & Latapy, 2004; Guimerà et al., 2007b; Newman, Watts, & Strogatz, 2002; Ramasco, Dorogovtsev, & Pastor-Satorras, 2004), so running community detection algorithms on bipartite representations of graphs could lead to better results. The general lack of clear guidance prompted a revisiting of the approach to tracking the evolving mesoscopic structure of a network.

That reexamination started with asking the general question: Is there some view of durability of relationships that could be determined as a function of time? If so, could that durability be used to construct network representations that account for scientists' reactions to the group structure of the network? In contrast to the network representations mentioned above (cumulative and effective networks) where edge weights are a sum of interactions over a specified time interval, time-based networks would see edge weights decay as a function of time. Functionally, this would allow us to more accurately predict the influence of a relationship based on its age and intensity, versus intensity alone.

A discussion of the theoretical guidance regarding the factors that influence the durability of collaborative relationships is provided in the next section, along with several hypotheses resulting from that discussion. Following those hypotheses, the methodological approach employed to test the hypotheses and compare different methods of constructing evolving networks is described.

4.3.1 Half-life of scientific collaboration

In the theoretical development chapter, it was argued that Complex Adaptive Social Systems (CASS) are a special type of complex adaptive system because the agents are capable of abstracting and communicating about their relationships and the structural patterns they form, and adjusting their behavior to their behaviors in response to those patterns. Furthermore, an argument was made that the abstraction and communication imply that CASS have *memory*, which allows historical interactions to guide future interactions. However, it was also argued that memory should fade over time, which allows agents within the system to give preference to more recent interactions over older interactions. It will be argued in this section that challenges

associated with coordinating scientific collaboration underpin memory because those challenges incentivize relationship maintenance while the desire to innovate pushes scientists to seek out new relationships. However, the affordances of repeat collaborations underpinning memory are expected to deteriorate over time, particularly if those relationships are not actively maintained. Thus, it can be argued that relationships have a *half-life* regarding the power of the memory of the relationship to influence future interactions, that is, relationships decay.

The process of identifying potential collaborators and assembling teams is a challenging one that tends to favor repeat relationships. There are several factors working against the formation of new collaborative relationships. First, there are significant trust issues involved in selecting new collaborative partners (Gonzalez-Brambila et al., 2008; Hara et al., 2003). Scientists often use competence and interpersonal criteria to filter out potential collaborators. Second, not all scientists are amenable to working with new collaborators, or are only willing to work with new collaborators if they are introduced through an existing collaborator (Jansen et al., 2010). As a result, scientists who primarily select collaborators based on special competencies are limited because their attempts to form a collaborative relationship are rebuffed. The phrase ‘unilateral initiative with reciprocal confirmation’ (Bunt & Groenewegen, 2007) is used to describe this interaction. The idea is that, in order for a collaborative relationship to form, one scientist must initially propose the relationship (unilateral initiative), but it must be confirmed by the target of the offer.

In addition to trust and interpersonal issues, there are cognitive and administrative hurdles to forming new collaborative relationships. First, collaborating on a research project involves developing a shared understanding of concepts. This is often referred to as a homogenization of knowledge (Guimerà, Uzzi, Spiro, & Amaral, 2005). Second, there are

logistical hurdles to overcome. Although it sounds trivial, many research groups develop different work procedures and tools to manage workflow; integrating discrepant practices can be quite challenging, particularly when those practices and tools are integrated into many other, equally important projects. Many of these issues are resolved through persistent effort; solutions to problems emerge through continued interaction, with the end result being increased productivity (D. H. Lee et al., 2012).

Because there are many obstacles to overcome when establishing collaborative relationships, and reactivating existing relationships often results in increased productivity, the tendency to favor existing over new relationships is particularly strong in collaboration networks. In essence, successful collaborative relationships have a form of momentum (Dahlander & McFarland, 2013), or propensity to continue on because of the benefits of working in established relationships. However, momentum should decay if not maintained, as the collaborators begin to focus on and develop alternative collaborative relationships and the perceived affordances of working with a known partner fade. Additionally, the desire to seek out new collaborative partners as a way to increase the likelihood of generating an innovative product (Uzzi et al., 2013; Whitfield, 2008) works against the momentum of relationships. This leads to the first hypothesis:

H1) Collaborative relationships are subject to decay, such that:

H1a) The probability of finding a collaborative relationship within the system that survives for a specific length of time (t) will be inversely proportional to t.

H1b) The probability of finding a collaborative relationship within the system that is reactivated after being dormant for a specific length of time (t) will be inversely proportional to t.

The decay of collaborative relationships within the complex system of collaboration can be thought of as a form of *half-life of scientific collaboration*, borrowing directly from the concept of *citation half-life* (Burton & Kebler, 1960). At the time, Burton and Kebler were interested in the general rate of obsolescence of scientific literature as a way to manage library collections. However, other researchers found that the concept was useful for estimating the rate of change in scientific fields (C. Chen, 2006; de Solla Price, 1965). Intuitively, this makes sense—if most papers stop getting cited within 2 years in one field, and 10 in a second field, it can be argued that the first field values more recent papers as the forefront of knowledge is changing rapidly while the second field continues to find older literature relevant for longer. This dissertation leverages the same intuition to ask whether the temporal stability of relationships can be used to estimate the general rate of change within the system and the strength of historical ties between scientists.

In particular, we can expect that the dynamic tension between scientists' need to seek out new relationships as a way of producing novel research and the desire to work with existing relationships where most of the coordination challenges are resolved (Stephan, 2012; Whitfield, 2008). In this dissertation, the utility of decay networks is tested as a way to counterbalance the weaknesses of cumulative networks which give equal preference to all relationships regardless of age, and effective networks which give preference to more recent relationships while ignoring the status acquired through a long history of interactions. Instead, decay networks mimic a gradual decline in tie strength over time, as newer relationships supplant older relationships, while still preserving traces of relationships that give rise to cumulative advantage observed in most networks.

4.3.2 Capturing the evolution of the mesoscopic structure of the network

The general approach to tracking the evolution of the mesoscopic structure of a collaboration network, where collaboration is operationalized as coauthorship, is to identify cut points in the temporal range of the data, use data up to and through the cut point to construct a network, and then run the community detection algorithm to extract the modular structure of the network. The main points to consider are the number of time slices, which relationships are included in the time slice, and how to evaluate the results.

There is no exact method for identifying cut points, so this dissertation used three-year intervals based on the argument that three years is equivalent to half the tenure clock or the doctoral education period, as well as roughly equivalent to the duration of a research grant. Tracking the changing structure of the network every year would seem to produce too much noise while tracking it over periods longer than 3 years might result in missing important changes. In terms of which data to include in each time slice, data up to and through the year of the time slice were included. For example, if the data covered publications from the years 1994–2000, and the cut point was 1997, the data from 1994 through 1997 were used. Finally, in order to evaluate the results, scientists' actions from the year after the cut point through the cut point plus three years were analyzed. For the remainder of this dissertation, the former time period is referred to as the *time slice* and second time period the *focal window* (Figure 4-1). Details of the analysis are described below.

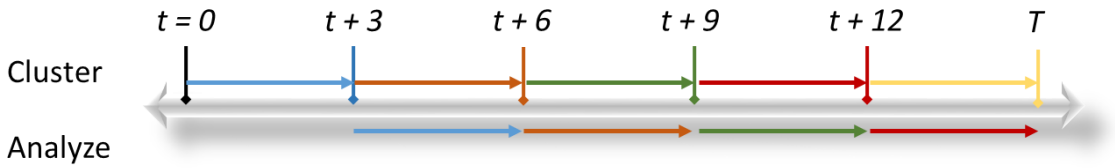


Figure 4-1: The clustering algorithm was run on publication data from the beginning of the time slice in question through the end of the time slice. Collaborative interactions were then analyzed immediately after the time slice for an entire 3-year interval (e.g., *t+4* through *t+6*).

As outlined in the previous section, constructing the network from coauthorship can be done using the cumulative or effective networks approach. This involves taking the coauthorship data and constructing a graph, where an edge between two scientists exists if they coauthored a paper together, and the weight of that edge is equal to the number of papers they coauthored together. For the cumulative network, that weight is equal to the sum of all prior collaborations. For the effective network, the weight is the sum of all collaborations within five years of the cut point, where the window was taken from (Tomassini & Luthi, 2007). In addition to those two networks, this dissertation explored decay-based networks, where the decay of relationships is based on a decay function identified in the analysis of the half-life of scientific collaboration (§4.3.1). Specifically, all relationships were used in a manner similar to the cumulative network, but their weights were discounted based on their age, as determined by the decay function (Equation 3).

$$W = \sum_{t=0}^{Years} \sum_{n=1}^{Pubs} \alpha^{t_0 - t}$$

Equation 3: Calculating the weight of relationships between scientists based on decay, where α is derived from the analysis of the half-life of collaborative relationships.

To give a concrete example, the decay parameter (α) will be set to 0.5. From there, calculating the weight of an edge involves multiplying the sum the number of publications published in any given year by 0.5 raised to the power of the age of the publications, per Equation 3. This was done for all publications between two scientists, with the results from each year summed together for the total weight. For example, if the network was being analyzed in 2000, and two scientists had published two articles together in 1998, 4 in 1999, and 2 in 2000, the weight of the edge between them would be (2 * 0.5 ^ 2) + (4 * 0.5 ^ 1) + (2 * 0.5 ^ 0) = (0.5 + 2.0 + 2.0) = 4.5. In comparison, the weight of the relationship in the cumulative and effective networks would be 8.

Three types of networks have been identified so far—cumulative, effective, and decay. In addition to these three network types, the network itself can be modeled as a bipartite/affiliation graph or as a unimodal graph, where all relationships are dyadic. The dyadic approach is by far the more popular, although intuitively and theoretically, we understand team-based scientific collaboration networks to be bipartite. The reason why is because collaborative relationships are organized around papers or projects, not as a collection of pairwise relationships of scientists who happen to work on a project together. With three ways to consider the weighting of relationships and two ways to consider the method of association, we have six types of networks, arranged in a 2 x 3 matrix (Table 5-1). Those six types include a cumulative unimodal (UC)

network where scientists are connected pairwise, and all relationships are given equal weight, regardless of their age; a cumulative bipartite (BC) network where all affiliations around publications are given the same weight, regardless of how much time has elapsed since the article was published; a unimodal decay (UD) network where the strength of an edge decays over time based on a decay function; a bipartite decay (BD) network where the weight of the affiliation decays over time based on a decay function. Finally, the effective networks (UE and BE) are similar to the cumulative networks, except the time window is limited to the preceding five years.

	CUMULATIVE	DECAY	EFFECTIVE
UNIMODAL	**UC**	**UD**	**UE**
BIPARTITE	**BC**	**BD**	**BE**

Table 4-3: Modeling the temporal and organizational aspects of relationships in the evolving mesoscopic structure of the network. Letters in boxes refer to the abbreviations used.

For each network type, the relationships up until the cut point of the time slice were input into the Infomap algorithm (Rosvall, 2014) for community structure, based on how the particular network should be constructed. The same seed was set for each run, so the results would be reproducible. The two-level hierarchy was not used; instead, the final module leaf on each cluster was treated as a distinct module.

The solutions generated by the different approaches were qualitatively evaluated based on three pieces of evidence—the departure of the solution from the null model of the network, the true positive/ false positive rate, and the size of the clusters generated. The first metric was used to evaluate the solutions based on how far the group structure identified through the detection algorithm deviated from the group structure on a randomized model of the network (Bassett et al., 2013). This was accomplished by extracting all of the collaborations of scientists

who were active in both the time slice and the focal window. Next, the ratio of those collaborations that contained at least one set of in-group relationships was calculated. Then the null model was created by shuffling the authors randomly amongst the publications with no replacement. This was done 1000 times for each network configuration, and the mean of the proportion of in-group collaborations was taken for the random results. Finally, the number of standard deviations for the distance between the observed results in the real-world network and the random network was calculated, which provided an estimate of the likelihood of finding a similar solution by a random process.

The second metric was the true positive/false positive rate, which was calculated by measuring the ratio of collaborations of scientists that included at least one scientist within the group as the true positive rate (TPR), and the ratio of within-group relationships who have not collaborated as the false positive rate (FPR). Neither metric is particularly useful on its own. However, the combination of both metrics and a descriptive analysis of group sizes proved to be very useful. More specifically, the TPR and FPR, the deviation from the null model, and the descriptive analysis helped tease out which approach to constructing the network performed better simply because the groups identified with the detection algorithm were larger and therefore, by sheer chance, captured a higher portion of the collaborations, versus those that provided more discriminatory power.

The beginning of the chapter focused on outlining the general framework of this exploratory study, including the operationalization of concepts and the experimental process used to test different approaches to tracking the evolution of the mesoscopic layer of the network. The following section describes the data source used in this dissertation.

4.4 Data source

Bioinformatics is an excellent example of a hierarchically organized discipline, where scientists are organized into institutes, centers, and labs or (e.g., http://www.broadinstitute.org/scientific-community/science/core-faculty-labs) groups (e.g., http://www.broadinstitute.org/chembio/lab_schreiber/home.php). Each organizational level is led by scientists at differing levels of seniority. Although there are established organizational divisions, researchers will frequently participate in team research across formal organizational boundaries. There are multiple areas of research within the genomics community, one of which focuses on the sequencing of DNA. DNA sequencing usually involves the integration of scientists with a variety of skill sets, including wet lab biologists, clinical researchers, statisticians, bioinformatics experts, chemists, theoretical biologists, and mathematicians. Additionally, the sequencing of DNA involves the use of expensive equipment that is owned by or assigned to the lab.

In many ways, the sequencing of DNA is an area where we would expect to see some dependence between researchers. Basic research in this area requires the integration of multiple skill sets and access to expensive equipment, each of which is brokered by established scientists. This environment also creates a scenario where lab leaders may become dependent on scientists with specific expertise because researchers with specific expertise may be in short supply, or because it is administratively simpler to rely on a single person, versus engaging in a reoccurring search process for new collaborators. Because genomics research matches the organizational structure, collaboration structure, and resource allocation expectations of a research field organized around the research group, this dissertation chose to focus on the international research community surrounding the sequencing of DNA.

The primary source of data for this dissertation was metadata related to the intellectual provenance of genome data submissions to GenBank (http://www.ncbi.nlm.nih.gov/genbank/), an international nucleotide sequence data repository serving the bioinformatics community. GenBank has been in continuous operation since 1984 and now houses over 130 million data submissions of DNA/RNA sequences. GenBank operates as part of an international consortium of genetic data repositories, with each database exchanging data on a daily basis. More detailed overviews of the data source were published in (Costa & Qin, 2012; Costa, Qin, & Bratt, In press; Costa, Qin, & Wang, 2014), with a summary included here. Of particular relevance is the nature of the submission process, where scientists will submit the sequence data and frequently attach a formal publication to the data set. Therefore, the publication data is directly related to the sequencing of a certain portion of the DNA or RNA. This dissertation focused on the authorship of those publications.

Metadata, along with the genetic sequence data, is accessible via GenBank's web interface. The data sets are also stored in compressed, semi-structured text files on an FTP server hosted by the National Center for Biotechnology (NCBI). Data for this study was collected from the text files as part of a larger study on the collaboration activities of scientists contributing to data repositories (Qin, 2014). The data set for the larger study was collected from the FTP site in August of 2013 and computationally processed. The compressed files were downloaded, decompressed, and processed, with the metadata extracted and the genetic sequence data dropped. Once the metadata was extracted, the data were re-parsed into a database and normalized for analysis. Author names were disambiguated by first normalizing the data, then stemming the names, which were then compared using similarity and proximity measures based

on shared coauthors and organism focus. Descriptive statistics of the data set are provided at the beginning of the results chapter.

4.5 Analytical approach

This network analytic framed exploratory study relied heavily on computational manipulation of trace data to support the analysis. The general approach used was to start with basic descriptive analysis of the population, using the results of the general descriptive analysis to motivate more detailed analysis. For example, the research on the dependence of scientists started with general descriptive statistics of the distribution of dependence across the population as well as some cross-tabulation analysis on the relationship between dependence and productivity over time.

From there, the entire population for the different time periods was divided into subgroups by roles, as described in §4.2.5. General descriptive statistics of the distribution of dependence by roles were provided, followed by more fine-grained analyses of role-to-role dependencies, and net dependence by role. Much of the analysis was guided by questions that arose from the results of the preceding analysis, usually in the form of "Why do we observe *x*?" or "We see *x* and one possible explanation is . . . so we should see *y*?" As an example, the distribution of changes in maximal dependence between two cohorts was W shaped, with no easily discernable pattern between the two cohorts. That observation prompted the question: How else can the data be studied to provide more insight into the concept of dependence given the limitations of analyzing maximal dependence alone? In this case, the answer was to look at median dependence as well (more detailed explanations are provided in Chapter 5).

An example of analysis that was motivated by the second question is where the clustering coefficient was used to study the relationships surrounding scientists in certain roles. That analysis was motivated by the interpretation of the relationship between maximum and minimum dependence, specifically, that certain roles seemed to be working with several independent groups infrequently with little overlap between groups. That suggested that the clustering coefficient should be lower around scientists with that specific pattern of dependence.

From a personal standpoint, I argue that the single most important thing to do when you're conducting exploratory analysis of large data sets is to always think of ways to test your interpretations of your analysis. There needs to be a constant process of referring back to the theory, thinking about how the observations confirm, refute, or expose gaps in the theory and the understanding of concepts in the theory. Any time observations are made, the researcher should always think about how the data can be manipulated to test the interpretations. This process is doubly important when the analytic process requires heavy coding because it is easy to make mistakes that result in inaccurate measurements. Thinking about how to test results, either through secondary analysis, or random sampling and hand-calculated verification of results, is an important part of the process. Basically, I argue that the last thing the field of computational social sciences needs is lazy data dredging (Smith & Ebrahim, 2002).

The code for the dissertation is available upon request, and will eventually be made public once it can be cleaned and scrubbed of security related information (e.g., authenticated calls to the database).

4.6 Limitations

The limitations of this study can be divided into two broad categories—reliability and validity. In terms of reliability, the data and analytic techniques are available and technically replicable. The only two points in this dissertation where quasi-randomized processes were leveraged were in the use of the Infomap clustering algorithm and the randomized networks for evaluating the results of the clustering algorithm. The Infomap algorithm does allow users to set the random seed, which facilitates replication; the latter process has no similar mechanism. However, it is expected that any researcher who uses the same approach will get similar results.

In contrast to reliability, the challenges to validity are more numerous:

Operationalization of collaboration as coauthorship—Other researchers have addressed the limitations of operationalizing collaboration as coauthorship (Glänzel & Schubert, 2005; Laudel, 2002; Melin & Persson, 1996). The main concerns are that (1) scientists collaborate on more than just papers; (2) coauthorship can overstate the level of interaction between the participants; (3) authorship can overstate the contributions of authors; and (4) collaboration falls on a spectrum and not all collaboration warrants coauthorship. De Haan (1997) identified six types of collaboration, only one of which was coauthorship of a paper. Laudel (2002) argues that about half of scientific collaboration is invisible because it does not culminate in formal acknowledgement or coauthorship. Laudel also argues that coauthorship can overstate collaboration because scientists are often given honorary coauthorships, either to leverage the name recognition of a prominent scientist or as a courtesy to a friend. Even the collaborations that receive formal acknowledgement (Cronin et al., 2003), but are not included in coauthorship, get excluded from analysis based on coauthorship.

As a rebuttal, publication is central to the activity of science; scientists are expected to be and do get evaluated on their contributions to the general body of knowledge. It's important to recognize that there are other forms of collaboration, and that it would be useful to consider them, but collaboration on a publication is the most important of collaborations (assuming that scientists are operating under the norms of science). After all, no one gets tenure by listing their *informal* contributions to papers on their curriculum vitae.

Arguably, the most significant drawback related to operationalizing collaboration as coauthorship of a scientific publication is the fact that it does not capture collaboration on other formal knowledge outputs, including data sets (Costa et al., 2015) and patents. There are many opportunities for commercialization in genomics; the GenBank repository contains metadata on approximately 25 million patents (Costa & Qin, 2012), which represents a significant amount of collaborative effort. Including metadata from submissions and patents should be part of a follow-up study.

GenBank as a publication repository—GenBank is not a publication repository. It is a data repository that contains metadata on publications related to the datasets stored in the repository. As such, it does not contain second-order publications. Scientists cannot submit publications that use or synthesize analysis on datasets already in GenBank; instead, each publication must be attached to a sequence submission. It's possible to identify scientists who have had publications in the GenBank repository who appear to have stopped publishing five or more years ago, yet are still actively researching and publishing on genomics.

Despite the limitations, GenBank was still considered to be a useful data source for several reasons. First, this dissertation is part of a larger project studying the collaborative

interactions of scientists around a large cyberinfrastructure investment. Also, there is an advantage to limiting the source of data to publications in GenBank in that it focuses on the sequencing genomic data, which is technically complex and requires the integration of multiple skill sets, which makes the community a relevant example of "team science." (Costa et al., 2016).

One possible way to address the potential limitations of using GenBank publication as the focal point for analysis is to test whether the patterns of coauthorship and social organization around GenBank are similar to or different than patterns in the broader research community. This can be done by expanding the analysis to include article metadata from a comprehensive publication repository (e.g., PubMed or Web of Science) in a follow-up study.

Modules and formal groups/labs—Community detection algorithms extract the modular structure of the network, either based on structural divisions that optimize the links within modules against links between modules, or simulated information flows and information compression performance. In either case, the modules differ from formal organizational groups, and under certain conditions, will erroneously place scientists in modules with which they have had little interaction (Danon et al., 2005; Lancichinetti & Fortunato, 2009; Velden et al., 2010). Furthermore, using community detection on temporal networks is still unproven—the results between time periods are not linked together, which may, depending on how the network is constructed, result is significant shuffling of module assignments that bear no resemblance to one another.

More importantly, in the process of trying to make the dissertation readable and accessible to readers who are not immersed in network analysis, there is a tendency to want to

switch back and forth between the terms modules and groups, or to use the latter term extensively. People understand the idea of a research group and using the term facilitates connecting the research to their current understanding of the phenomenon being studied. The problem is, use of the word *group* may unintentionally seed the idea that modules *are* formal affiliations, which they are not. Modules do not always coincide with formal affiliations—in some sense, the algorithms pick up on relationships that exist out of comfort, results, or necessity, and may easily cross formal affiliation boundaries. Having said that, there is some sense that modules should be related to formal affiliations, as we expect the formal lab or research group to be heavily influential on collaboration patterns. The metadata related to formal affiliations of authors was not available for this dissertation, so the module assignments could not be validated against those affiliations.

Publication dates as dates of collaboration—Throughout the dissertation it was assumed that the temporal ordering of publications matched the temporal ordering of collaboration. This assumption is safe if there is little variability in the lag between collaboration and publication. If there *are* significant differences either between publication outlets, or time to publication in different years, then the validity of the temporal analysis is reduced.

4.7 Summary

The core question guiding this exploratory study was:

What is the relationship between a scientist's position within the group structure and their dependence on other scientists, and how has the relationship between position and dependence changed over time?

The focus of this research was on scientific collaboration, which was operationalized as coauthorship of an article (Glänzel & Schubert, 2005). Data on coauthorship was used to create a collaboration network, with scientists as nodes that were connected by joint authorship of a paper. A scientist's dependence on another scientist was operationalized as the fraction of the scientist's papers coauthored with the second scientist compared to all the scientist's papers. Interdependence describes a two-way relationship, as any two authors who coauthor one or more papers together exhibited some degree of dependence on one another. The Infomap community detection algorithm (Rosvall, 2014) was used on the collaboration network to extract the group structure and assign scientists to their respective groups. After scientists were placed into groups, their role within the group structure was calculated using the node role framework developed in (Guimerà & Amaral, 2005; Guimerà et al., 2007b). Table 4-4 outlines the concepts and their operationalization.

Concept	Operationalization	Examples
Scientific collaboration	**Coauthorship**	**(Glänzel & Schubert, 2005)**
Dependence	**Fraction of papers coauthored with other**	
Collaboration network	**Coauthorship network**	**(Newman, 2001b)**
Groups	**Modules**	**(Rosvall & Bergstrom, 2007)**
Position within the group structure	**Role classification**	**(Guimerà & Amaral, 2005)**
Adaptive system	**Temporal networks**	**(Holme & Saramäki, 2012)**

Table 4-4: Core concepts and their operationalization

Studying temporal networks is challenging with no standardized approaches other than using snapshots or windows to analyze the network at various points in its history (Aggarwal & Subbian, 2014; Holme & Saramäki, 2012). The actual construction of the windows can be done in several different ways, with no clear guidance on which approach is preferred, particularly

when the researcher wants to study the evolution of the mesoscopic layer of the network. The lack of clear guidance motivated an experiment to constructing the network using different approaches, testing two known approaches and an approach developed in this dissertation. The approach developed in this dissertation was based on a framework focused on the factors influencing scientists' desire to maintain or reactivate a prior collaborative relationship. The primary hypothesis guiding the experiment was that collaborative relationships are subject to decay, the decay can be modeled, and the decay model can be used in the temporal analysis of the network.

The final part of the experiment involved analyzing six network construction techniques for temporal network analysis. The experimental design outlined several tests that were used to compare the outputs of the community detection algorithm on the networks created with the different techniques.

The chapter concluded with a description of a research community organized around the research group—Genomics. A general description of the organizational structure, skill sets, and resources involved in conducting DNA sequencing research was provided to demonstrate the field's similarity to what was expected in the theoretical framework. The data set and extraction process were also included. Descriptive analysis of the data set is in the next chapter, along with the results of the exploratory analysis.

5 Results

5.1 Half-Life of scientific collaboration

To briefly recap the motivation behind the analysis discussed in this section, the literature review and theoretical development sections of the dissertation have addressed some of the current challenges associated with temporal networks. In particular, there is no established way to identify which of the historical relationships are influencing the actors in the network. Two approaches to network construction are commonly used: cumulative networks and effective networks. The general question was raised: Is it possible to identify a temporal property related to collaborative relationships that can be used to construct network representations that account for scientists' reactions to the group structure of the network?

The first step toward answering that question was to explore the effect time has on collaborative relationships. The background argument led to the following hypothesis:

H2) Collaborative relationships are subject to decay, such that:

H2a) The probability of finding a collaborative relationship within the system that survives for a specific length of time (*t*) will be inversely proportional to *t*.

H2b) The probability of finding a collaborative relationship within the system that is reactivated after being dormant for a specific length of time (*t*) will be inversely proportional to *t*.

For H1a, we can identify the probability that a relationship will last for *t* years by first calculating the duration of all relationships within the community, then determining the

proportion of all those relationships that have lasted for *t* years. Figure 5-1 shows the results if we track the ratio of all coauthorship relationships that produce publications for *t* years under two conditions. The first condition looks at all unique coauthorship relationships (no repeats), including those where the two scientists never coauthored again. The second condition includes only those relationships where the collaborators coauthored a paper in two consecutive years. For this analysis, data through 2009 (instead of 2012) were used because no relationship after 2009 could continue for more than 3 years within the timeframe for this study, perhaps skewing the results to the lower end of the distribution.

There are three major limitations to note. First, the analysis relies on the assumption that the publication ordering reflects the temporal ordering of the collaborative relationship. Second, if the two authors had more than a one year break between their publications, that coauthorship relationship was not considered to be continuous. The third limitation is that every time two scientists had a break in coauthorship of one year or more, then reactivated that coauthorship relationship, it was considered a new collaboration. So the second limitation resulted in undercounting the number of continuous relationships, while the third limitation resulted in overcounting relationships.

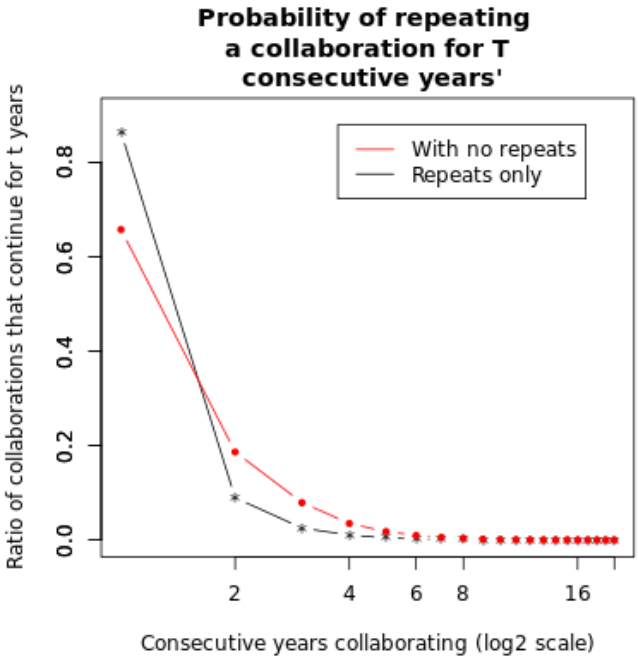


Figure 5-1: The probability any two scientists will continue a collaboration for x years. The red line includes figures for those who never repeat a collaboration; the black line includes figures for those who collaborate in two or more separate years.

Of the scientists who coauthored, 86.4% never coauthored again, 8.9% coauthored for one additional year, and less than 0.6% coauthored for six or more continuous years. Focusing on all the scientists who coauthored again at least one time in two different years, 65.7% coauthored for only one additional year, another 18.6% coauthored for two more consecutive years, and no more than 0.8% coauthored for seven or more continuous years. Both plots, for nonrepeating relationships, and those relationships that did reoccur, monotonically decrease and can be modeled using an exponential decay function $f(t) = ae^{bt}$. However, the two plots differ in their intercepts and slopes, where b = -1.71 and a = 2.11 for the model including repeats only, and a = 0.864 and b=-2.23 ($r^2 > 0.98$) for all relationships, including those that did not collaborate again. The actual probability of finding a coauthorship relationship that published for *t* continuous years is roughly half that of one that published together for *t-1* years, where the minimum *t* is 3 years.

For H2a, we look at the probability of finding a relationship that gets reactivated after a *t* year hiatus. Once again, the red line in Figure 5-2 includes relationships that were never repeated, while the black line maps the data of relationships that produced a coauthored paper in two different years or more.

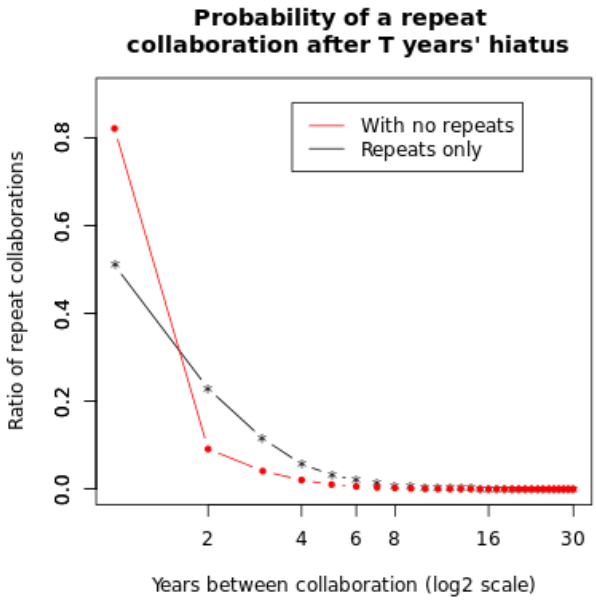


Figure 5-2: Probability of two scientists reactivating a collaborative relationship after not collaborating for t years. Red includes all collaborations, including those who never repeated (0), while the black line only includes relationships that were repeated at least once.

Of all the coauthorship relationships in the dataset, 82.1% were never reactivated, which is nearly identical to the observed values in (X. F. Liu, Xu, Small, & Tse, 2011). Another 9.11% were reactivated within one year, and cumulatively, less than 0.7% of relationships were reactivated after six years. Of the relationships that were reactivated, 51.1% were reactivated within one year, and another 22.9% were reactivated within two years. By the time six years passed, 96.3% of the scientists who were expected to coauthor again coauthored, and less than 0.8% of coauthorship relationships were rekindled after a 10-year hiatus. The probability of finding a relationship that is reactivated after a *t* year hiatus decreases with *t* and can be modeled

with the exponential decay function $f(t) = ae^{-0.75t}$ ($r^2 > 0.99$), with differing intercepts for the plot with no repeats ($a = 1.07$) and for the plot with repeats only ($a = 0.51$). Once we get past $t = 2$, the actual probability of finding a relationship reactivated after *t* years is roughly half that of finding a relationship that has been reactivated after *t-1* years.

Using coauthorship as an operationalization of collaboration, and using the dates on those publications as an operationalization of the temporal sequencing and duration of collaborative relationships, we find support for both H1a and H1b. The data related to both hypotheses produce monotonically decreasing functions that are well-modeled by an exponential decay function. Modeling stable relationships, or those that published together for *t* continuous years, provides some insight into the effect time has on collaborative relationships.

However, the more important model is the hiatus model, where we try to model the likelihood a relationship will be reactivated after *t* years not having published together. The hiatus model is more important for this dissertation because it directly relates to the concept of a collaboration half-life, and it helps us get to the problem of creating temporal networks—determining the probability that any given relationship out of a set of historical relationships is relevant to the dynamics of the current network. Put another way, we are trying to model the strength of the relationship in terms of how it influences an actor's collaboration relationship over the duration of the relationship.

From both perspectives of looking at relationships, including analyzing all relationships, and only those where they do publish again, there is congruence on the decay function. Furthermore, the rate of decay approaches a half-life of one year after year 2, which means it is possible to use a decay function heuristic of one year for creating decay networks (as described

in §4.3.2 and analyzed in §5.2). It will be demonstrated in the following section that the decay model of creating networks has several distinct advantages over other methods for tracking changes in the network at the mesoscopic level.

5.2 Tracking the evolving mesoscopic structure of scientific collaboration networks

Recall that the theoretical framework predicts that the current configuration of relationships within the research community serves as both a constraint and point of reference for the scientists within the community (Ladyman et al., 2012a; Wagner & Leydesdorff, 2009). Furthermore, based on the framework proposed in this dissertation, we expect that the group structure influences the scientists within the community based on the concept of bidirectional causality (Sawyer, 2005). However, extracting the group structure of a network as it evolves over time is not a mature methodology with well-established approaches.

Six different approaches to constructing temporal networks were used, divided into two dimensions. There are two categories in one dimension, and three in another (Table 5-1). In the first dimension, the two categories are two types of networks—unimodal networks where all relationships are viewed as dyadic, and a bipartite network where all relationships are organized around an affiliation, which in this network is a publication. The former type is more common, the latter more closely aligned with the theoretical framework guiding this dissertation. The categories in the second dimension consist of three ways of determining the effects of time on the strength of relationships. The cumulative approach (Holme & Saramäki, 2012) uses all trace data to construct a network, ignoring the effects associated with time. The second approach is to use effective networks (Tomassini & Luthi, 2007), taking relationship data from only the previous five years under the assumption that the effects of time are negligible within that five

year window and older relationships have little to no effect on the current interactions of scientists. Finally, decay networks are introduced, where the expected strength of a relationship decays over time and can be calculated according to the formula equation described in §4.3.2.

	CUMULATIVE	DECAY	EFFECTIVE
UNIMODAL	**UC**	**UD**	**UE**
BIPARTITE	**BC**	**BD**	**BE**

Table 5-1: Modeling the temporal and organizational aspects of relationships in the evolving mesoscopic structure of the network. Letters in boxes refer to the abbreviations used.

Table 5-2 provides summaries of the results, organized first by year, then by network construction type. The bipartite network constructions consistently produced clustering solutions with larger maximum group sizes. However, that pattern changed once transients were accounted for (Braun et al., 2001; de Solla Price & Gürsey, 1975), with transiency operationalized as scientists who published together for one year only. The maximum size of groups decreased the most for the bipartite groups once transients were accounted for. Put another way, the ratio of maximum group size with transients to maximum group size without transients was consistently higher for the bipartite solutions than the unimodal solutions, and the solutions based on decayed weights (Figure 5-3).

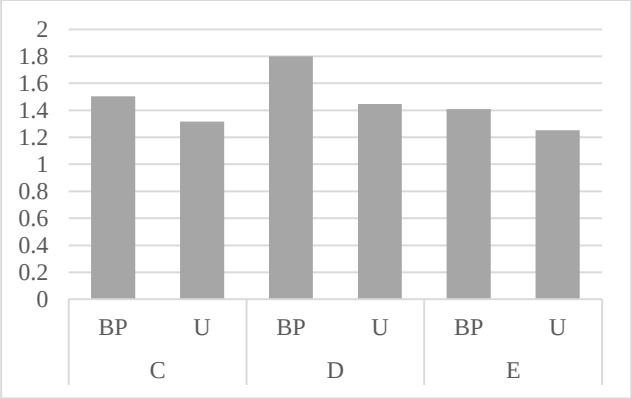


Figure 5-3: Ratio of maximum group sizes with transients versus without transients.

YEAR	TYPE	ACTIVE MODULES	FIRST QTR	MEDIAN	MEAN	SD	THIRD QTR	MAX
1994	BC	7485	4 (3)	8 (6)	11.167 (8.354)	12.31 (9.128)	13 (10)	222 (175)
1994	BD	9229	4 (3)	6 (5)	9.048 (6.961)	9.73 (7.366)	11 (8)	204 (130)
1994	BE	2285	12 (8)	27 (19)	31.084 (22.689)	70.982 (57.041)	41 (30)	3231 (2607)
1994	UC	8095	4 (3)	8 (6)	10.326 (7.793)	8.868 (6.806)	13 (10)	193 (151)
1994	UD	13125	3 (3)	5 (4)	6.334 (5.171)	5.077 (4.032)	8 (6)	119 (84)
1994	UE	8155	4 (3)	7 (5)	8.688 (6.841)	7.122 (5.782)	11 (9)	180 (146)
1997	BC	9470	5 (3)	8 (6)	12.652 (9.114)	40.092 (30.063)	15 (11)	3681 (2640)
1997	BD	11840	4 (3)	6 (5)	10.113 (7.58)	13.424 (9.753)	12 (9)	475 (286)
1997	BE	9524	4 (3)	7 (5)	9.906 (7.584)	16.246 (12.878)	12 (9)	1197 (940)
1997	UC	10524	5 (3)	8 (6)	11.386 (8.28)	10.356 (7.684)	15 (10)	206 (150)
1997	UD	16739	3 (3)	5 (4)	7.126 (5.664)	6.739 (5.279)	9 (7)	174 (130)
1997	UE	10142	4 (3)	7 (5)	9.3 (7.168)	8.287 (6.551)	12 (9)	181 (143)
2000	BC	11031	5 (4)	9 (6)	15.656 (10.806)	113.17 (84.133)	17 (12)	11688 (8182)
2000	BD	15524	4 (3)	7 (5)	11.118 (8.075)	20.899 (14.6)	13 (9)	1709 (1116)
2000	BE	11856	4 (3)	7 (5)	10.747 (7.991)	29.815 (22.916)	12 (9)	2984 (2203)
2000	UC	14011	5 (4)	9 (6)	12.328 (8.531)	12.07 (8.799)	16 (11)	357 (286)
2000	UD	21822	3 (3)	5 (4)	7.876 (6.054)	8.623 (6.598)	10 (7)	418 (307)
2000	UE	12867	4 (3)	7 (5)	9.903 (7.448)	9.699 (7.672)	12 (9)	352 (290)

2003	BC	11556	5 (4)	10 (7)	19.363 (12.878)	167.524 (114.605)	21 (14)	16740 (10958)
2003	BD	18122	4 (3)	7 (5)	12.335 (8.688)	45.431 (27.617)	13 (10)	5512 (2944)
2003	BE	12838	5 (3)	8 (6)	11.974 (8.716)	44.867 (32.737)	13 (10)	4100 (2898)
2003	UC	16079	5 (4)	10 (6)	13.919 (9.241)	15.11 (10.731)	18 (11)	543 (387)
2003	UD	25988	3 (3)	6 (4)	8.557 (6.395)	10.077 (7.423)	10 (8)	410 (275)
2003	UE	14310	5 (3)	8 (6)	10.743 (7.888)	12.08 (9.468)	13 (10)	519 (375)
2006	BC	12915	5 (4)	11 (8)	21.605 (13.887)	202.964 (127.722)	23 (15)	20330 (12376)
2006	BD	20661	4 (3)	7 (5)	13.465 (9.293)	90.562 (51.616)	14 (10)	11529 (5604)
2006	BE	13335	5 (3)	8 (6)	12.708 (9.189)	67.262 (48.021)	14 (10)	6028 (3876)
2006	UC	18529	6 (4)	10 (7)	15.061 (9.687)	18.949 (13.638)	19 (12)	850 (680)
2006	UD	29622	3 (3)	6 (4)	9.332 (6.854)	12.71 (9.102)	11 (8)	588 (383)
2006	UE	14971	5 (3)	8 (6)	11.318 (8.296)	14.041 (11.154)	14 (10)	454 (366)
2009	BC	13359	6 (4)	12 (9)	25.032 (15.508)	256.775 (146.45)	26 (16)	23745 (13502)
2009	BD	22739	4 (3)	7 (5)	14.61 (9.821)	129.344 (71.698)	15 (10)	19007 (9039)
2009	BE	13236	5 (3)	8 (6)	13.785 (9.771)	92.81 (65.614)	15 (11)	8094 (5009)
2009	UC	18472	6 (4)	12 (7)	18.106 (11.202)	26.57 (19.05)	23 (13)	2093 (1561)
2009	UD	32029	3 (3)	6 (5)	10.317 (7.454)	19.604 (14.283)	12 (9)	2491 (1622)
2009	UE	14869	5 (3)	8 (6)	12.27 (8.826)	20.174 (16.976)	15 (10)	1686 (1431)

Table 5-2: A summary of the solutions generated by the clustering algorithm. TYPE is the approach used to create the network (see Table 5-1); ACTIVE MODULES include those with 2 or more scientists (those with 1 include unconnected transients); Numbers in parentheses refer to module populations without transients.

The bipartite solutions do a better job of revealing the fact that certain groups of scientists leveraged transients more frequently than other scientists. The analysis of group sizes was rerun a second time, calculating maximum group size when all scientists that had two years' of experience or less, or two or fewer publications were dropped. The results changed the most dramatically for the bipartite solutions. This change can be seen most clearly by looking at the 2009 data again, where the group sizes from the various solutions were the largest (Table 5-3). The mean group sizes and standard deviations all collapse to a much tighter range once the least active of scientists are excluded from the calculations.

TYPE	FIRST QTR.	MEDIAN	MEAN	STD. DEV	THIRD QTR.	MAX
BC	6 (3)	12 (4)	25.032 (5.881)	256.775 (4.982)	26 (7)	23745 (119)
BD	4 (3)	7 (4)	14.61 (6.086)	129.344 (5.546)	15 (7)	19007 (111)
BE	5 (3)	8 (4)	13.785 (6.207)	92.81 (6.488)	15 (7)	8094 (243)
UC	6 (3)	12 (5)	18.106 (6.542)	26.57 (7.993)	23 (8)	2093 (277)
UD	3 (3)	6 (5)	10.317 (6.907)	19.604 (9.526)	12 (8)	2491 (310)
UE	5 (3)	8 (5)	12.27 (7.232)	20.174 (15.483)	15 (8)	1686 (1270)

Table 5-3: Group population characteristics with all scientists with < 2 years' activity or publications excluded (revised figures in parentheses)

Using the method described above, approaches that produce larger groups also capture a larger percentage of collaborations within a group. However, it is likely that this accuracy is achieved by chance alone because larger groups will inherently capture a larger percentage of the collaborations. The most extreme example would be to place all scientists in one group, which would functionally assign all classifiable collaborations into the same group. But that approach provides no discriminatory power, so how do we evaluate the performance of the algorithm under the theory when the theory suggests that scientists should be more likely to collaborate with group members than not collaborate? It does not suggest that scientists should not

collaborate with scientists outside of their group, which we know happens because intergroup collaboration is relatively common. Nor does the theory suggest that scientists collaborate with everyone in their group, yet if the group is so large that it is unlikely two scientists within the group would ever meet each other, then the concept of group loses its utility.

Looking at Table 5-4, we see data on the ratio of collaborations that contain at least two members of the same group by network construction type. The bipartite network configurations capture a higher proportion of collaborations within the group than the unimodal configurations (column 1 of Table 5-4). However, some of that increased performance may be due to chance because the groups are larger. To test whether the ability of the different approaches to capture within-group coauthorship was due to chance, each approach was compared to null models of the network reconstructed in the exact same manner as the network under analysis. The methodology is fully described in §4.3.2, but briefly, the results observed in the real network were compared to the mean within-module collaboration ratio for 1000 randomized trials (column 3 of Table 5-4). Also, the number of standard deviations that the observed mean was from the mean of the null models was calculated as a way to help differentiate the extent to which the observed values differed from the random values observed in the null models (column 4).

The bipartite cumulative network, which produced the largest groups and the most accurate partitioning also produced the solution that was closest to random out of any solution (although it was still 10 s.d. away from random). Conversely, the unimodal effective network tended to produce the smallest groups, was the farthest away from randomized networks, yet was also the least accurate in terms of classifying scientists into groups that accounted for their collaborations.

NETWORK TYPE	RATIO WITHIN	MEAN RATIO WITHIN, RANDOMIZED	SD DIFF
BC	0.535	0.087	10.522
BE	0.483	0.037	14.747
BD	0.467	0.023	15.968
UD	0.420	0.016	18.419
UC	0.424	0.016	19.197
UE	0.407	0.010	30.304

Table 5-4: Performance of the different clustering configurations.

Under the evaluation scheme above, there are no true negatives to evaluate the clustering algorithm because the theory does not suggest that scientists should not collaborate with people outside of their research group, but instead that they should not collaborate without someone in their group. However, it is possible to penalize solutions for producing larger groups by determining the false positive rate based on the number of relationships in a group that are not present. Larger groups often include pairs of scientists who did not interact; we can use this pattern to penalize solutions that achieve higher true positive rates simply by grouping more scientists together into larger modules. We can compare the result of the different solutions by looking at the ratio of collaborations of scientists that included at least one pair of scientists within the group as the true positive rate (TPR), and the ratio of within-group relationships that have not collaborated as the False Positive Rate (FPR) (Table 5-5).

NETWORK	FPR	TPR
BC	0.994	0.535
BE	0.969	0.483
BD	0.920	0.467
UD	0.837	0.420
UE	0.868	0.424
UC	0.893	0.407

Table 5-5: True and false positive rates of the different solutions

The results, in terms of identifying the best approach for constructing networks, are not as conclusive as hoped for, particularly along the time–weight dimension. Effective networks consistently had false and true positive rates in the middle of the pack, while cumulative approaches were mixed based on whether the network type was bipartite versus unimodal. Unimodal cumulative networks, by far the most popular approach in the literature, performed the worst in terms of capturing forward collaborations. Decay networks had the lowest false positive rates in the bipartite and unimodal groups, but also had the lowest true positive rate in the bipartite group. The community detection algorithm generated larger groups on the unimodal networks after the least active scientists were screened out. This suggests that it is harder to identify core groups of scientists that interact with one another using unimodal networks. The modular solutions on unimodal networks do appear to be the furthest from the null model, although no solution on any network type was anywhere near the observed values in their respective null models.

Bipartite approaches captured a larger percentage of collaborations within group than the unimodal networks. The module sizes in bipartite cumulative (BC) networks were consistently larger and appear to have captured a larger percentage of forward collaborations within group simply by chance because the groups were larger. The group sizes for the bipartite decay networks were smaller than groups in any other approach once transients were accounted for, making it the preferred solution for identifying the relevant group structure in terms of our understanding of research groups.

Although the results are not conclusive, this experiment was still worth conducting. The results demonstrate that breaking from the common approach (unimodal cumulative) to constructing networks is not detrimental to studying scientific collaboration, and in fact may be

better because the networks consistently capture a greater percentage of collaborations within group, with a marginally worse number of false-positives. If the goal of the research project is to track the evolution of a network over time and how that network structure influences the collaborative interactions of scientists, then bipartite analysis of the network is a better choice if the community under analysis can be characterized as being organized around the research group. The language we use to describe a phenomenon shapes our understanding of the phenomenon. For those who model scientific collaboration with network analytic approaches, using terminology and measurements that treat scientists as independent actors creates a divide between their views on collaboration and sociologically grounded views on the organization of scientific fields (Moed, 2006; Ziman, 1994).

5.3 Dependence

Sections 5.1 and 5.2 outlined research related to constructing networks for the purpose of tracking the evolving mesoscopic structure of those networks and how those structures influence the collaborative interactions of scientists. In particular, this dissertation focused on scientists' dependence on their groups and on other scientists, and how that dependence changes over time. Here, dependence is operationalized as the proportion of papers coauthored with at least one other person in the group (dependence on the group), or the proportion of papers coauthored with each of their coauthors (dependence on others).

5.3.1 Dependence and productivity

If a scientist has *x* coauthors for a specific period of time, then they have *x* dependence scores, one for each coauthor. The maximum dependence score gives us some indication of how dependent the scientist's publishing activity is on at least one other scientist, although it does not

tell us why that dependency exists. Although it is not possible to tell why a scientist is dependent on another scientist, it is possible to explore whether it takes scientists who have entered the network recently a longer time to become less dependent on the person they rely on than scientists who entered the network in previous years. Put another way, if two cohorts of scientists who entered the network at different times are compared, were scientists from the earlier cohort more likely to reduce their maximum dependence sooner than scientists from the second cohort? To answer that question, two cohorts of scientists were analyzed, one from 1994 and the other from 2003. Each cohort included only scientists who remained in the network for the entire period of analysis, which was 9 years following the year the cohort entered the network. There were 3857 scientists in the 1994 cohort, 3850 in the 2003 cohort. For both cohorts, the mean, median, and mode for starting maximum dependence was 1.00. For those who started in 1994, 35.6% exhibited no change in the intervening years, while 33.4% of the second group exhibit no change. Mean change in maximum dependence over the 9-year period was -0.289 for the group starting in 1994, -0.281 for the group starting in 2000.

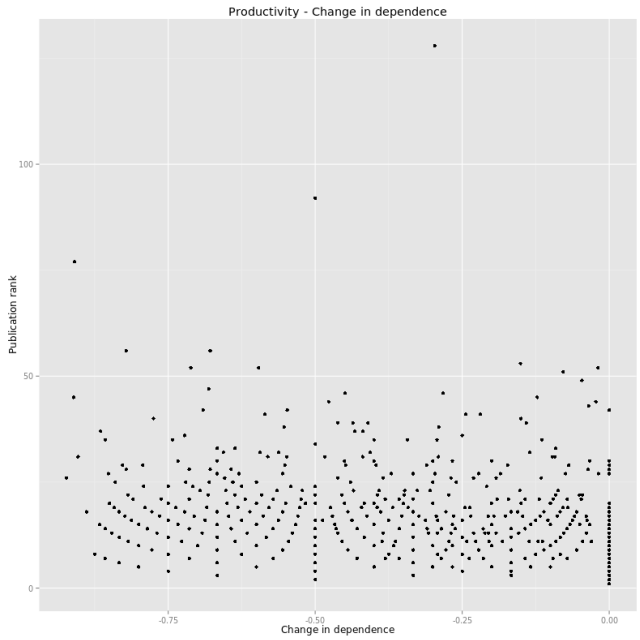


Figure 5-4: Relationship between maximum dependence and productivity for the two groups who were active for 9 or more years.

There was no statistically significant difference between the two cohorts, so maximum dependence is not a useful indicator in and of itself. As a follow-up question, we can explore the relationship between change in dependence and productivity, asking whether more productive scientists become less dependent, or if the relationship between productivity and dependence takes some other form (or none at all). A negative, statistically significant correlation, $r(7705) = -0.258$, 0.95 CI[-.027, -0.237], between productivity and change in dependence was identified when outliers who had an increase in dependence were omitted (~1.4% of the population). A plot of the relationship between productivity and change in dependence suggests that the relationship is more complex; scientists with the highest productivity (10+ publications, 17% of the population) are distributed over the entire range of the spectrum in terms of change in maximum dependence.

Even though there is a statistically significant negative correlation between maximum dependence and productivity, a graph of the relationship between the two variables shows that the relationship is much more complex. To explore the relationship between maximum dependence and productivity a bit more, we can look at the relationship between productivity and dependence and how it changes over time. One way to do that is to separate scientists into groups based on their productivity over a three-year span, and then identify how maximum dependence is distributed over each group. This descriptive, cross-tabulation analysis was done for four three-year intervals: 1994–1997, 2000–2003, 2006–2009, and 2009–2012, and then the data were compared for potential changes over time.

Max dependency range	Number of publications					
	1	2	3-4	5-8	9-16	17+
==1.00	1.00	0.785	0.504	0.269	0.085	0.043
>=0.50 & < 1.00	0	0.215	0.435	0.524	0.435	0.247
>=0.25 & < 0.50	0	0	0.061	0.195	0.393	0.312
< 0.25	0	0	0	0.012	0.088	0.398

Table 5-6: Maximum dependency by number of publications for the years (1994-1997]

Table 5-6 summarizes the relationship between productivity and maximum dependency for the years (1994-1997]. As scientists become more productive they are less likely to publish with the same people repeatedly. Approximately 40% of the relationships of scientists with 17 or more publications were present on less than 25% of their papers. Only 52% of scientists who published between 9 and 16 papers, and 29% of those who published more than 17 articles, had at least one scientist with whom they published more than half their papers. The numbers for the most productive groups changed by 2000.

Max dependency range	Number of publications					
	1	2	3-4	5-8	9-16	17+
==1.00	1.00	0.77	0.496	0.271	0.138	0.079
>=0.50 & < 1.00	0	0.23	0.437	0.503	0.468	0.425
>=0.25 & < 0.50	0	0	0.067	0.214	0.316	0.231
< 0.25	0	0	0	0.012	0.079	0.266

Table 5-7: Maximum dependency by number of publications for the year (2000-2003]

Table 5-7 provides the same information as Table 5-6 for the years (2000-2003]. The figures for the least productive scientists are very similar to the (1994-1997] time period. Somewhat surprisingly, approximately 8% of the scientists who had 17 or more publications during the time period had at least one colleague who was a coauthor on every paper. For any given range of publications, over 50% of the authors were more than 50% dependent on at least one other scientist. For any short period of time, scientists will tend to work with relatively stable groups, often relying on a set of individuals to support their productivity. This suggests that highly prolific authors have a partner who contributes to their productivity, or that within groups that are arranged hierarchically, authors at the middle management level are more likely to collaborate on larger portions of papers. For example, the most productive author coauthored only one paper with 199 authors, 2 papers with another 20 authors, 3 papers with another 52 authors, and had a common coauthor on 79, 80, 660, and 1574 papers during the time period. The scientist who coauthored 660 papers with the most productive author in turn coauthored with another scientist on 658 out of 660 papers, 2 other authors on 76 papers each, and 77 authors only once. In this situation, the subordinate scientist collaborated with the more senior scientist on 100% of her papers, but was present on only 40% of the more senior scientist's papers. The pattern repeats itself for two other authors in the lab, suggesting that in very productive labs, collaboration is coordinated around a handful of individuals.

The numbers change slightly if the year range is expanded, with most of the changes occurring in the upper ranges (Table 5-8). For the middle ranges, the percentage of scientists who exhibit higher maximum dependency increased slightly, indicating that scientists whose productivity per year is lower (e.g., 5 publications in 9 years vs. 9 publications in 3 years) are more reliant on at least one other scientist than scientists with higher productivity rates. Scientists with higher productivity rates see a decrease in their maximum dependency, indicating that they expand their collaboration network over time.

	Number of publications					
Max dependency range	1	2	3-4	5-8	9-16	17+
==1.00	1.00	0.773	0.508	0.263	0.126	0.046
>=0.50 & < 1.00	0	0.227	0.435	0.555	0.514	0.375
>=0.25 & < 0.50	0	0	0.057	0.172	0.314	0.345
< 0.25	0	0	0	0.009	0.046	0.233

Table 5-8: Maximum dependency by productivity for the years (2006-2009]

The years 2000–2003 were important years as scientists rushed to sequence the human and mouse genomes. The mean number of authors per paper were very high for these years, declining slightly afterwards (Table 5-9). A larger percentage of scientists in the midrange of productivity were dependent on at least one other scientist. In contrast, a smaller percentage of scientists at the upper end of the productivity range were dependent on someone for a majority of

their publications, although the percentage is much higher (44% vs 29%) for these years than for the years 1994–1997].

Max dependency range	Number of publications					
	1	2	3-4	5-8	9-16	17+
==1.00	1	0.780	0.515	0.273	0.107	0.068
>=0.50 & < 1.00	0	0.220	0.425	0.529	0.484	0.372
>=0.25 & < 0.50	0	0	0.061	0.187	0.335	0.230
< 0.25	0	0	0	0.01	0.074	0.330

Table 5-9: Maximum dependency by number of publications for the year (2009–2012]

Based on the data above, not much changed for the dependence of scientists at the lower end of the spectrum in terms of productivity. Scientists at the upper end of the spectrum in terms of productivity, exhibited, as a group, some oscillations in the distribution of maximum dependence with no clear pattern emerging. That is to say, the relationship between how scientists structure their coauthorship relationships and their resultant productivity levels is ambiguous.

Instead of looking at the relationship of maximum dependence to productivity, it might be useful to explore its relationship to a measure of the general distribution of dependence, with the scientist as the unit of analysis. The median dependence gives us some insight into the general distribution of a scientist's dependence on others to publish. If the scientist works with a core group of coauthors, he or she more often than not will publish with only those authors, and his or her dependence scores will be skewed toward the higher end. Conversely, if a scientist tends to publish papers with a diverse set of actors and rarely publishes with the same person twice, then the distribution of his or her dependence scores will be skewed toward the lower end of the spectrum. We'll be able to see the interplay between maximum and median dependence

throughout this chapter, and how taken together they provide insight into the publication networks of scientists.

A plot of the relationship between the maximum and median dependence of scientists reveals that there are several common mixes of maximum-median dependencies, demonstrating that scientists distribute their relationships in different ways. Figure 5-5 is a plot of the relationship between median and maximum dependence of scientists' cumulative relationships, with node sizes scaled based on number of instances. The most common intersection of median and maximum dependence was at 1.00, which covers 55.5% (673,169) of instances. The common points of intersection are at the reciprocals of common productivity numbers (2–5 publications).

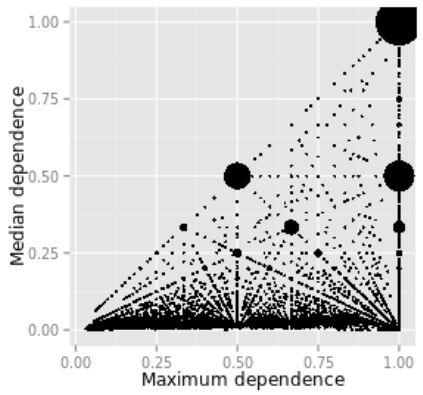


Figure 5-5: Relationship between maximum and median dependence, aggregated with point size proportional to the number of instances.

A comparison of median and maximum dependence scores shows that there are common intersection points describing the distribution of dependence across all of a scientist's collaborators. However, the results up to this point do not help us differentiate between scientists in terms of how they experience dependence. The theoretical framework suggests that there is a

strong modular structure in any collaboration network, and that a scientist's dependence should be related to their position within that group structure. The investigation into the interplay between the modular structure, scientists' positions within that structure, and their dependence are discussed in the next section.

5.3.2 Position within the group structure and dependence

The full description of the methodological approach to calculating a scientist's position within the group structure is outlined in §4.2.5, but a summary here is useful. First, the group structure of the network is identified using the community detection algorithm, and with each scientists' home group. Each scientist has a total number of connections, with some going to others within his or her home group, and others possibly extending to scientists outside the home group. Two variables are calculated from the distribution of links within and between groups—the within-module degree and the participation coefficient (Guimerà et al., 2007a). The scores of those two variables are used to classify scientists into one of seven roles that fall into two broad categories—hubs and non-hubs; the classification scheme is reprinted from the methodology section in Table 5-10 for ease of reference.

		P	z_i
NON-HUBS	(R1) Ultra-peripheral nodes	$P \leq 0.05$	< 2.5
	(R2) Peripheral nodes	$0.05 < P \leq 0.62$	< 2.5
	(R3) Satellite connectors	$0.62 < P \leq 0.80$	< 2.5
	(R4) Kinless nodes	$P > 0.80$	< 2.5
HUBS	(R5) Provincial hubs	$P \leq 0.30$	>= 2.5
	(R6) Connector hubs	$0.30 < P \leq 0.75$	>= 2.5
	(R7) Global hubs	$P > 0.75$	>= 2.5

Table 5-10: Node role assignment based on the Participation coefficient (P) and within-module degree (z_i)

Figure 5-6 presents the distribution of node roles for the years 1994–2009, taken every three years. For this figure, node role 0 is a special case where the intra-module degree has zero

variance (so the denominator is 0), making the within-module degree impossible to calculate. The distributions observed in this study differ from previous studies (Velden et al., 2010), with far larger percentages in Roles 3 and 4, as well as in 6 and 7. The observed distributions support the argument that the tendency of intergroup collaboration increases all participants' connections within and between groups. This is further supported by the fact that Role 1's connections declined as a percentage of the entire population steadily, while Role 3 and Role 4 increased steadily as a percentage of the population, with the former moving from approximately 17% in 1994 to 25% in 2009, and the latter 5% to 13%. Simultaneously, hub roles remained flat as a relative portion of the population.

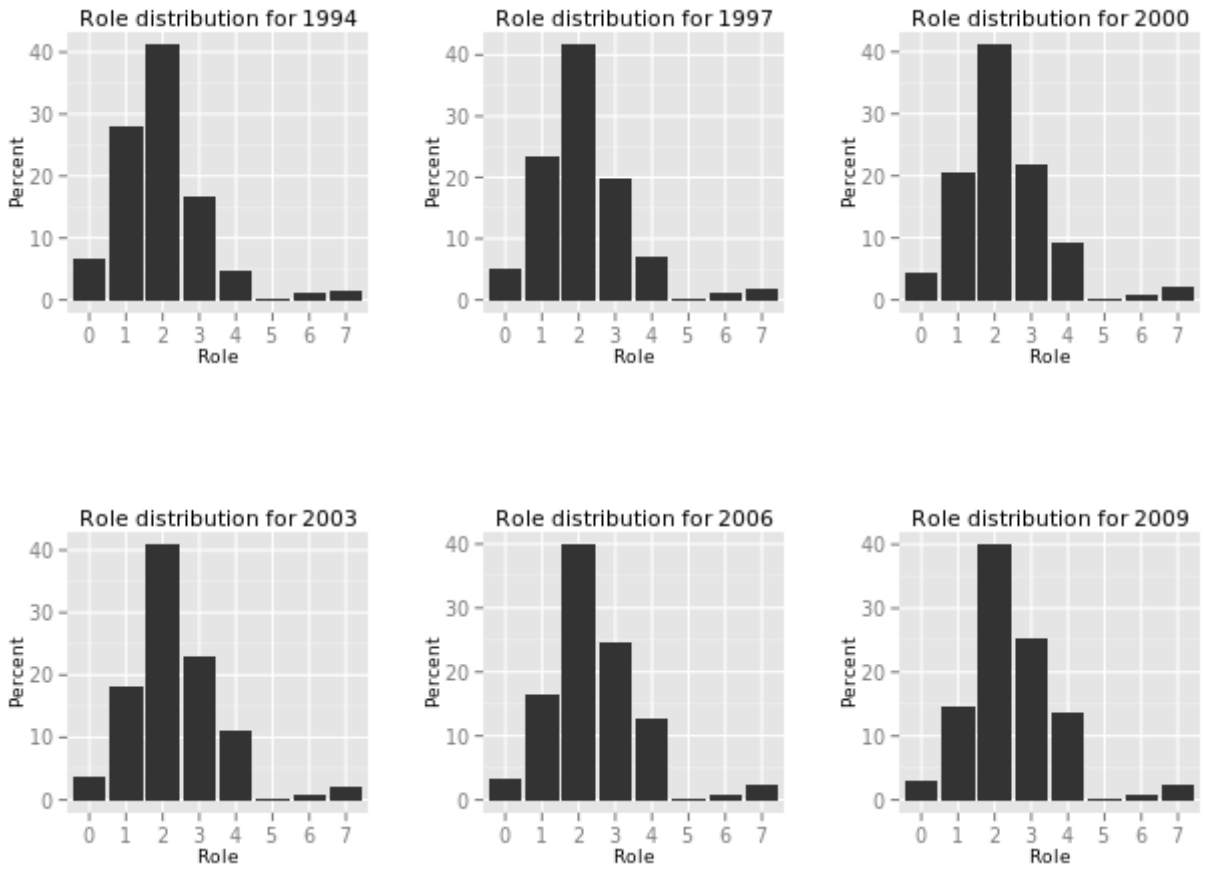


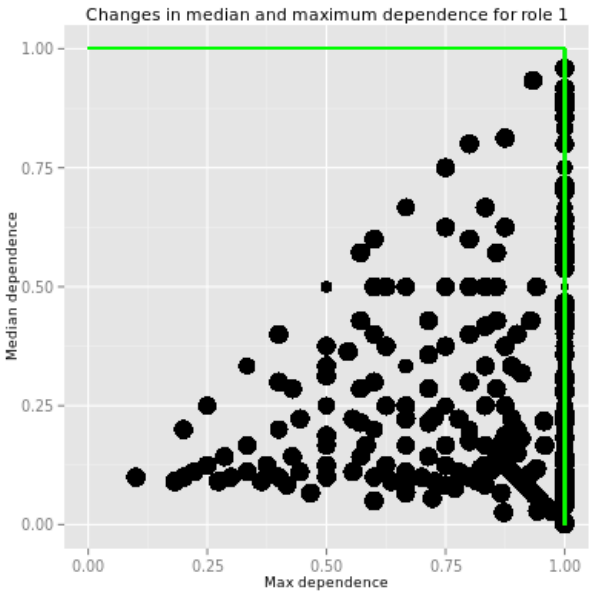
Figure 5-6: Node role distributions for years 1994–2009

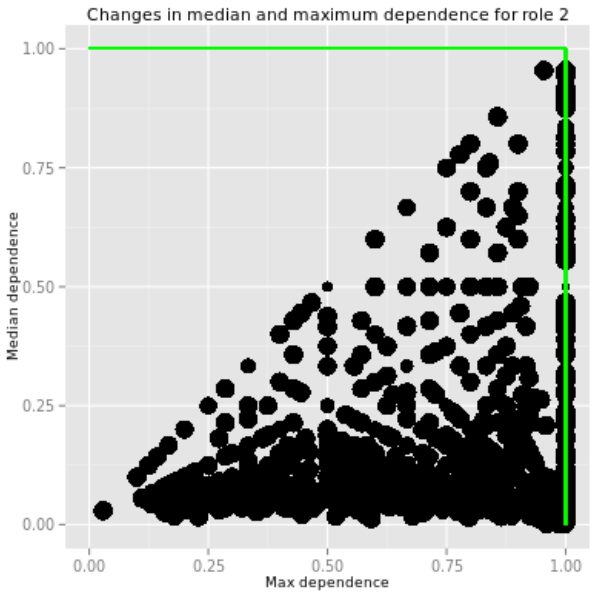
Role distributions changed over time, with a pronounced shift in the distribution of non-hubs toward roles with higher participation coefficients. The observed shift in role distribution can be attributed to increasing intergroup collaborations, which can be seen in the analysis in §5.2, where approximately 50% of papers had no coauthors from the same group. The participation coefficient is calculated directly on intergroup coauthorships, and is not normalized.

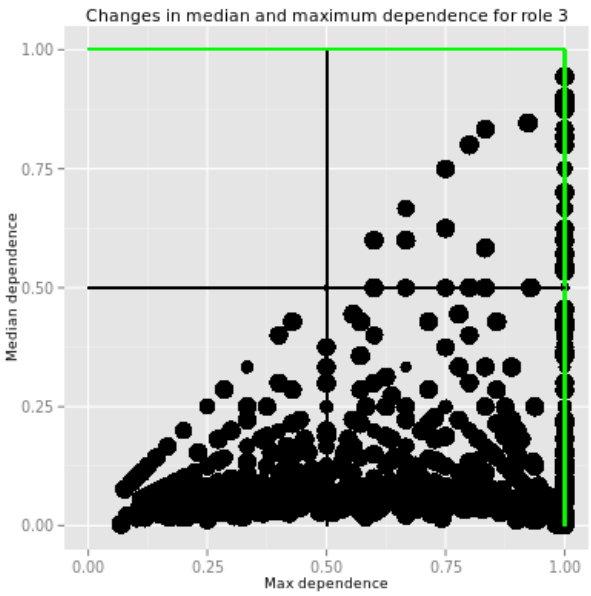
Comparing the median and maximum dependence of scientists is a useful way of answering the questions posed above. Figure 5-7contains plots of median versus maximum dependence for each role, for the years 1994 (black), 2000 (red), and 2009 (green). The

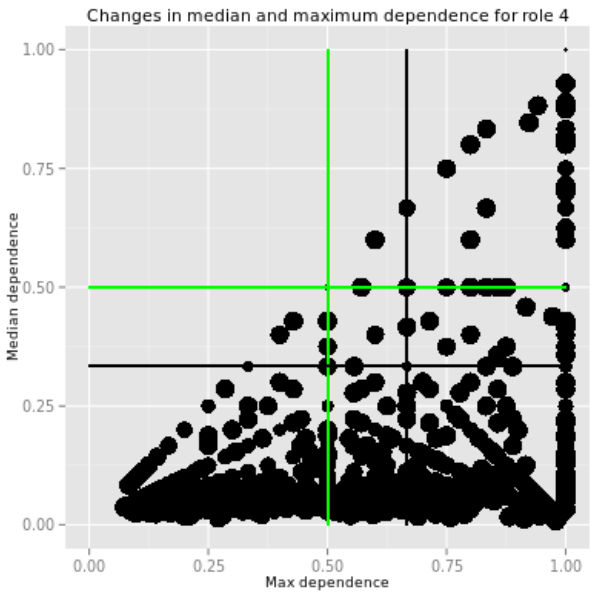
intersection of the lines creates quadrants on the graph; the upper right quadrant contains 50% of the population for that role. An important point to note is that the dependence values observed are for forward collaborations. Node roles were calculated on historical interactions up until the year listed, but dependence scores were calculated on behaviors from the year in question up to, but not including, three years in the future. As an example, the dependence scores observed for 1994 are for the coauthorships observed from 1994–1997.

Scientists in Roles 1 and 2 were consistently highly dependent on their network of connections, with a vast majority coauthoring every paper with the same group of scientists. A subset of scientists in the ultra-peripheral role were not as dependent on others, but they only make up 10.1% of scientists in Role 1 and 19.1% in Role 2. It is important to note that newcomers are not classified into either role, but instead are classified separately. Only 4.1% of newcomers were not fully dependent on their relationships, while 11.4% of the unclassifiable scientists (see above) did not get placed in the upper right corner.


Changes in median and maximum dependence for role 1
Median dependence
Max dependence
0.00
0.25
0.50
0.75
1.00


Changes in median and maximum dependence for role 2
Median dependence
Max dependence
0.00
0.25
0.50
0.75
1.00


Changes in median and maximum dependence for role 3
Median dependence
Max dependence
0.00
0.25
0.50
0.75
1.00


Changes in median and maximum dependence for role 4
Median dependence
Max dependence
0.00
0.25
0.50
0.75
1.00

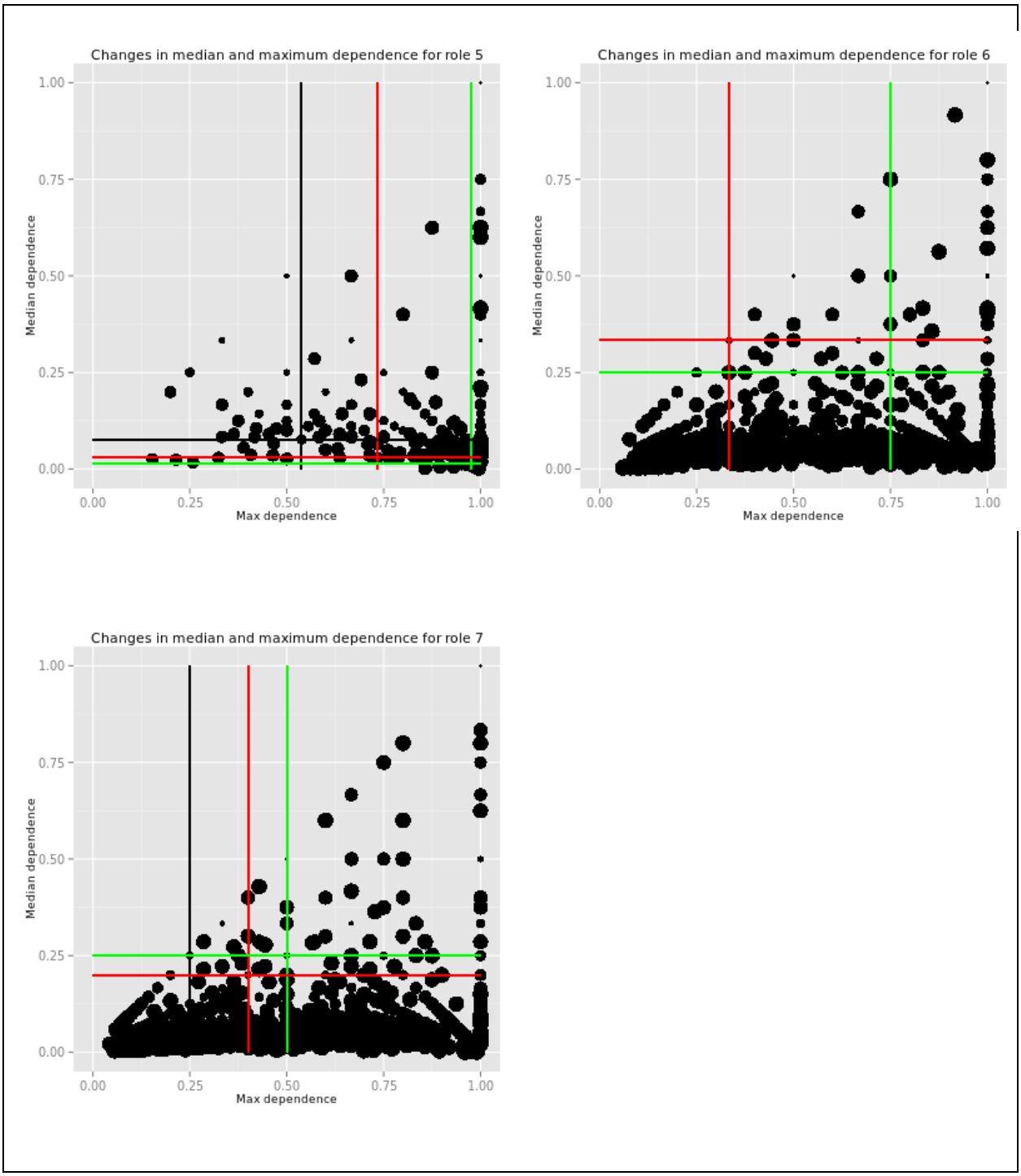


Figure 5-7: Maximum vs median dependence of scientists, by role. The lines represent the median of the range of values, where the upper right quadrant contains 50% of the population. Black is for the year 1994, Red 2000, Green 2009. 1994 & 2000 overlap for Role 6,

Scientists in Role 3 had one of the most dramatic shifts, moving from partial dependency, being in general no more than 50% dependent on any one contact, with a relatively evenly

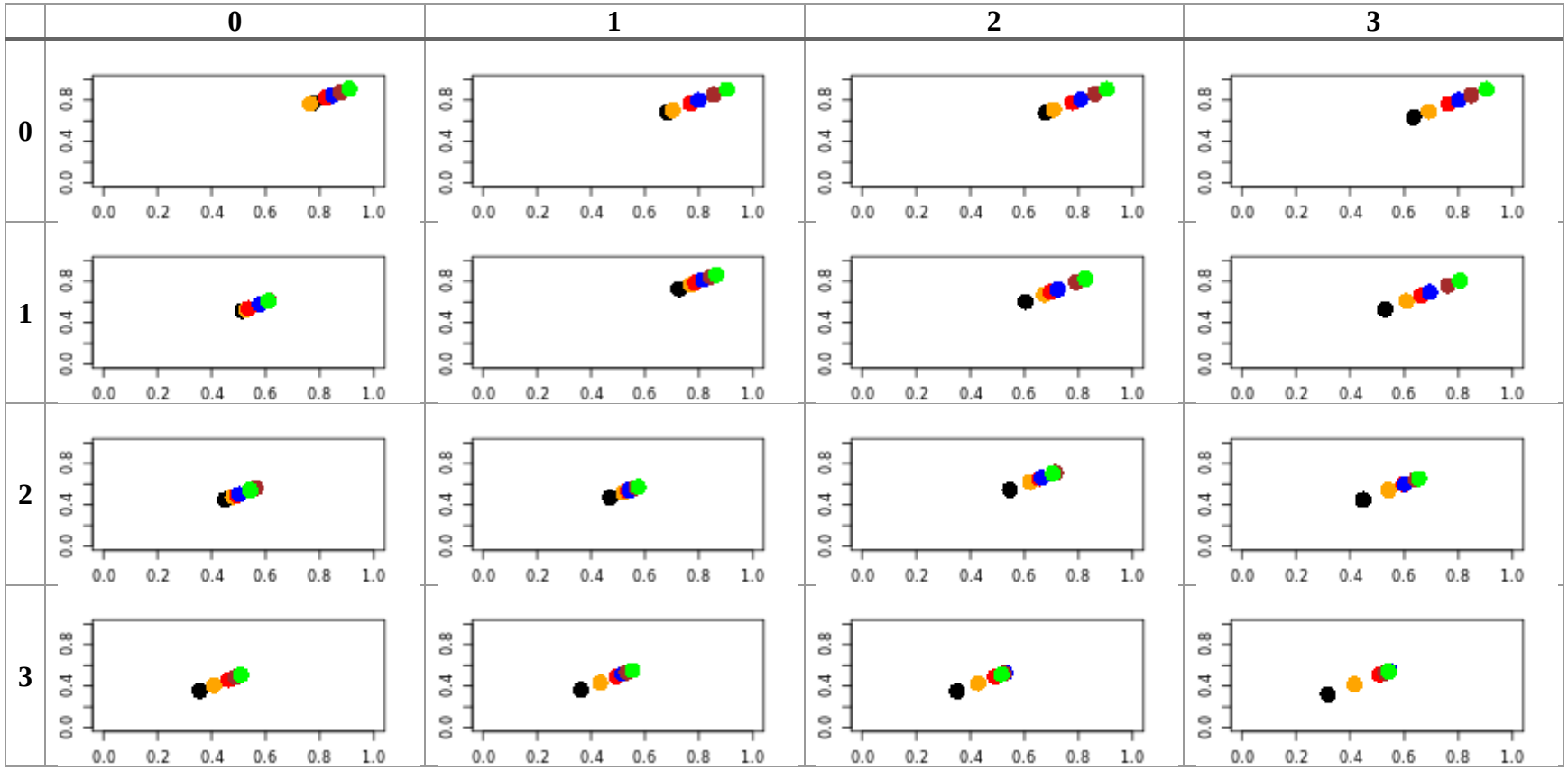
distributed set of dependencies, to a state of high dependence on a majority of their connections. This means that even though scientists within this group were likely to collaborate with scientists from different groups, they were more likely to do so frequently and exclusively with those scientists. Scientists in Role 4 became less dependent on any single person, but in general experienced a smoothing of the distribution of their connections. That is, they were more likely to be dependent on a larger portion of their connections. Scientists within this role may have had more opportunities to collaborate widely, but were also more likely to reuse those relationships more often.

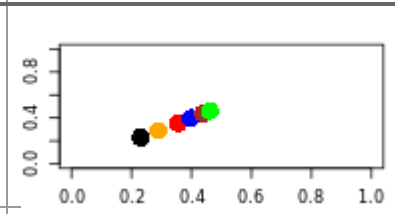
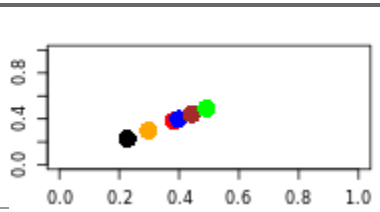
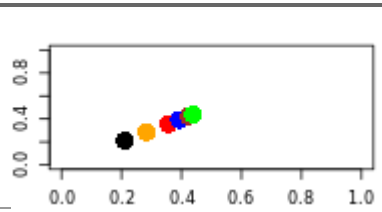
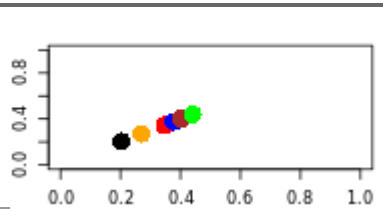
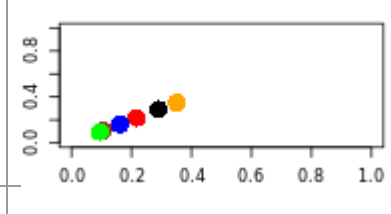
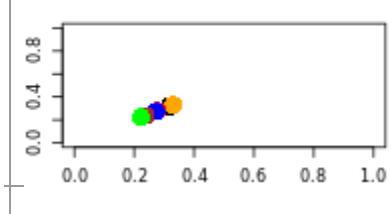
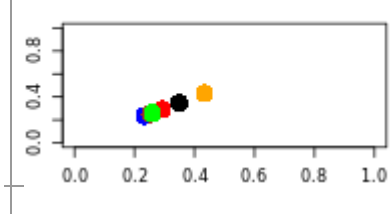
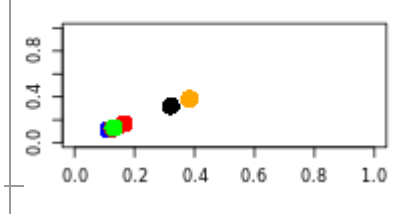
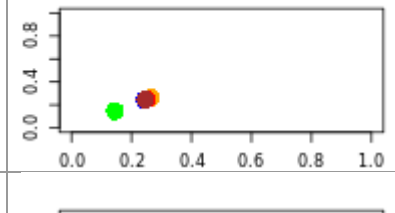
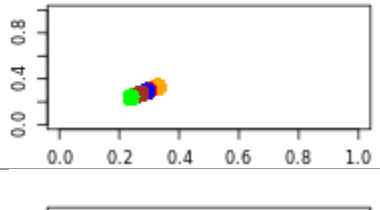
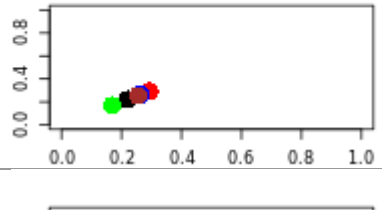
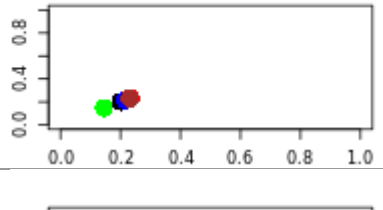
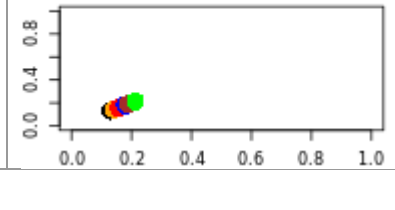
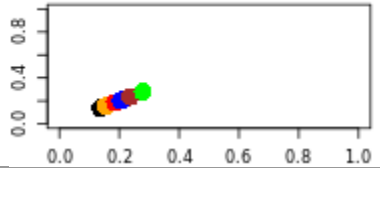
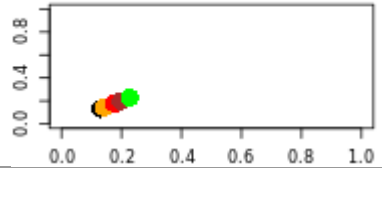
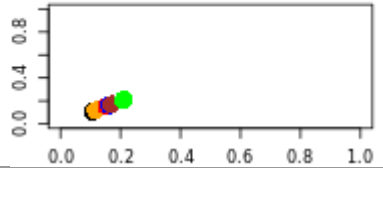
The patterns for the hub roles are much different than those for the non-hub roles. Role 5 is the smallest in terms of percentage of the population, and it appears to be comprised of scientists who became increasingly dependent on a small subset of their coauthorship network, but less dependent overall on their other coauthors. A similar, yet less pronounced pattern emerges for Roles 6 and 7. Scientists in Role 6 were moderately dependent on a subset of their coauthors and exhibit a relatively even distribution of dependencies on all their coauthors through 2000. However, by 2009 scientists in Role 6 became less dependent on a majority of their connections, but more dependent on a smaller core group of coauthors. The distribution for Role 7 shows a slightly different change. Scientists in this role became more dependent on a core group of scientists, but their overall dependence declined in 2000, which coincided with the increased activity around the sequencing of the mouse and human genomes, only to increase again by 2009. The importance of the core group of collaborators increased for hubs, but the role of the transitioning scientists returned to a relatively constant level.

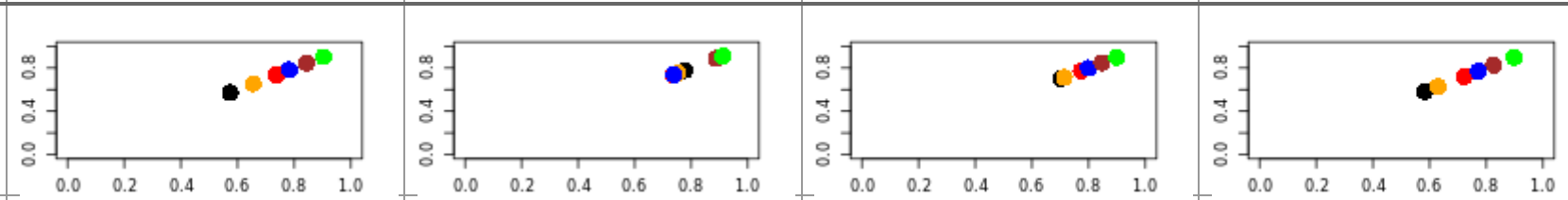
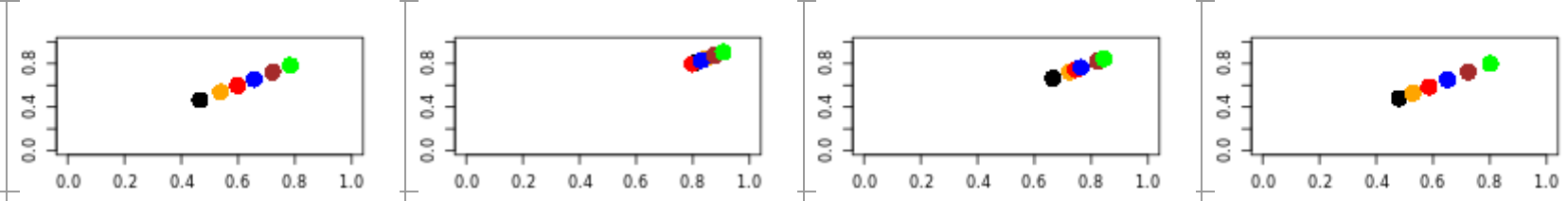
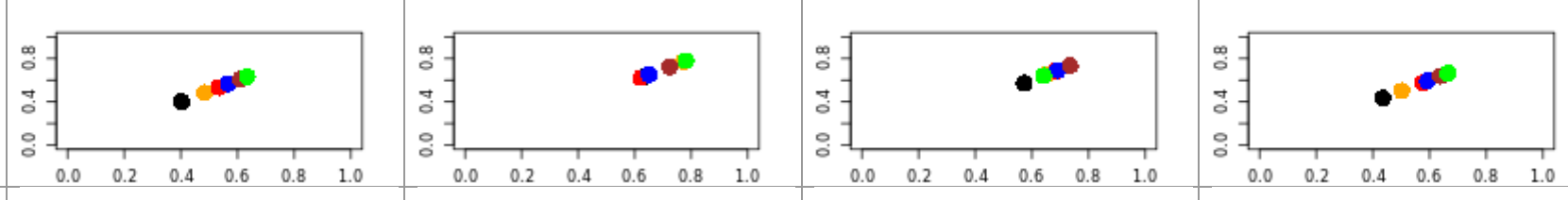
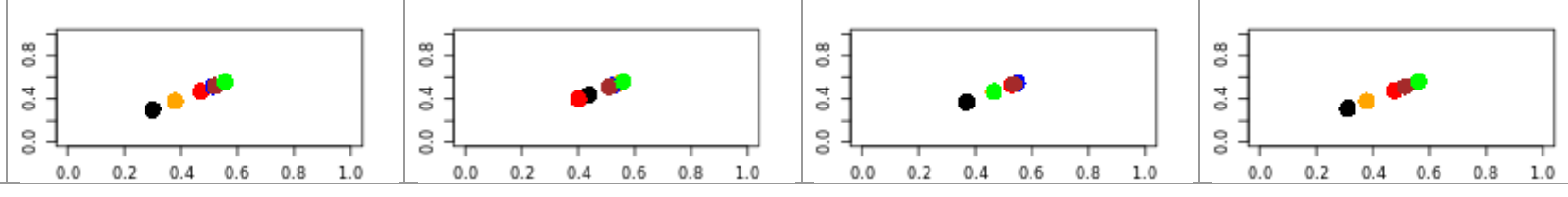
Figure 5-8 highlights the role-to-role dependencies, using the median of median scores and median of maximum scores for scientists in role *x* (in the rows) on role *y* (in the columns).

The intersection of median and maximum scores are color coded by year. There is a substantial amount of information in the graphs; only a summary will be provided here. The interpretation is covered in greater detail in Chapter 6, where the data in the figures help clarify the interpretation of other data.

Figure 5-8: Role-to-role dependencies, by year; y-axis is median dependence, x-axis is mean dependence. Rows are the source, Columns are the target roles. Role 0 indicates newcomers. Key: Black (1994), orange (1997), red (2000), blue (2003), brown (2006), green (2009).



	0	1	2	3
4				
5				
6				
7				

	4	5	6	7
0				
1				
2				
3				

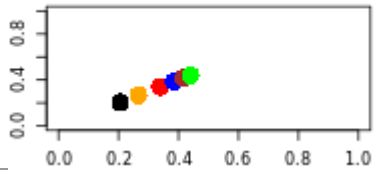
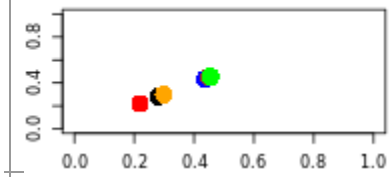
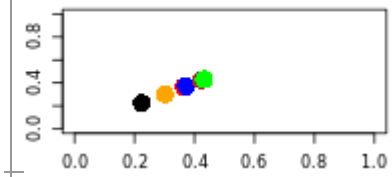
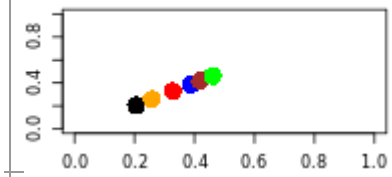
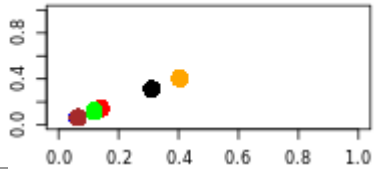
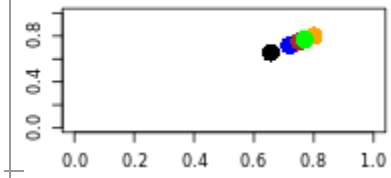
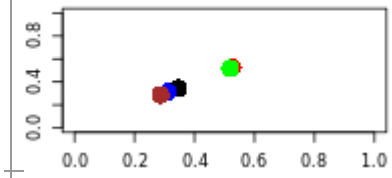
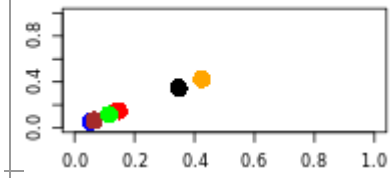
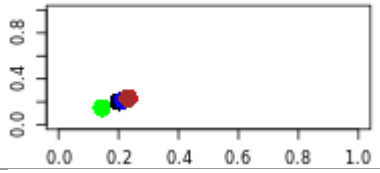
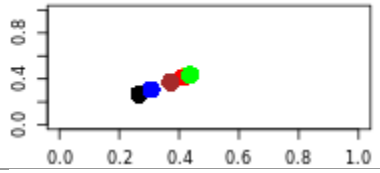
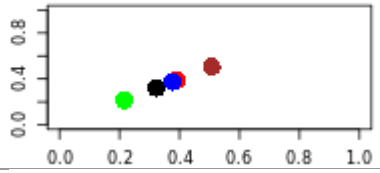
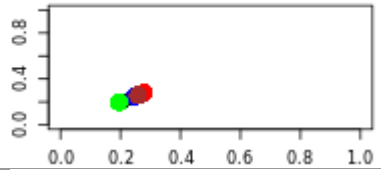
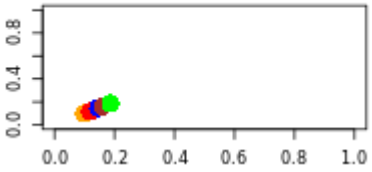
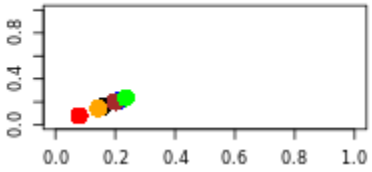
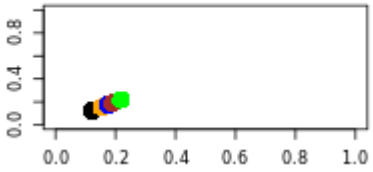
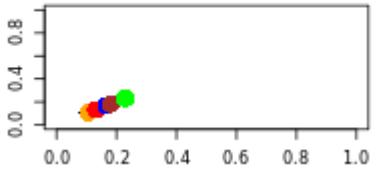
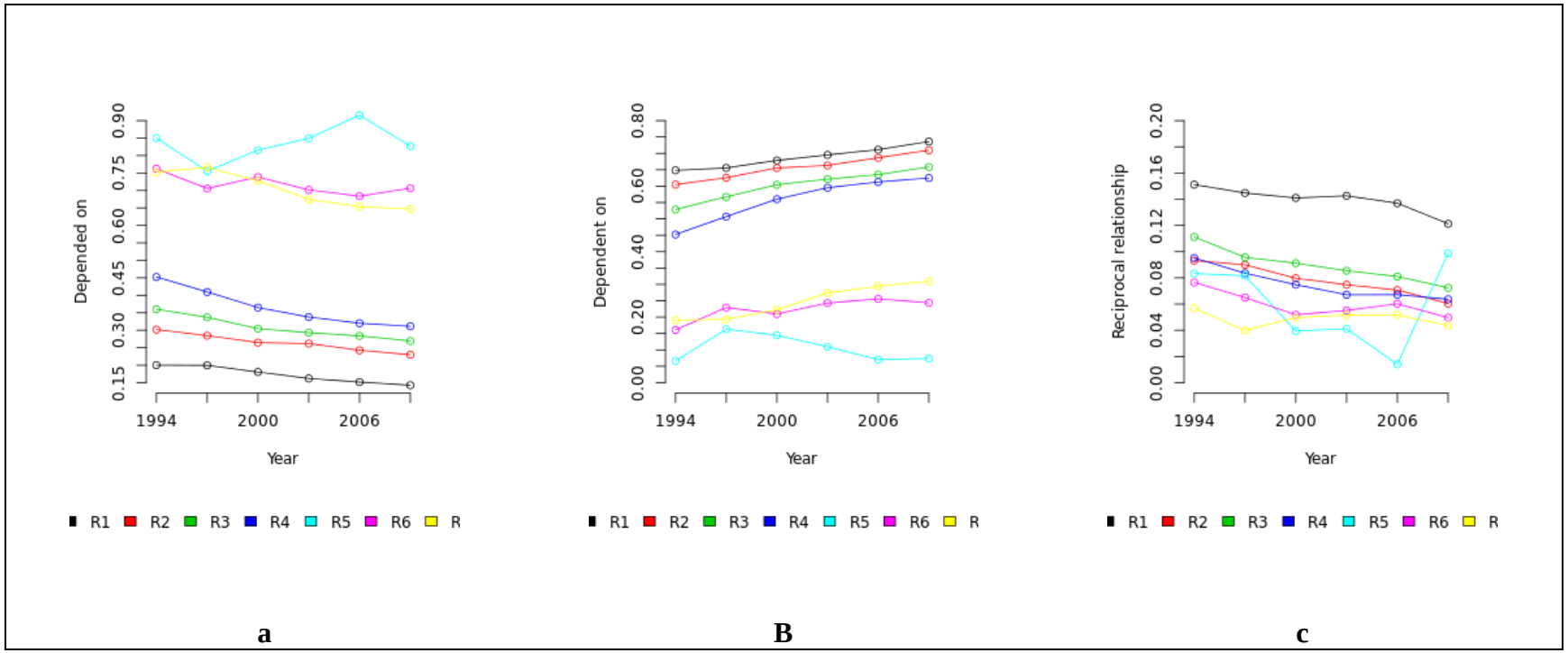
	4	5	6	7
4				
5				
6				
7				

Figure 5-9: The relative proportion of scientists who have more relationships in which they are dependent on (a); dependent on (b); or have a reciprocal relationship with (c).



Continuing with Figure 5-8, scientists in Roles 0, 1, and 2 became increasingly dependent on at least one other scientist, and saw an overall upward shift in their distributions of dependence over the years. Scientists in Role 2 did not experience a significant uptick in dependence on scientists in Role 6 and Role 1. Scientists in Role 3 saw a general increase in dependence on scientists in all Roles along both dimensions. Scientists in Role 4 also saw an increase in dependence along both dimensions, but that increase ended about midrange of the scores for all inter-role calculations. Scientists in Role 5 had a different pattern than other Roles, decreasing in dependence on scientists in all non-hubs roles and on global hubs (Role 7s), but increasing in dependence on scientists in Roles 5 and 6. Scientists in Role 6 were generally less dependent on non-hubs and Role 7s, but more dependent on scientists in Role 5. Scientists in Role 7 were less dependent on scientists in non-hub roles, but slightly more dependent on scientists in hub roles.

Dependence is always bidirectional, but not equivalent. If *Scientist A* writes a paper with *Scientist B*, *A* is a coauthor of *B*, and vice versa. However, *A* may be on more publications of *B*'s than the other way around. It is possible to model the disparities in dependence by creating a directed, instead of undirected, network. In a directed network, a relationship is only present from *A* to *B* if *A* is more dependent on *B* than the other way around. In cases where the two scientists are codependent, two edges exist between the scientists, traveling in the opposite direction. Using this model, a scientist can be viewed as *depended on* (Figure 5-9a) if she has more edges coming in then leaving. In contrast, a scientist is *dependent on* (Figure 5-9b) others if she has more edges leaving than coming in. In certain cases, the relationships can be *reciprocal*, where the number of edges coming and going balance out (Figure 5-9c).

Two patterns emerge when tracking the relative distribution of dependence over time, by role (Figure 5-9). First, hubs and non-hubs occupy two distinct regions of the graphs, except for the relative proportion of relationships that are reciprocal in nature. The fact that the roles occupy two distinct regions supports the argument that the node role framework is a useful tool for exploring scientists' dependence on one another. Scientists in Role 1 were much more likely to be in reciprocal relationships. The reason for this is, as less experienced scientists, they were more likely to be on only a few publications with the same set of people, some of who were other newcomers. Hubs were much more likely to be depended on (Figure 5-9a), but the trend was for the disparity in dependence to decrease from 1994–2009. That is to say, scientists were more likely to rely on a group of collaborators more frequently; as a result, the relationships became more balanced among all actors. Scientists in each Role were more likely to be depended on more frequently (Figure 5-9 b), except for scientists in Role 5. Put another way, scientists were more likely to participate on a larger proportion of their peers' work in 2009 than in 1994.

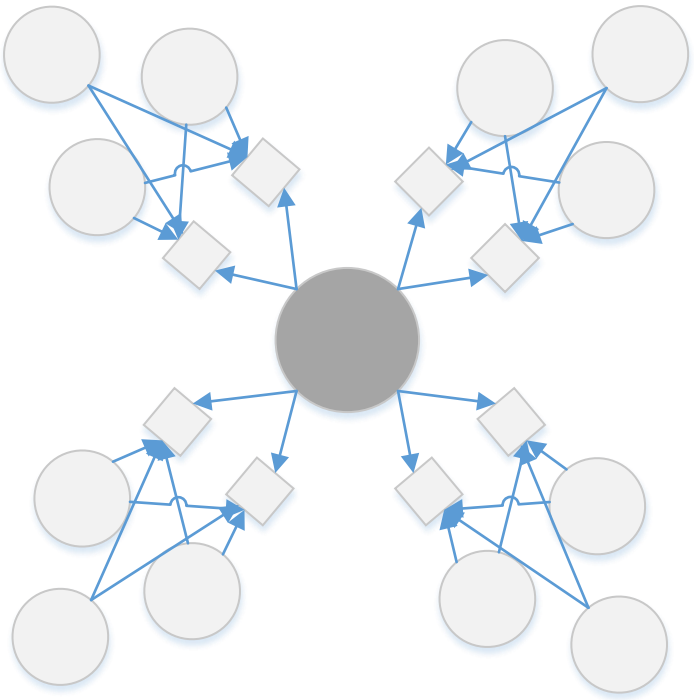
The emerging picture is that scientists within this network became more interdependent, but also that there is variance within and between roles. The following section goes into greater detail on dependence, looking at the asymmetrical nature of the measurement and how that asymmetry can be exploited to develop a more nuanced picture of the relationship between dependence and role within the group structure of the network.

5.3.3 Net dependence and the clustering coefficient

If a scientist is depended on by a large portion of her collaborators, then those collaborators rarely publish without the scientist (Figure 5-10). If the relationship is asymmetrical, then the scientist collaborates with many different people, but does not collaborate

with any one person or group frequently. In the above figure, no single scientist accounted for more than 25% of the central scientist's collaborations, while the central scientist accounted for 100% of all her neighbors' collaborations. Because the other groups of scientists rarely published without her, they had fewer relationships to other scientists. This results in a lower clustering coefficient around scientists who had asymmetrical relationships. Another way of thinking about this is that if a scientist frequently publishes with transients, her collaborators never publish again, contributing to the asymmetrical nature of dependence. Because transients do not publish again, the clustering coefficient around the scientist will be lower because transients never have the opportunity to form additional relationships.

Figure 5-10: A depiction of a local network where the primary scientist's neighbors are highly dependent on the scientist (dark gray) and isolated from one another.



It is possible to test the relationship between dependence and clustering coefficient by looking at the local, undirected relationships around scientists within a network. The most

straightforward way is to compare scientists who are largely *dependent on* others versus scientists who are *depended upon*. There should be a negative correlation between being depended on and the local clustering coefficient around the scientist. Using Pearson's correlation, r(59247) = -0.69, $p < 0.001$ for 1994 and r(123734) = -0.66, $p < 0.001$ for 2009, there is a strong negative correlation between dependence and local clustering coefficient. Scientists can also be divided into two groups based on whether they are net dependent or depended on, and then have their local clustering coefficients compared (Table 5-11).

	1994	2000	2009
NET DEPENDENT	**1.00 / 0.956**	**1.00 / 0.951**	**1.00 / 0.937**
NET DEPENDED ON	**0.467 / 0.496**	**0.467 / 0.490**	**0.455 / 0.468**

Table 5-11: Clustering coefficients for scientists based on whether they are net dependent or net depended on. Numbers reflect median and mean, respectively.

There are clear differences between scientists who were net dependent versus those who were net depended on. The clustering coefficient for dependent scientists is clustered around 1.00, with the mean and median values all above 0.93. In contrast, the clustering coefficient around scientists who were depended on resembles a Poisson distribution, with both mean and median values in the vicinity of 0.45 for all time periods. From a visual perspective, the distribution of the clustering coefficient plotted against dependence, by role, were similar for all time periods. Figure 5-12 shows the data for 2009; for the *x*-axis, negative values indicate that the scientist was net dependent on other scientists, with the absolute value indicating what fraction of relationships were dependent. Positive values indicate net depended on, with values indicating what fraction of that scientist's relationships were dependent on him or her.

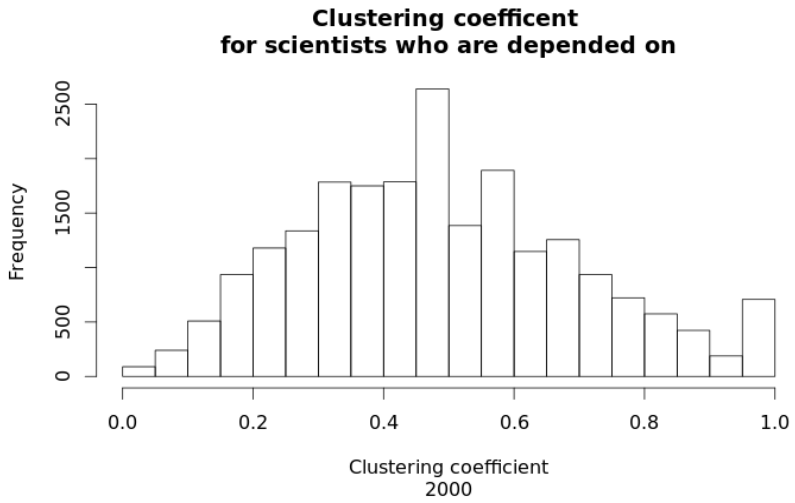
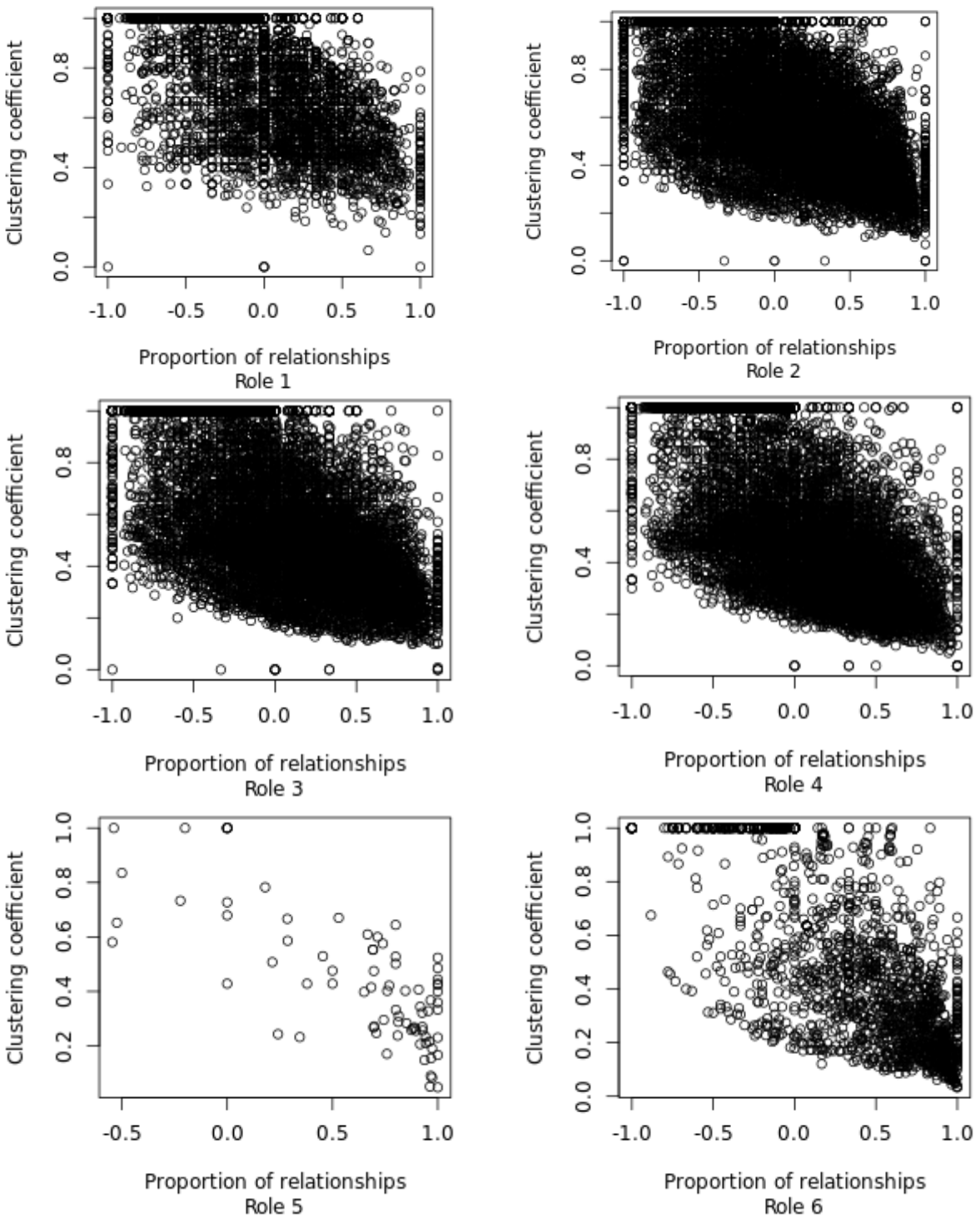


Figure 5-11: Distribution of the Clustering coefficient for scientists who are depended on, in 2000.

As expected, scientists in the hub roles were much more likely to be depended on. The clustering coefficients for scientists in the hub roles are clustered more heavily in the lower right part of the graph (Fig. 5-12), suggesting high dependence and low clustering coefficient. The lower left and upper right portions of all graphs are sparsely populated or completely uninhabited, indicating that no scientists functioned independently (lower left), or were highly depended on and part of a dense network of people who were dependent on that scientist.


Clustering coefficient
0.0
0.4
0.8
-1.0
-0.5
0.0
0.5
1.0
Proportion of relationships
Role 1
Clustering coefficient
0.0
0.4
0.8
-1.0
-0.5
0.0
0.5
1.0
Proportion of relationships
Role 2
Clustering coefficient
0.0
0.4
0.8
-1.0
-0.5
0.0
0.5
1.0
Proportion of relationships
Role 3
Clustering coefficient
0.0
0.4
0.8
-1.0
-0.5
0.0
0.5
1.0
Proportion of relationships
Role 4
Clustering coefficient
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5
1.0
Proportion of relationships
Role 5
Clustering coefficient
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0
Proportion of relationships
Role 6

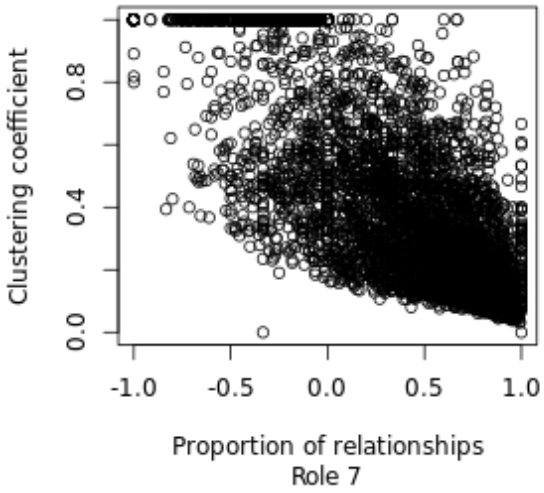


Figure 5-12: Distribution of the clustering coefficient, by role, for 2009

The role-to-role interdependence, interdependence by role, and the local clustering coefficient of scientists, by dependence, all indicate that the general trend was for scientists to become more interdependent from 1994 to 2009, for this network. However, the analysis up to this point has focused on the interdependence of scientists as dyadic actors, leaving the unanswered question: How has the dependence of scientists on the group changed for the same time period?

5.3.4 Dependence on group

Five different types of collaboration related to intergroup collaboration were identified from the perspective of the individual scientist– *between, in only, new only, out only, and out and new*. *Between* collaborations include collaborations that involved at least one other scientist from the scientist's group, as well as at least one other scientist from another research group. *In only* collaborations involved collaborations with only scientists in the same group. *New only* collaborations are those that involved the scientist and newcomers, with no other scientists from within or between groups. *Out* only collaborations involved no other scientists from within the group, and no newcomers. *Out and new* collaborations involved no other scientists from within

the group, at least one scientist external to the group, and one or more newcomer scientists. Newcomers were not classified into groups because groups were identified from collaborations up through a certain year; newcomers thus were not present in the dataset for clustering. Otherwise, all data were tabulated from the individual perspective, i.e., each collaboration was analyzed from the perspective of all participating scientists. The *In all* column includes all *Between* and *In* collaborations, following the theory that the scientists should act with their groups, but not always exclusively with their groups. The *In high* category was an estimate of the fraction of collaborations that were most likely within the research group, assuming that the *New only* category captured a scientist's collaboration with new members of his group. The same could be said for the *Out and new* category, however the assumption is not as safe because the newcomers could be associated with the other group. Table 5-12 includes information on different collaboration types and the proportion of collaborations that fall into each type, by role and year.

The trend was for scientists to frequently collaborate with at least some other scientist within their group. This trend increased moving from Role 1 to Role 4, with scientists in Roles 3 and 4 not having a majority of their collaborations within group. The pattern resets itself moving into the hubs, with scientists in Role 5 exhibiting the strongest tendency to collaborate within group. Scientists within Role 5 also participated in more between group collaborations as a relative proportion of their collaborations, despite, by definition, having a lower participation coefficient (connections between groups). To test whether there were differences in the tendency to collaborate within group over time, non-paired, two-tailed t-tests were conducted, by role, for the years 1994 and 2009. All scientists that were in the same

ROLE	YEAR	BETWEEN	IN	NEW ONLY	OUT	OUT & NEW	IN ALL	IN HIGH
1	1994	0.270	0.275	0.068	0.067	0.316	0.544	0.612
1	2000	0.315	0.261	0.050	0.069	0.297	0.576	0.626
1	2009	0.368	0.251	0.027	0.072	0.280	0.619	0.646
	t(13453) = -8.453, 0.95 CI[-0.083, -0.052]*							
2	1994	0.336	0.161	0.061	0.081	0.357	0.496	0.558
2	2000	0.402	0.147	0.042	0.077	0.328	0.548	0.590
2	2009	0.494	0.136	0.018	0.082	0.267	0.630	0.649
	T(19428) = -17.75, 0.95 CI[-0.107,-0.086]*							
3	1994	0.251	0.083	0.064	0.121	0.476	0.335	0.399
3	2000	0.300	0.101	0.042	0.123	0.430	0.400	0.442
3	2009	0.397	0.085	0.027	0.140	0.348	0.482	0.509
	T(7428) = -12.69, 0.95[-0.104,-0.076]*							
4	1994	0.176	0.045	0.038	0.178	0.559	0.222	0.260
4	2000	0.219	0.037	0.027	0.219	0.490	0.257	0.284
4	2009	0.300	0.045	0.015	0.211	0.426	0.345	0.361
	T(877) = -7.082, 0.95[-0.113,-0.064]*							
5	1994	0.514	0.291	0.064	0.008	0.120	0.805	0.869
5	2000	0.443	0.281	0.019	0.066	0.191	0.724	0.743
5	2009	0.870	0.096	0.000	0.010	0.024	0.966	0.966
	t(80) = -2.86, 0.95 CI[-0.298, -0.054]*							
6	1994	0.412	0.172	0.063	0.066	0.284	0.583	0.646
6	2000	0.438	0.179	0.040	0.072	0.268	0.617	0.658
6	2009	0.621	0.173	0.006	0.081	0.118	0.793	0.800
	t(1420) = -3.36, 0.95 CI[-0.092, -0.024]*							
7	1994	0.249	0.064	0.036	0.159	0.490	0.314	0.349
7	2000	0.202	0.199	0.015	0.193	0.387	0.401	0.416
7	2009	0.369	0.073	0.011	0.207	0.338	0.442	0.454
	t(494) = -1.42, 0.95 CI[-0.068, 0.011]**							

Table 5-12: Types of collaborations, by role. BETWEEN indicates that the scientist within that role collaborated with scientists in his/her module as well as an external module; IN includes collaborations where all participants were internal to the group; NEW ONLY means the scientist collaborated only with newcomers; OUT includes only scientists external to the group; OUT & NEW includes scientists external to the group and unassigned newcomers; IN ALL includes all collaborations that involve between group collaborations and within group collaborations; The IN HIGH is equal to IN ALL + NEW ONLY. T-tests were between the years 1994 and 2009, no paired samples. **p< 0.01, ** not statistically significant*

role at both time points were excluded for analysis. Results are also in Table 5-12, underneath the results for each role. All results are significant to $p < .001$, except for those of Role 7, which are not statistically significant. All confidence intervals are within the negative range, indicating

that the mean within group collaboration was lower for the (1994–1997) period than for the (2009–2012) time period. There was a statistically significant tendency for scientists in all Roles, except for the global hubs, to collaborate more frequently in their own group in the years (2009–2012).

The fact that between group collaborations make up a large portion of collaborations supports Beaver's (2001) argument that the mode of collaboration is 2, if one were to consider groups the fundamental actors within a research community. Intergroup collaborations were common in this network; however, participation on intergroup research varied by both Role and year. Scientists in Roles 3, 4, and 7 were more likely to participate in the *Out and new* type of collaborations. It is difficult to tell from this data, but this could imply that scientists in these roles arranged between-group collaborations while relying heavily on newcomers within their own groups to provide the labor.

5.4 Summary

The strength of ties in a coauthorship network can be modeled using an exponential decay function, providing support for the hypothesis that collaborative relationships are subject to decay. The data also indicate that the decay rate closely fits a half-life of one year for relationships that were two or more years old. The purpose for asking whether collaborative relationships are subject to decay was to see if we can use the concept of half-life to create evolving network representations that more accurately reflect the structure that influences the collaborative interactions of scientists. Here, 'more accurately' refers to existing approaches in the literature, which include cumulative (Holme & Saramäki, 2012) and effective (Tomassini & Luthi, 2007) network representations.

An experiment was conducted to test which of the three approaches to constructing networks—decay based on half-life, cumulative, and effective—would serve as a better foundation for tracking the evolving mesoscopic layer of the network. The Infomap (Rosvall, 2014; Rosvall & Bergstrom, 2007) community detection algorithm was used to extract the group structure on six types of network representations ([cumulative, effective, decay] x [bipartite, unimodal]) at different points in time in the network's history. The modular solution was tested for its ability to predict future collaborations, where prediction meant capturing a larger portion of collaborations within module (true positive rate) and placing fewer scientists into the same module if they did not collaborate (false positive rate). The results of this experiment do not clearly demonstrate the superiority of one approach over another. Bipartite solutions had higher TPR and FPR, but also produced modular solutions that were closer in size to what we would intuitively expect for a research group (a maximum of ~100–200 versus ~300–1300). Although the results are not conclusive, they are still useful because the results demonstrate that breaking with the tradition of using unimodal cumulative networks in favor of the more theoretically consistent approach of using bipartite networks is not detrimental to subsequent analysis.

The bipartite decay approach to constructing networks was used to study the nature of dependence in a scientific coauthorship network emerging around GenBank, the international nucleotide sequencing databank. Here, a scientist's dependence on another scientist was operationalized as the ratio of the first scientist's papers that the second scientist was a coauthor on. The basic idea was to measure the portion of a scientist's publication productivity that could be attributed to the presence of another scientist, under the assumption that the presence indicated that the second scientist was providing either resources, access to equipment, specialized skills, or labor. Maximum dependence was used to measure the extent to which a

scientist relied on at least one other individual, or put another way, maximum dependence reflects that most of the scientist's productivity could be attributed to coauthoring with their most frequent coauthor.

In terms of productivity, less productive scientists had much higher rates of maximum dependence than more productive scientists for all years studied. More productive authors, in contrast, did not exhibit consistent patterns of maximum dependence, nor was there a consistent trend over the years. Because the relationship between productivity and maximum dependence was not straightforward, median dependence was introduced as a way to describe scientists' distribution of dependence among all their coauthors. A graph of the relationship between median and maximum dependence revealed that there were common areas of intersection, but no further detail on what those areas of intersection implied could be extracted from these measures alone.

The thesis of this dissertation was that using the node role framework to classify scientists based on their position within the mesoscopic structure of the network would help us understand the nature of dependence in a collaboration network (where collaboration was operationalized as coauthorship). Basic descriptive analysis of the trends in distribution of roles within the population of this network indicate that there was a significant shift in role assignment. From 1994 to 2009, the distribution into various roles moved rightward, placing more scientists into Roles 3 and 4, which, taken with other measures, indicates that more intergroup collaborations occurred and that even the most junior of scientists participated in those collaborations. This observation is not surprising, given the comments by (Beaver, 2001) and the findings in the empirical work by (Velden et al., 2010).

In terms of trends, scientists in the peripheral roles were heavily dependent on almost their entire network of collaborators, indicating that they tended to publish with the same set of coauthors (Figure 5-7Scientists in Role 3 had one of the most pronounced shifts, moving from a mixed distribution of dependence in 1994 where no author tended to account for more than 50% of their publications, and their general dependence was distributed across their entire set of coauthors, to a pattern where they were highly dependent on all of their coauthors in 2009. Scientists in Role 4 became less dependent on any one coauthor, but their general dependence across all coauthors increased. Maximum dependence increased for all hubs (Roles 5–7) from 1994–2009, while median dependence decreased. This suggests that scientists in hub roles are more likely to collaborate with one or more coauthors frequently, but less likely to collaborate with the remainder of their coauthors frequently. Furthermore, it suggests that scientists in the hub-roles are much more likely to be depended on (although the trend was slightly negative for Roles 6 and 7) but become more dependent on others as time passes. Also suggested is that scientists in non-hub roles are less-depended on over the years, and much more dependent on others as time passes (Figure 5-9).

Analysis of the clustering coefficient around scientists in different roles supports the interpretation of the observations on median and maximum dependence. That is, non-hubs had high clustering coefficients, indicating that most of their coauthors were connected to each other. The clustering coefficient around hubs was more evenly distributed, closely resembling a Poisson distribution. The more dependent a scientist was on others, the higher their local clustering coefficient was, with hubs' clustering coefficients concentrated toward the lower end.

Analysis of dependence by roles over time, as well as net dependence and the clustering coefficient, all suggest that there was a hollowing out of the middle. Scientists in the lower roles

increasingly coauthored more of their papers with the same set of individuals, and those individuals were more likely to be connected with one another. Scientists in hub roles were more likely to coauthor a majority of their papers with a small group of individuals, with the rest of their network accounting for very little of their productivity. Furthermore, scientists in all roles were much more likely over time to see a larger portion of their collaborations happen within group, with statistically significant changes observed for all Roles except 7. The shift was especially pronounced for scientists who fell into Role 5. Interesting patterns in out-group collaborations were observed, with some non-hubs having a higher percentage of out-group collaborations than scientists who were in Roles 5 and 6.

Interpretation and analysis of these patterns are provided in Chapter 6.

6 Discussion

6.1 Overview

Two theoretical models of scientific collaboration were outlined in Chapter 3: one based on dyadic interactions, the other based on the assumption that collaboration is organized around the research group. The former model assumes that scientists negotiate collaborative relationships pairwise, even if the research team on any given project involves three or more people. The latter model assumes that research teams are larger and organized around either stable research groups or short term assemblages of two or more groups. What has been argued in the research presented in this dissertation is that the group model would create dependencies in the system, where scientists who did not possess the ability or desire to organize group projects would rely on those who could arrange and provide access to the projects.

The reality is that dependence takes on many forms, and it is only possible to estimate the extent to which the different forms are present in the current system. In terms of dependence, we can identify at least three kinds of dependence: financial/technical, social, and cognitive. Financial and technical dependence is best described in Stephan's (2012) work on the economics of science, where a scientist's ability to negotiate for the financial and technical resources needed to conduct research affect his or her ability to produce research. In exchange for bringing together those resources, the scientist (PI) is able to stake some claim on all the intellectual output of that lab, as Stephan puts it "My lab, my article" (Stephan, 2012, p. 74). There is also social dependence, where we expect that a subset of scientists have the ability and desire to coordinate research projects (Bozeman & Corley, 2004). A PI of a lab may coordinate intragroup projects, but may depend on other PIs to coordinate intergroup projects, or run a larger lab and

rely on other scientists within his or her lab to coordinate intragroup projects. Finally, there is cognitive dependence, where scientists depend on other scientists to either provide basic mental labor (the worker bees), or a very specialized set of skills that are required for projects (Bozeman & Corley, 2004; Melin, 2000).

Although it is not possible to determine which type of dependency is in play for any given relationship, or between any two roles, we can draw potential hypotheses for future research based on the data and prior research. The inability to test causal relationships is a known weakness of exploratory studies; in exchange for that weakness, we get the opportunity to generate deductively testable hypotheses for the future. The remainder of the discussion in this section will elaborate and synthesize the analysis, tie the analysis to prior literature, and hypothesize potential causal relationships or future research questions that would clarify or further test the interpretation provided.

The remainder of this Chapter is organized around the roles described in Chapter 4: Methodology. The node role framework was very useful for teasing out dependencies, and may be even more useful for identifying the types of dependencies in future research. More specifically, the node role framework was useful for summarizing historical interactions and predicting near-term collaborative patterns. In terms of the flow of the discussion, we will start at the top, with Role 7, in acknowledgment of what Stephan (2012) refers to as the pyramid-like structure of modern scientific fields.

ROLE	1994	1997	2000	2003	2006	2009
1	**2.9 / 2**	**3 / 2**	**3.6 / 2**	**3.6 / 2**	**4.3 / 3**	**5.1 / 3**
2	**4.8 / 4**	**5.1 / 4**	**5.9 / 5**	**5.9 / 5**	**6.7 / 6**	**7.6 / 7**
3	**7.4 / 7**	**8.2 / 8**	**9.3 / 9**	**9.3 / 9**	**10.4 / 10**	**11.5 / 11**
4	**9.5 / 10**	**10.7 / 11**	**12.1 / 12**	**12.1 / 12**	**13.7 / 14**	**15.2 / 15**
5	**7.2 / 7**	**7.4 / 8**	**9.1 / 8**	**9.1 / 8**	**9.7 / 8**	**9.6 / 9**
6	**9.5 / 10**	**10.1 / 10**	**11.2 / 11**	**11.2 / 11**	**12.3 / 12**	**13.5 / 13**
7	**11.2 / 12**	**12.9 / 13**	**14.5 / 15**	**14.5 / 15**	**16 / 16**	**18 / 18**

Table 6-1: Mean/median years active in the network, by role and year

Before discussing the individual roles, we will bring in one additional piece of data—the mean/median years that scientists were active in the network, by role (Table 6-1). The data in the table will be discussed in greater detail throughout the remainder of the chapter, but what can be seen at first glance is that the mean and median number of years' experience for scientists in each of the roles increases from 1994–2009, suggesting that it takes longer for scientists to move up through the roles. It is important to note that the analysis is for scientists who were active from the year listed in the column header through the next three years, so the data do not include scientists who are older and inactive, which would skew the results heavily toward the higher end. Having said that, some of the observed increase may be due to the fact that by 2009 there were more scientists who had been in the network longer. It would be useful in a follow-up study to trace the scientists' transitions between roles, as it appears that scientists progress through the non-hubs, and somewhere between 7–10 years, transition to either Role 4 or to Role 6.

One final piece of evidence to keep in mind here is that the proportion of scientists in the hub roles did not change significantly over the years (Figure 5-6), which indicates that scientists do not passively accumulate the types of connections necessary to be considered a hub over time. Instead, achieving that hub-like status involves some effort, and is related to the efforts of others within the community.

6.2 Role analysis

The formal definition of a Role 7 in a collaboration network is someone whose intragroup ties are 2.5 z-scores greater than the mean z-score of intragroup ties for other group members and whose ties distribution extends to scientists in many other groups. In the node role framework, they are referred to as global hubs. Within the system of science, in this particular field, they were likely to be lab managers running very large labs that either have active collaborations with many other labs or take in visiting researchers from many other labs. Within the field studied in this dissertation, they are more common than any other hub role (Figure 5-6), which differs from past findings (Velden et al., 2010). This indicates that larger, intergroup collaborations are more common in this field than in other fields; a finding that is supported by the team sciences of the field, which have increased steadily over the years (Costa et al., 2015).

The first piece of evidence to suggest this interpretation is the extent to which Role 7 scientists collaborate within and between research groups (see Table 5-12), where scientists in Role 7 had the second lowest high-end estimates of in-group collaboration. Scientists who've developed the broad networks that define a Role 7 had more experience, and were likely to have their names affixed to many papers, suggesting that they are responsible for coordinating the resources and personnel necessary to conduct the research, but probably did very little actual hands-on work (Beaver, 2001).

Scientists in every other role became more dependent on scientists in Role 7 (Figure 5-8), suggesting that scientists at the top of the pyramid were able to stake an intellectual claim on a larger portion of the research output. One possible reason for this observation is that the equipment needed to conduct sequencing is expensive, and the high throughput instrumentation

is concentrated in a handful of labs worldwide (Stephan, 2012). The collaborative interactions of scientists in Role 7 were star-shaped in the sense that they collaborated with, or were more likely to provide resources to distinct groups of researchers who had little interaction between them, as evidenced by the lower clustering coefficient around Role 7s (Figure 5-12). Scientists in Role 7 were more likely over time to become dependent on at least a core group of collaborators, suggesting that stable teams are preferable as teams get larger. That is to say, as teams get larger, it is more efficient for researchers to assemble teams in chunks, instead of piecewise.

In summary, we can argue that Role 7 scientists become more depended upon over time, and to some extent, become more dependent on others to generate research output (Figure 5-9). Scientists in this role likely benefit from cumulative advantage, with their publication history and accumulated status drawing offers from potential collaborators and junior researchers (H. Jeong et al., 2003; Merton, 1968).

In contrast to scientists who can be described as Role 7s, scientists in Role 6 are those who have established themselves in the research community, but lack, by definition, the breadth of connections to be considered global hubs. Based on the mean years' experience of scientists in this role, we can estimate that they would be equivalent to an associate professor (Table 6-1), with about 12 years' worth of publishing history behind them.

Before going into the difference in dependence patterns between Role 7s and Role 6s, we can see some similarities—mostly around the extent to which the less experienced scientists in Roles 1 and 2 depend on them (Figure 5-8). In comparison to scientists in Role 7, those in Role 6 are much more reliant on their core set of collaborators (Figure 5-7), are more likely to concentrate their collaborative interactions within their own research group (Table 5-12), and

less likely to be depended on by scientists in Role 3, who have roughly the same number of years' experience as those in Role 6 (Table 6-1).

The data suggest that scientists in Role 6 are those who've established functional and stable labs but lack the resources to strongly influence the collaborative interactions of the groups around them. It is clear, based on the definition of a Role 6, that they do collaborate with other research groups. However, the extent to which they are depended on by the nomadic researchers, characterized by the labels Role 3 and Role 4 (see below), indicates that they do not wield sufficient resources to account for large portions of those researchers' time. One thing that would be useful to understand is whether scientists in Role 6 are learning how to become hubs who coordinate large-scale activity, or if they are scientists who prefer to run smaller labs that fill certain niches, and only occasionally coordinate with other labs (Chang & Huang, 2013).

Scientists in Roles 6 and 7 have strong ties within and between groups, while a scientist in Role 5, by definition, has strong ties within his or her own research group, but is weakly tied to researchers in other groups. Based on the mean years' experience of researchers in this group, they are roughly equivalent to either an assistant professor or experienced postdoc (Table 6-1). However, they are not as dependent on scientists in Roles 6 and 7 as other hubs, which would suggest that Role 5s are those who are just establishing their labs and working on collaborating with the other, less experienced members of their lab (Figure 5-8). Having said that, scientists in Role 5 are much more likely to publish a larger percentage of their papers between groups, and to publish a higher percentage of their papers with at least one group member (Table 5-12), than scientists in any other role.

They, too, have lower local clustering coefficients, suggesting that their coauthors are less likely to coauthor with one another. The data suggest that scientists in Role 5 manage several independent projects, keeping the teams separate from one another. Keeping in mind that dependence was calculated in three-year intervals, it is possible that scientists in Role 5 work on a series of publications with non-overlapping lab staff. In terms of the theoretical framework, scientists who are strongly connected to others within their research group and weakly connected to other groups (i.e., Role 5s), are heavily dependent on their own group going forward, but are also much more likely to collaborate between groups. Their tendency to work between groups would result in a change in role over time, as hubs who develop stronger intergroup ties will be, by definition, Role 6s or 7s.

Up to this point, the discussion has focused on the hubs in the mesoscopic structure of the network. For non-hubs, scientists in Role 4 appear to be the most senior, having more experience on average than scientists in any other role other than Role 7s and Role 6s, who they were roughly equal to (Table 6-1). Scientists in this role were also more likely to collaborate outside of their own group and had the lowest in-group collaboration rates (Table 5-12). Additionally, their dependence scores were evenly distributed (Figure 5-7). What the data tells us is that scientists in this role were not assigned to the role based on the community detection algorithm's inability to assign them the proper home (Lancichinetti & Fortunato, 2009), or detect their movement from one group to another. Instead, compared to others, scientists in Role 4 engage in a consistent pattern of collaboration that makes them the least dependent on their own research group.

Additionally, scientists in Role 4 are not heavily dependent on scientists in any other role (Figure 5-8). In terms of the theoretical framework, we can argue that scientists in Role 4 are the

most distinct of scientists because they exhibit the least amount of dependence. One possible reason is that scientists in Role 4 possess a specialized skill set that allows them to work as independent agents between several groups. That is to say, scientists in Role 4 do not have to maintain their own research labs, students, or equipment. Instead, based on their levels of dependence on other scientists (i.e., midrange levels of both being depended on and depending on others in Roles 1-7), we can say that they are more likely to collaborate with other groups with established personnel. Scientists in this role did become more dependent over time, suggesting that they were more likely to publish with fewer teams over the three-year period over which dependence was calculated. One possible reason for the collaboration with fewer teams is that faster sequencing machines enabled higher productivity (Stephan, 2012), which gave these specialists more opportunities to publish with scientists in their home group.

Scientists in Role 3 had about as much experience as scientists in Role 6 but exhibited different patterns of dependence. From 1994–2009, scientists in this role became highly dependent on their core group of coauthors, from publishing with the same coauthor no more than 50% of the time in 1994, to publishing with their core network 100% of the time in 2009 (Figure 5-7). They became more dependent on scientists in every other role but were not heavily depended on by scientists in any role except Role 5. They, too, were more likely to be dependent on their research group over time but were less likely to focus exclusively on in-group collaborations (Table 5-12).

All of the analysis discussed in this section would greatly benefit from a trajectory study, but such a study would probably shed the most light on scientists in Role 3. Based on the years' experience of scientists in this role, we can hypothesize one of two things—scientists in this role are either transitioning between groups due to changes in career stage (e.g., postdoc to assistant

professor), or are a special set of researchers who are more likely to frequently coauthor on teams that span multiple groups.

Scientists in Roles 1 and 2 were highly dependent upon both their groups and other scientists in their immediate network. That is to say, they either wrote only one paper, and thus were entirely dependent on their network (by operational definition), or collaborated with the same set of coauthors on multiple papers. Scientists in Roles 1 and 2 had the least amount of experience. The distribution of their dependencies did not change over the years (Figure 5-7), but they were more likely to be dependent on their group over time. Some of the scientists in Roles 1 and 2 were, as Stephan (2012) labels them, "worker bees"—the undergraduate and graduate students who provide the bulk of the labor in the lab.

One important point to note is that the data suggest intergroup collaboration was much more common as time passed. Scientists in Roles 1 and 2 were more likely to collaborate between groups (Table 5-12), and the observation that more scientists were classified as Role 2 than Role 1, by definition, suggest that these scientists participated in intergroup collaborations. However, it is worth noting that scientists in Roles 1 and 2 were those who were active in the past and the near future (based on the operationalization of dependence), and therefore, had enough experience to occasionally (~ 7% of the time) collaborate with others independently of the group.

6.3 Interpretation

Two major research questions guided this research:

1) How is the increasing prominence of the research group and the associated team-based research impacting scientists' dependence on one another and the research group?
2) What is the relationship between a scientist's distribution of relationships within the group structure and their dependence on other scientists, and how has the relationship between position and dependence changed over time?

With respect to the first question, the data indicates that scientists were more dependent on their research group in the time period analyzed, and that dependency varied by position within the group structure (Table 5-12). Addressing the second question, the role framework revealed clear differences in dependence between most roles, and for scientists in every role, that dependence moved toward interdependence over time.

There are several implications of these findings from the Complex Adaptive Systems perspective. We find support for the theoretical position that interactions among agents in complex systems give rise to stable patterns of interaction that later serve as building blocks for more complex emergent structures (Holland, 1992; Simon, 1991). More specifically, what we are seeing is the emergence of the functional research group as a stable building block for more complex patterns of team-based collaboration. By reducing the cognitive effort spent on organizing the labor that requires the least amount of expertise to conduct, scientists are able to focus on the aspect of team assembly that matters most to the successful outcome of a project—identifying collaborators whose skills and resources complement their skills.

Over time, the research group has a stronger influence on the coordination activities of scientists. Functionally, it becomes more useful for scientists who are focused on larger goals to

leverage the stability of groups to plan and execute projects. Basically, they can use coordination mechanisms and incentives to guide the actions of many research groups toward a single goal. One prominent example of this type of organization is the Human Genome Project, which involved 20 international consortia and cost almost 3 billion USD. What can be argued, and what the results of this dissertation supports, is that stable groups allow scientists to create more complex organizational arrangements that can tackle more complex problems. Another example of a large scale project enabled through the coordinated efforts of multiple research groups is the project that searches for gravitational waves, which involves over 90 institutions across the globe. At Syracuse University, the LIGO team working on the gravitational wave project had three faculty working under the direction of a senior researcher, plus dozens of students and other research staff.

The pyramid scheme described in the LIGO project is similar to what we see across the bioinformatics community—large groups led by established scientists, who in turn coordinate the actions of other scientists below them toward a shared goal. The emergence of stable patterns of collaboration up and down the chain make it possible for larger coordination efforts like this. The node role framework (Guimerà et al., 2007a) used in this dissertation to classify scientists into roles based on their connections within and between groups served to identify scientists who play different parts in keeping the larger organizational system functioning, which is why we saw variable dependence between scientists. One of the weaknesses of this study is that it did not tease out different projects to see if coordination around larger projects (e.g., the Human Genome Project) created different patterns of dependencies than collaboration around other efforts. We see traces of differential sizes in collaborative efforts in the module outputs of the community detection algorithm in (Table 5-2), which produced skewed distributions of module sizes, where

the mean size ranged between 10–14 scientists over the years, and the maximum size was over 100 active scientists (after dropping transients). We could test this interpretation in a subsequent study with a highly curated dataset, where the efforts of individual scientists could be tied to specific projects.

The emergence of the research group as a stable pattern has implications for the interactions of agents. More specifically, the interactions that give rise to the structure of the complex network are strongly influenced by the emergent structure of the group, such that scientists fully embedded within those groups are less likely to interact outside the group independently of group members who already have connections outside the group. It's not as if there is a fundamental rule saying that scientists cannot operate outside of their group independently; instead, the downward causality, described by Sawyer (2005), is where agents in a system see the emergent pattern and react to that pattern.

Functionally, the reaction to the emergent pattern manifests as an alteration of what Wagner and Leydesdorff (2009) describe as the local and global search process. Scientists who have fewer connections in the community are not sought out by other scientists outside their group and limit their searches to scientists within their group. More importantly, what the increasing dependence suggests is that it takes longer for scientists, both in terms of years of publication activity and intergroup connection accrued, to function independently of their groups. The increased dependency makes the stabilization of the basic building blocks of modern collaboration possible. The fact that more experienced scientists are becoming more dependent implies that the system is moving toward larger stable building blocks, which in turn would enable larger scale coordinated efforts, which is what complexity theorists like Weaver (1948) and Simon (1991) suggested would happen.

What the results of this study mean for people who study Complex Networks and Adaptive systems is that the addition of edges between nodes is not done pairwise, and that the ability of a node to add edges outside of its home module is dependent on the actions of other nodes in the module who already have connections outside the module. This would suggest that there is a threshold of connections outside the module beyond which enable the node to add connections independently of other nodes in its module and that the threshold is moving over time as the modules become larger and more integrated.

7 Conclusions and future work

There are strong temporal patterns related to the coauthorship of scientists in the GenBank community. The probability of finding a collaboration that is reactivated after a hiatus decreases by approximately half for every year that passes, while the probability of a coauthorship relationship continuing for an additional year also drops by approximately half for each year that passes. The observed temporal patterns provide support for, but do not prove, that collaborative relationships have momentum due to increased efficiencies and the natural process of research whereby findings generate additional research questions. Furthermore, the momentum appears to decline quickly if the collaborative relationship is not maintained.

From a methodological perspective, the observed temporal patterns can be used to construct network representations of coauthorship. The study of temporal networks is a nascent field (Holme & Saramäki, 2012) with best practices in development. The existing approaches to tracking the changing nature of networks are to assemble either cumulative networks over a given time period or effective networks with a defined window size. Cumulative networks are relatively easy to implement but give equal weight to relationships regardless of their currency. In contrast, effective networks capture the current network at the expense of information loss regarding accumulated status and relationships. Using effective networks for smaller communities is more difficult because the boundaries are more porous; it is easier to mistakenly identify an established scientist as a newcomer when in fact she is an occasional contributor to an area. If clear temporal patterns can be found in other coauthorship networks, the practice of estimating the *decay* of relationships may prove to be a more effective approach to modeling an evolving network.

Looking at collaboration networks through the lens of mesoscopic network analysis continues to be a fruitful line of research. However, methodological approaches to studying the mesoscopic layer still require refinement, especially in temporal networks. Chaining together clustering solutions is difficult to do in abstract networks (Y. Chen, Kawadia, & Urgaonkar, 2013; Kawadia & Sreenivasan, 2012), even more so since there is a requirement to ground the results in social observations. Employing community detection algorithms to study coauthorship networks usually involves a secondary confirmation mechanism (Velden et al., 2010), which is quite difficult to do at larger scales. Nevertheless, the results from the community detection algorithms were consistent with the theory that the existing configuration of relationships serves as both a reference point and constraint for scientists' actions.

Each of the six approaches to clustering scientists into communities produced results that were well outside the range of what was observed in null model networks. The least accurate approach still managed to put over 40% of near term collaborations within-module, and that figure is a lower-end estimate because it assumes that any collaboration with newcomers is not an intra-module collaboration. Of the different approaches to identifying the mesoscopic layer of the network, the bipartite approaches were more accurate in comparison to the unimodal approaches, and they better fit the underlying theory of this research. Bipartite approaches to analyzing collaboration networks have not been popular, most likely due to the fact that their initial research relied on unimodal networks, and there are fewer tools available to study bipartite networks; nevertheless, bipartite networks are likely to be the better choice moving forward. Even from a basic conceptual perspective, unimodal networks involve using language that obscures the reality of the team-based nature of scientific collaboration. Unimodal networks offer one advantage—they are better tools for testing hypotheses related to the interpersonal

dynamics of lab managers and their effects on intergroup collaborations, or studying fields where the dominant form of collaboration is based on individual interaction and not group coordination.

Using the mesoscopic lens to study collaboration networks does reveal some weaknesses in theories that are implicitly or explicitly entangled with complex network analytic models and frameworks. Mechanisms guiding the formation of links and properties of nodes as independent entities within the macroscopic structure that are commonly referred to when discussing complex systems or networks, such as the small-world phenomenon or preferential attachment, or even bridges (Abbasi et al., 2011) and the related concept of structural holes (Burt, 2001), miss an important component of such systems, and that is the prominence of the group structure. As an example, the idea of a bridge node, or a scientist who connects two or more distinct groups, would not provide the level of detail as do the combined concepts of participation coefficient and within-module degree. Is a person floating between groups, or strongly connected in one location and diffusely connected elsewhere? What does the distribution of connections between the modules in the mesoscopic layer mean for agents' actions within the community? It is more effective to use mesoscopic network analysis as a lens to answer these questions, as it provides a richer context detail regarding a scientist's position than unidimensional centrality measures.

The theory laid out in (Wagner & Leydesdorff, 2009) does provide an effective way of abstractly thinking about the reasons for and approaches to forming collaborative relationships, but it needs to be extended to account for the group structure and team-based nature of scientific collaboration. The suggestion is not to strip notions of autonomy and agency away from individuals, but instead to adjust the language to account for the ways in which collaborations are commonly arranged—not through dyadic interactions of all parties, but as interactions coordinated by smaller subsets of actors, which in turn guide the actions of those within the

coordinators' spheres of influence. The observed patterns regarding increasing dependence within the GenBank community support this argument—the distribution of dependence is highly skewed such that a majority of scientists within the community at any given time are dependent on a minority of actors.

Although the distribution of dependence is skewed in the favor of more established scientists, the nature of scientists' interdependence is more complex. The added benefit of stable relationships noted in (D. H. Lee et al., 2012; Whitfield, 2008) appears to foster mutual dependence. The data would not support an argument that more established scientists treat all junior researchers as interchangeable parts. Instead, the results from this dissertation indicate that scientists benefit from cultivating stable relationships with junior researchers. It is not possible to isolate the reason why in this study, but possible reasons could include: increased comfort and familiarity (which was part of the hypothesis related to the half-life of collaboration) and the difficulty associated with identifying skilled professionals.

Not only were scientists more interdependent in general, but they were also more likely to engage in within-group collaboration over the years. It could be said, within the theoretical framework, that the structure of existing relationships serves as a more powerful constraint to scientists' actions in later years. All roles had statistically significant changes in their within-group research participation over the years except for Role 7. Only global hubs exhibited no significant changes to the way in which they distributed their collaborative efforts across and within groups. Put another way, the most established of scientists continue to coordinate intergroup activity in relatively the same manner, while scientists in all other roles were more likely to put more of their collaborative efforts toward established relationships.

7.1 Limitations

The limitations to this study, in terms of validity and reliability, were discussed in §4.6 but are worth repeating to put the discussion, interpretation, and conclusion into context. The limitations fall into three categories—validity, reliability, and bias (Pannucci & Wilkins, 2010; Punch, 1998). The types of limitations are not discussed independently; instead, they are brought together to show how certain assumptions and approaches to the analysis used in this dissertation result in the limitations.

The measurements used in this dissertation present certain validity, reliability, and bias challenges. For example, there is already an established debate on the validity of coauthorship as a measure of scientific collaboration (Glänzel & Schubert, 2005; Laudel, 2002; Melin & Persson, 1996). Coauthorship only captures a portion of scientific collaboration (Glänzel & Schubert, 2005), and sometimes overstates collaborative interactions (Melin & Persson, 1996). Having said that, this dissertation explicitly focused on the production of formal knowledge outputs, and not scientific collaboration in general.

Whether the measure of dependence used in this dissertation is a valid measure of the construct is debatable as well because it is not a widely tested measure. The operationalization of the concept is based on certain assumptions (i.e., that authorship represents contributions and that authorship claims are systematically applied) that leave it susceptible to certain systematic biases that may or may not be consistent within this field, or in other fields as well. Specifically, it is possible that intellectual claims in the subcommunity of bioinformatics studied in this dissertation adhere to a different pattern than the broader bioinformatics community or other research communities. There is prior literature to draw on (Stephan, 2012; Whitely, 2000)

supporting the argument that the field chosen fits the description of a field that is oriented around the research group and has a high degree of task dependency, but no systematic analysis of the community has been conducted.

In one sense, the research presented in this dissertation is highly reliable, in that another researcher could use the code to obtain the same results (Punch, 1998). However, the interpretation of the results is highly dependent on the researcher (see the section on bias below), which could result in a form of low interrater reliability. There are also issues with reliability in terms of the data set—there may be systematic biases in the data set related to the way certain subsets of the community stake their intellectual claims. There is also the chance that author attribution and/or practices have changed over the years, which would reduce the reliability of the measure longitudinally.

In addition to the systematic bias discussed above, there is also researcher bias to contend with. In particular, the interactive effect between this dissertation being an exploratory study and the researcher looking for evidence of the core concept may have introduced systematic bias into the analysis, particularly at the interpretation level. I tried to investigate the concept from multiple perspectives to search for supporting or contradictory evidence, but there is still a chance that bias might be present. To a certain extent, this type of bias is an inherent limitation of the study type, and would be best addressed by conducting a subsequent, deductive, confirmatory analysis-based study.

Another type of systematic error that may be present stems from errors in data processing and manipulation. This dissertation relied heavily on computational methods of analysis, including numerous chained steps of data transformation (e.g., author–publication lists to author–

publication matrices to condensed matrices over time), which opens up the possibility for error. The analysis also relied heavily on open source analysis packages in R, which were referenced and are available in the code. Extensive work went into identifying and countering coding error, including the use of hand calculations, analysis on subsets of data that were amenable to direct inspection and verification, and the use of simpler algorithms that were known to produce the correct results, but were not scalable to the larger data set. The code is available upon request for verification, and will be made publicly available in the near future after sensitive information related to the storage of the data on servers is expunged.

The operationalization of dependence masks the nature of the relationship that gives rise to dependence. As a result, no empirical evidence can be provided as to why the coauthorship patterns that served as the basis for the operationalization of dependence came about. To that end, the measure of dependence may not be reliable in terms of measuring dependence in other contexts, and its validity cannot be tested any further using the methods in this dissertation.

In terms of changes over time, using the publication dates to determine the chronological ordering of relationships is prone to bias, particularly if one were to try and repeat this study on another community. This is due to the fact that lag between submission for publication and actual publication differ between fields, and may even differ over the years.

The issues surrounding the validity and reliability of this study were expected, either because other researchers have explored the measures before (e.g., coauthorship as a measure of collaboration), or because of the exploratory nature of the study. This research looked at dependence as it relates to publication output, which is not a tested approach. In order to test the

validity and reliability of the findings, the experiment would have to be replicated on other scientific communities.

7.2 Future work

There are at least three areas to focus on in the future. First, the theoretical concept of local and global searches should involve exploring the reputation embedded in the body of literature. Scientists use their publications to stake a claim on research areas; the attention their publications gather contributes to the reputation of the researcher. Although a significant portion of scientists rely on interpersonal relationships and word-of-mouth to identify potential collaborators, many others use the publication record to identify expertise. A question of how the collaboration and publication networks co-evolve has been explored before (Börner et al., 2004; De Domenico, Lancichinetti, Arenas, & Rosvall, 2014), but not extensively. It would be useful to explore how long lists of authors affects the identification of expertise in the publication record, and what the implications are for individual scientists who are trying to establish their careers.

This study should also be expanded to include other fields in order to improve the generalizability of the findings. It is difficult to properly model the relationship between team size, interactions between research groups, and individual scientists' dependence without being able to compare genomics research to other disciplines. What portion of the variance is truly captured by these factors versus other unknown exogenous or endogenous factors? Using large datasets similar to what was used in (Uzzi et al., 2013; Wuchty et al., 2007) would be appropriate for this task. Extending the concept of dependence to the notions of team dynamics (e.g., mixing of repeat and new collaborations) outlined in those publications would also be useful.

A more inclusive, better-curated dataset would also be of tremendous use. In addition to comparing the results across fields, it would be beneficial to compare team structure and dynamics within the same community in different nations. Whitely (2001) suggests that different funding regimes and types of science both have an impact on the social organization of the sciences. From a policy perspective, it would be interesting to see how different approaches to funding research affects the interplay between individual action and social organization. It is also important to note that the analysis contained in this dissertation did not include collaborations on datasets or patents. Bioinformatics is a highly commercialized field—there are over 25 million patents in the GenBank database alone. Additionally, there are publications related to bioinformatics that are not submitted to GenBank because they do not directly address the sequencing of the genome. Second-order publications that use the sequencing data still include active members of the genomics and genetics community; collaborations that can be identified in those publications would improve the validity of this research. Finally, a dataset that contained information on the career movements and formal affiliations of researchers would be beneficial. Trying to ground the results from the community detection algorithm to formal affiliations, similar to what was done in (Velden et al., 2010), but computationally, would improve the validity of this research as well.

One last area for future research lies at the intersection of the results of specific funding and system dynamics. How does the awarding of grants change the nature of scientists' dependence? Or, put another way, how would the awarding of a grant change the structure of the community? The research is not at a point where prescriptive guidance can be provided through network analysis (Whitfield, 2008), but understanding how different decisions can affect the capacity of a research area over time could help programs allocate resources more effectively.

Research through simulations would probably be an effective method to explore these relationships. However, before that can happen more work needs to be done to connect the concepts of team science, dependence, and resilience within a complex adaptive systems framework.

References

Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, *6*(3), 403–412. http://doi.org/10.1016/j.joi.2012.01.002

Abbasi, A., Hossain, L., Uddin, S., & Rasmussen, K. J. R. (2011). Evolutionary dynamics of scientific collaboration networks: multi-levels and cross-time analysis. *Scientometrics*, *89*(2), 687–710. http://doi.org/10.1007/s11192-011-0463-1

Abramo, G., D'Angelo, C. A., & Solazzi, M. (2011). The relationship between scientists' research performance and the degree of internationalization of their research. *Scientometrics*, *86*(3), 629–643. http://doi.org/10.1007/s11192-010-0284-7

Aggarwal, C., & Subbian, K. (2014). Evolutionary Network Analysis: A Survey. *ACM Comput. Surv.*, *47*(1), 10:1–10:36. http://doi.org/10.1145/2601412

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. http://doi.org/10.1103/RevModPhys.74.47

Ardanuy, J. (2011). Scientific collaboration in Library and Information Science viewed through the Web of Knowledge: the Spanish case. *Scientometrics*, *90*(3), 877–890. http://doi.org/10.1007/s11192-011-0552-1

Arenas, A., Danon, L., Díaz-Guilera, A., Gleiser, P. M., & Guimerá, R. (2004). Community analysis in social networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, *38*(2), 373–380. http://doi.org/10.1140/epjb/e2004-00130-1

Arthur, W. B. (1999). Complexity and the Economy. *Science*, *284*(5411), 107–109. http://doi.org/10.1126/science.284.5411.107

Axelrod, R. M. (1997). *The Complexity of Cooperation: Agent-based Models of Competition and Collaboration*. Princeton University Press.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512. http://doi.org/10.1126/science.286.5439.509

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and Its Applications*, *311*(3), 590–614.

Bassett, D. S., Porter, M. A., Wymbs, N. F., Grafton, S. T., Carlson, J. M., & Mucha, P. J. (2013). Robust Detection of Dynamic Community Structure in Networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *23*(1), 013142. http://doi.org/10.1063/1.4790830

Beaver, D. (2001). Reflections on Scientific Collaboration (and its study): Past, Present, and Future. *Scientometrics*, *52*(3), 365–377. http://doi.org/10.1023/A:1014254214337

Beaver, D., & Rosen, R. (1978). Studies in scientific collaboration: Part I. The professional origins of scientific co-authorship. *Scientometrics*, *1*(1), 65–84. http://doi.org/10.1007/BF02016840

Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration: Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830. *Scientometrics*, *1*(2), 133–149. http://doi.org/10.1007/BF02016966

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., … Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, *59*, 1–26. http://doi.org/10.1111/j.1467-9922.2009.00533.x

Bonaccorsi, A. (2008). Search Regimes and the Industrial Dynamics of Science. *Minerva*, *46*(3), 285–315. http://doi.org/10.1007/s11024-008-9101-3

Borgatti, S. P., Jones, C., & Everett, M. G. (1998). Network measures of social capital. *Connections*, *21*(2), 27–36.

Börner, K., Maru, J., & Goldstone, R. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101 Suppl*, 5266–73. http://doi.org/10.1073/pnas.0307625100

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy*, *33*(4), 599–616.

Bozeman, B., Dietz, J. S., & Gaughan, M. (2001). Scientific and technical human capital: an alternative model for research evaluation. *International Journal of Technology Management*, *22*(7-8), 716–740.

Braun, T., Glänzel, W., & Schubert, A. (2001). Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics*, *51*(3), 499–510.

Brunson, J. C., Fassino, S., McInnes, A., Narayan, M., Richardson, B., Franck, C., … Laubenbacher, R. (2013). Evolutionary events in a mathematical sciences research collaboration network. *Scientometrics*, *99*(3), 973–998. http://doi.org/10.1007/s11192-013-1209-z

Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., & Havlin, S. (2010). Catastrophic cascade of failures in interdependent networks. *Nature*, *464*(7291), 1025–1028. http://doi.org/10.1038/nature08932

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198. http://doi.org/10.1038/nrn2575

Bunt, G. G. van de, & Groenewegen, P. (2007). An Actor-Oriented Dynamic Network Approach The Case of Interorganizational Network Evolution. *Organizational Research Methods*, *10*(3), 463–482. http://doi.org/10.1177/1094428107300203

Burton, R. E., & Kebler, R. W. (1960). The “half-life” of some scientific and technical literatures. *American Documentation*, *11*(1), 18–22. http://doi.org/10.1002/asi.5090110105

Burt, R. S. (2001). The social capital of structural holes. In Guillen, Mauro F., R. Collins, P. England, & M. Meyer (Eds.), *The new economic sociology: Developments in an emerging field* (pp. 201–247). New York: Russel Sage.

Calero, C., Buter, R., Valdés, C. C., & Noyons, E. (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. *Scientometrics*, *66*(2), 365–376. http://doi.org/10.1007/s11192-006-0026-z

Chang, H.-W., & Huang, M.-H. (2013). Prominent institutions in international collaboration network in astronomy and astrophysics. *Scientometrics*, *97*(2), 443–460. http://doi.org/10.1007/s11192-013-0976-x

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. http://doi.org/10.1002/asi.20317

Chen, Y., Kawadia, V., & Urgaonkar, R. (2013). Detecting Overlapping Temporal Community Structure in Time-Evolving Networks. *arXiv:1303.7226 [physics, Stat]*. Retrieved from http://arxiv.org/abs/1303.7226

Cilliers, P., & Spurrett, D. (1999). Complexity and post-modernism: Understanding complex systems. *South African Journal of Philosophy*, *18*(2), 258–274.

Cole, S., & Cole, J. R. (1968). Visibility and the Structural Bases of Awareness of Scientific Research. *American Sociological Review*, *33*(3), 397. http://doi.org/10.2307/2091914

Colizza, V., Barrat, A., Barthélemy, M., & Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(7), 2015–2020. http://doi.org/10.1073/pnas.0510525103

Costa, M. R., & Qin, J. (2012). Analysis of networks in cyberinfrastructure-enabled research communities: A pilot study. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–4. http://doi.org/10.1002/meet.14504901244

Costa, M. R., Qin, J., & Bratt, S. (In press). Emergence of collaboration networks around large scale data repositories: A study of the genomics community using GenBank. *Scientometrics, forthcoming*.

Costa, M. R., Qin, J., & Wang, J. (2014). Research Networks in Data Repositories. Presented at the Digital Libraries 2014: ACM/IEEE Joint Conference on Digital Libraries, London: ACM.

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology*, *54*(9), 855–871. http://doi.org/10.1002/asi.10278

Dahlander, L., & McFarland, D. A. (2013). Ties that last: Tie formation and persistence in research collaborations over time. *Administrative Science Quarterly*, *58*(1), 69–110. http://doi.org/10.1177/0001839212474272

Danon, L., Duch, J., Diaz-Guilera, A., & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, *2005*(09), P09008–P09008. http://doi.org/10.1088/1742-5468/2005/09/P09008

De Domenico, M., Lancichinetti, A., Arenas, A., & Rosvall, M. (2014). Identifying modular flows on multilayer networks reveals highly overlapping organization in social systems. *arXiv:1408.2925 [physics]*. Retrieved from http://arxiv.org/abs/1408.2925

Defazio, D., Lockett, A., & Wright, M. (2009). Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. *Research Policy, 38*(2), 293–305. http://doi.org/10.1016/j.respol.2008.11.008

de Solla Price, D. J. (1963). *Little science, big science*. New York: Columbia University Press.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science, 149*(3683), 510.

de Solla Price, D. J., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist, 21*(11), 1011–1018. http://doi.org/10.1037/h0024051

de Solla Price, D. J., & Gürsey, S. (1975). Studies in Scientometrics I Transience and Continuance in Scientific Authorship. *Ciência da Informação, 4*(1). Retrieved from http://revista.ibict.br/cienciadainformacao/index.php/ciinf/article/view/1611

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics, 5*(1), 187–203. http://doi.org/10.1016/j.joi.2010.10.008

Easley, D., & Kleinberg, J. (2010). Graphs. In *Networks, Crowds, and Markets: Reasoning about a Highly Connected World* (Vol. 26, pp. 23–46). Retrieved from http://djjr-courses.wdfiles.com/local--files/soc180%3Adigital-reference-library/Ch2_Kleinberg2010-networks-book.pdf

Estrada, E., & Rodriguez-Velazquez, J. A. (2006). Complex Networks as Hypergraphs. *Physica A: Statistical Mechanics and Its Applications*, *364*, 581–594. http://doi.org/10.1016/j.physa.2005.12.002

Fan, Y., Li, M., Chen, J., Gao, L., Di, Z., & Wu, J. (2004). Network of Econophysicists: A weighted network to investigate the development of Econophysics. *International Journal of Modern Physics B*, *18*(17n19), 2505–2511. http://doi.org/10.1142/S0217979204025579

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurman, P. W., Barrett, J. C., & Birkinshaw, J. (2006). Scientific Collaboration Results in Higher Citation Rates of Published Articles. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *26*(6), 759–767. http://doi.org/10.1592/phco.26.6.759

Frame, J. D., & Carpenter, M. P. (1979). International Research Collaboration. *Social Studies of Science*, *9*(4), 481–497. http://doi.org/10.1177/030631277900900405

Freeman, R. B., & Huang, W. (2014). *Collaborating With People Like Me: Ethnic co-authorship within the US* (Working Paper No. 19905). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19905

Gauvin, L., Panisson, A., & Cattuto, C. (2014). Detecting the Community Structure and Activity Patterns of Temporal Networks: A Non-Negative Tensor Factorization Approach. *PLoS ONE*, *9*(1), e86028. http://doi.org/10.1371/journal.pone.0086028

Gell-Mann, M. (2002). What Is Complexity? In P. A. Q. Curzio & P. M. Fortis (Eds.), *Complexity and Industrial Clusters* (pp. 13–24). Physica-Verlag HD. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-50007-7_2

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 7821–7826. http://doi.org/10.1073/pnas.122653799

Glänzel, W., & Lange, C. de. (2002). A distributional approach to multinationality measures of international scientific collaboration. *Scientometrics*, *54*(1), 75–89. http://doi.org/10.1023/A:1015684505035

Glänzel, W., Leta, J., & Thijs, B. (2006). Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics*, *67*(1), 67–86. http://doi.org/10.1007/s11192-006-0055-7

Glänzel, W., & Schubert, A. (2001). Double effort = Double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, *50*(2), 199–214. http://doi.org/10.1023/A:1010561321723

Glänzel, W., & Schubert, A. (2005). Analysing scientific networks through co-authorship. In *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S & T Systems* (pp. 257–276). Dordrecht: Kluwer.

Glänzel, W., & Winterhager, M. (1992). International collaboration of three east European countries with Germany in the sciences, 1980–1989. *Scientometrics*, *25*(2), 219–227. http://doi.org/10.1007/BF02028083

Goh, K.-I., & Barabási, A.-L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters), 81*(4), 48002. http://doi.org/10.1209/0295-5075/81/48002

Gonzalez-Brambila, C. N., Veloso, F. M., & Krackhardt, D. (2008). Social capital and the Creation of Knowledge [Industry Studies Working Paper]. Retrieved June 6, 2014, from http://isapapers.pitt.edu/101/

Gorraiz, J., Reimann, R., & Gumpenberger, C. (2011). Key factors and considerations in the assessment of international collaboration: a case study for Austria and six countries. *Scientometrics*, *91*(2), 417–433. http://doi.org/10.1007/s11192-011-0579-3

Gregorio, D., & Shane, S. (2003). Why do some universities generate more start-ups than others? *Research Policy*, *32*(2), 209–227. http://doi.org/10.1016/S0048-7333(02)00097-5

Guillaume, J.-L., & Latapy, M. (2004). Bipartite structure of all complex networks. *Information Processing Letters*, *90*(5), 215–221. http://doi.org/10.1016/j.ipl.2004.03.007

Guimerà, R., & Amaral, L. A. N. (2005). Cartography of complex networks: modules and universal roles. *Journal of Statistical Mechanics (Online), 2005*(P02001), P02001–1–P02001–13. http://doi.org/10.1088/1742-5468/2005/02/P02001

Guimerà, R., & Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, *433*(7028), 895–900. http://doi.org/10.1038/nature03288

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2007a). Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics*, *3*(1), 63–69. http://doi.org/10.1038/nphys489

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2007b). Module identification in bipartite and directed networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *76*(3 Pt 2), 036102.

Guimerà, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, *308*(5722), 697–702. http://doi.org/10.1126/science.1106340

Haan, J. D. (1997). Authorship patterns in Dutch sociology. *Scientometrics*, *39*(2), 197–208. http://doi.org/10.1007/BF02457448

Hackett, E. J. (2005). Essential Tensions Identity, Control, and Risk in Research. *Social Studies of Science*, *35*(5), 787–826. http://doi.org/10.1177/0306312705056045

Hara, N., Solomon, P., Kim, S.-L., & Sonnenwald, D. H. (2003). An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information Science and Technology*, *54*(10), 952–965. http://doi.org/10.1002/asi.10291

Heylighen, F. (1989). Self-organization, emergence and the architecture of complexity. In *Proceedings of the 1st European conference on System Science* (Vol. 18, pp. 23–32). Paris: AFCET. Retrieved from http://pespmc1.vub.ac.be/Papers/Self-OrgArchComplexity.pdf

Hill, V. A. (2008). Collaboration in an Academic Setting: Does the Network Structure Matter?

Holland, J. H. (1992). Complex Adaptive Systems. *Daedalus*, *121*(1), 17–30.

Holland, J. H. (2006). Studying Complex Adaptive Systems. *Journal of Systems Science and Complexity*, *19*(1), 1–8. http://doi.org/10.1007/s11424-006-0001-z

Holme, P., & Saramäki, J. (2012). Temporal networks. *Physics Reports*, *519*(3), 97–125. http://doi.org/10.1016/j.physrep.2012.03.001

Jansen, D., Görtz, R. von, & Heidler, R. (2010). Knowledge production and the structure of collaboration networks in two scientific fields. *Scientometrics*, *83*(1), 219–241. http://doi.org/10.1007/s11192-009-0022-1

Jeong, H., Néda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, *61*(4), 567. http://doi.org/10.1209/epl/i2003-00166-9

Jeong, S., Choi, J. Y., & Kim, J. (2011). The determinants of research collaboration modes: exploring the effects of research and researcher characteristics on co-authorship. *Scientometrics*, *89*(3), 967–983. http://doi.org/10.1007/s11192-011-0474-y

Johnson, N. F. (2007). *Two's company, three is complexity: A simple guide to the science of all sciences*. Oneworld Pubns Ltd.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-University Research Teams: Shifting Impact, Geography, and Stratification in. *Science*, *322*(5905), 1259–1262. http://doi.org/10.1126/science.1158357

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, *31*(1), 31–43.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. http://doi.org/10.1016/S0048-7333(96)00917-1

Kawadia, V., & Sreenivasan, S. (2012). Online detection of temporal communities in evolving networks by estrangement confinement. *Scientific Reports*, *2*. http://doi.org/10.1038/srep00794

King, C. (2012). Multiauthor Papers: Onward and Upward. *Science Watch News*. Retrieved from http://archive.sciencewatch.com/newsletter/2012/201207/multiauthor_papers/

Klemm, K., & Eguíluz, V. M. (2002). Highly clustered scale-free networks. *Physical Review E*, *65*(3), 036123. http://doi.org/10.1103/PhysRevE.65.036123

Kraut, R., & Egido, C. (1988). Patterns of Contact and Communication Collaboration in Scientific. In I. Greif (Ed.), *CSCW '88 Proceedings of the 1988 ACM Conference on Computer-supported Cooperative Work* (pp. 1–12). New York, New York, USA: ACM.

Ladyman, J., Lambert, J., & Wiesner, K. (2012a). What is a complex system? *European Journal for Philosophy of Science*, *3*(1), 33–67. http://doi.org/10.1007/s13194-012-0056-8

Ladyman, J., Lambert, J., & Wiesner, K. (2012b). What is a complex system? *European Journal for Philosophy of Science*, *3*(1), 33–67. http://doi.org/10.1007/s13194-012-0056-8

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, *80*(5), 056117. http://doi.org/10.1103/PhysRevE.80.056117

Larivière, V., Gingras, Y., & Archambault, É. (2013). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, *68*(3), 519–533. http://doi.org/10.1007/s11192-006-0127-8

Laudel, G. (2001). Collaboration, creativity and rewards: why and how scientists collaborate. *International Journal of Technology Management*, *22*(7), 762–781.

Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation*, *11*(1), 3–15. http://doi.org/10.3152/147154402781776961

Lee, D. H., Seo, I. W., Choe, H. C., & Kim, H. D. (2012). Collaboration network patterns and research performance: the case of Korean public research institutions. *Scientometrics*, *91*(3), 925–942. http://doi.org/10.1007/s11192-011-0602-8

Lee, S., & Bozeman, B. (2005). The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science*, *35*(5), 673–702. http://doi.org/10.1177/0306312705052359

Leimu, R., & Koricheva, J. (2005). Does Scientific Collaboration Increase the Impact of Ecological Articles? *BioScience*, *55*(5), 438–443. http://doi.org/10.1641/0006-3568(2005)055[0438:DSCITI]2.0.CO;2

Leydesdorff, L. (2003). A sociological theory of communication. *The Self-Organization*.

Li, M., Fan, Y., Chen, J., Gao, L., Di, Z., & Wu, J. (2005). Weighted networks of scientific communication: the measurement and topological role of weight. *Physica A: Statistical*

Mechanics and Its Applications, *350*(2–4), 643–656. http://doi.org/10.1016/j.physa.2004.11.039

Lin, N. (1999). Building a network theory of social capital.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, *41*(6), 1462–1480.

Liu, X. F., Xu, X.-K., Small, M., & Tse, C. K. (2011). Attack Resilience of the Evolving Scientific Collaboration Network. *PLoS ONE*, *6*(10), e26271. http://doi.org/10.1371/journal.pone.0026271

Logan, E. L., & Pao, M. L. (1990). Analytic and empirical measures of key authors in schistosomiasis. In *PROCEEDINGS OF THE ASIS ANNUAL MEETING* (Vol. 27, pp. 213–219). INFORMATION TODAY INC 143 OLD MARLTON PIKE, MEDFORD, NJ 08055-8750.

Logan, E. L., & Pao, M. L. (1991). Identification of Key Authors in a Collaborative Network. *Proceedings of the ASIS Annual Meeting*, *28*, 261–66.

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of International Scientific Collaboration. *Science, Technology, & Human Values*, *17*(1), 101–126.

Luukkonen, T., Tjissen, R., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics*, *28*(1), 15–36.

Mali, F., Kronegger, L., Doreian, P., & Ferligoj, A. (2012). Dynamic Scientific Co-Authorship Networks. In A. Scharnhorst, K. Börner, & P. van den Besselaar (Eds.), *Models of Science Dynamics* (pp. 195–232). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-23068-4_6

Melin, G. (2000). Pragmatism and self-organization. *Research Policy*, *29*(1), 31–40.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, *36*(3), 363–377. http://doi.org/10.1007/BF02129600

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. http://doi.org/10.1126/science.159.3810.56

Merton, R. K. (1973). The normative structure of science. In N. W. Storer (Ed.), *The sociology of science* (pp. 267–280). Chicago: University of Chicago Press. Retrieved from http://europepmc.org/abstract/MED/17737466

Merton, R. K. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis*, *79*(4), 606–623. http://doi.org/10.2307/234750

Mitchell, M. (2006). Complex systems: Network thinking. *Artificial Intelligence*, *170*(18), 1194–1212.

Moed, H. F. (2006). *Citation Analysis in Research Evaluation*. Springer.

Mohrman, S. A., Galbraith, J. R., & Monge, P. (2006). Network attributes impacting the generation and flow of knowledge within and from the basic science community. *Innovation Science and Industrial Change: A Research Handbook*, 196–216.

Nahapiet, J., & Ghoshal, S. (1998). Social Capital, Intellectual Capital, and the Organizational Advantage. *Academy of Management Review*, *23*(2), 242–266. http://doi.org/10.5465/AMR.1998.533225

Newman, M. E. J. (2001a). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132. http://doi.org/10.1103/PhysRevE.64.016132

Newman, M. E. J. (2001b). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, *64*(1), 016131. http://doi.org/10.1103/PhysRevE.64.016131

Newman, M. E. J. (2001c). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409. http://doi.org/10.1073/pnas.98.2.404

Newman, M. E. J. (2002). Assortative Mixing in Networks. *Physical Review Letters*, *89*(20), 208701. http://doi.org/10.1103/PhysRevLett.89.208701

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256. http://doi.org/10.1137/S003614450342480

Newman, M. E. J. (2004a). Fast algorithm for detecting community structure in networks. *Physical Review E*, *69*(6), 066133. http://doi.org/10.1103/PhysRevE.69.066133

Newman, M. E. J. (2004b). Who is the best connected scientist? A study of scientific coauthorship networks. *Complex Networks*, 337–370.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. http://doi.org/10.1103/PhysRevE.69.026113

Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(Suppl 1), 2566–2572. http://doi.org/10.1073/pnas.012582999

Niazi, M. A. K. (2011). Towards A Novel Unified Framework for Developing Formal, Network and Validated Agent-Based Simulation Models of Complex Adaptive Systems. Retrieved from http://dspace.stir.ac.uk/handle/1893/3365

Ozel, B. (2012a). Collaboration structure and knowledge diffusion in Turkish management academia. *Scientometrics*, *93*(1), 183–206. http://doi.org/10.1007/s11192-012-0641-9

Ozel, B. (2012b). Individual cognitive structures and collaboration patterns in academia. *Scientometrics*, *91*(2), 539–555. http://doi.org/10.1007/s11192-012-0624-x

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and Avoiding Bias in Research. *Plastic and Reconstructive Surgery*, *126*(2), 619–625. http://doi.org/10.1097/PRS.0b013e3181de24bc

Pao, M. L. (1982). Collaboration in computational musicology. *Journal of the American Society for Information Science*, *33*(1), 38–43. http://doi.org/10.1002/asi.4630330107

Pao, M. L. (1992). Global and local collaborators: A study of scientific collaboration. *Information Processing & Management, 28*(1), 99–109. http://doi.org/10.1016/0306-4573(92)90096-I

Parker, M., & Welch, E. W. (2013). Professional networks, science ability, and gender determinants of three types of leadership in academic science and engineering. *The Leadership Quarterly, 24*(2), 332–348. http://doi.org/10.1016/j.leaqua.2013.01.001

Pečlin, S., Južnič, P., Blagus, R., Sajko, M. Č., & Stare, J. (2012). Effects of international collaboration and status of journal on impact of papers. *Scientometrics*, *93*(3), 937–948. http://doi.org/10.1007/s11192-012-0768-8

Perianes-Rodríguez, A., Olmeda-Gómez, C., & Moya-Anegón, F. (2010). Detecting, identifying and visualizing research groups in co-authorship networks. *Scientometrics, 82*(2), 307–319. http://doi.org/10.1007/s11192-009-0040-z

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421–432. http://doi.org/10.1023/B:SCIE.0000034384.35498.7d

Piette, M. J., & Ross, K. L. (1992). An analysis of the determinants of co-authorship in economics. *The Journal of Economic Education, 23*(3), 277. http://doi.org/10.2307/1183230

Pravdic, N., & Oluić-Vuković, V. (1986). Dual approach to multiple authorship in the study of collaboration/scientific output relationship. *Scientometric, 10*(5-6), 259–280.

Punch, K. F. (1998). *Introduction to social research: Quantitative and qualitative approaches*. Thousand Oaks, CA: Sage.

Qin, J. (2014). Introduction. Retrieved November 12, 2014, from http://metadatalab.syr.edu/

Qin, J., Lancaster, F. W., & Allen, B. (1997). Types and levels of collaboration in interdisciplinary research in the sciences. *Journal of the American Society for Information Science*, *48*(10), 893–916. http://doi.org/10.1002/(SICI)1097-4571(199710)48:10<893::AID-ASI5>3.0.CO;2-X

Ramasco, J. J., Dorogovtsev, S. N., & Pastor-Satorras, R. (2004). Self-organization of collaboration networks. *Physical Review E, 70*(3), 036106. http://doi.org/10.1103/PhysRevE.70.036106

Rigby, J., & Edler, J. (2005). Peering inside research networks: Some observations on the effect of the intensity of collaboration on the variability of research quality. *Research Policy*, *34*(6), 784–794. http://doi.org/10.1016/j.respol.2005.02.004

Rosenberg, N. (1998). Chemical engineering as a general purpose technology. In E. Helpman (Ed.), *General purpose technologies and economic growth* (pp. 167–192). Cambridge: MIT Press.

Rosvall, M. (2014). mapequation.org - code. Retrieved December 2, 2014, from http://www.mapequation.org/code.html

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. http://doi.org/10.1140/epjst/e2010-01179-1

Rosvall, M., & Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, *104*(18), 7327–7331. http://doi.org/10.1073/pnas.0611034104

Rosvall, M., Esquivel, A. V., Lancichinetti, A., West, J. D., & Lambiotte, R. (2014). Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*, *5*. http://doi.org/10.1038/ncomms5630

Sætre, A. S., Wiggins, J., Atkinson, O. T., & Atkinson, B. K. E. (2009). University Spin-Offs as Technology Transfer: A Comparative Study among Norway, the United States, and Sweden. *Comparative Technology Transfer and Society*, *7*(2), 115–145. http://doi.org/10.1353/ctt.0.0036

Sawyer, R. K. (2005). *Social Emergence: Societies as Complex Systems*. Cambridge University Press.

Schott, T. (1998). Ties between center and periphery in the scientific world-sytem: Accumulation of rewards, dominance and self-reliance in the center. *Journal of World Systems Research*, *4*(2), 112–144.

Schutt, R. K. (2006). *Investigating the social world: The process and practice of research* (Fifth). Thousand Oaks: Sage.

Seglen, P. O., & Aksnes, D. W. (2000). Scientific Productivity and Group Size: A Bibliometric Analysis of Norwegian Microbiological Research. *Scientometrics*, *49*(1), 125–143. http://doi.org/10.1023/A:1005665309719

Simon, H. A. (1991). The Architecture of Complexity. In *Facets of Systems Science* (pp. 457–476). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4899-0718-9_31

Smith, G. D., & Ebrahim, S. (2002). Data dredging, bias, or confounding. *BMJ*, *325*(7378), 1437–1438. http://doi.org/10.1136/bmj.325.7378.1437

Stephan, P. E. (2012). *How economics shapes science*. Cambridge, MA: Harvard University Press.

Szalay, A., & Blakeley, J. (2009). Gray's laws: Database-centric computing in science. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm: Data-intensive scientific discovery* (1.1 ed., pp. 5–12). United States: Microsoft Corporation. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Taramasco, C., Cointet, J.-P., & Roth, C. (2010). Academic team formation as evolving hypergraphs. *Scientometrics*, *85*(3), 721–740. http://doi.org/10.1007/s11192-010-0226-4

Tomassini, M., & Luthi, L. (2007). Empirical analysis of the evolution of a scientific collaboration network. *Physica A: Statistical Mechanics and Its Applications*, *385*(2), 750–764. http://doi.org/10.1016/j.physa.2007.07.028

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, *342*(6157), 468–472. http://doi.org/10.1126/science.1240474

Vafeas, N. (2010). Determinants of single authorship. *EuroMed Journal of Business*, *5*(3), 332–344. http://doi.org/10.1108/14502191011080845

Van Raan, A. F. J. (1998). The influence of international collaboration on the impact of research results: Some simple mathematical considerations concerning the role of self-citations. *Scientometrics*, *42*(3), 423–428. http://doi.org/10.1007/BF02458380

Van Raan, A. F. J. van. (2000). On Growth, Ageing, and Fractal Differentiation of Science. *Scientometrics*, *47*(2), 347–362. http://doi.org/10.1023/A:1005647328460

van Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy*, *40*(3), 463–472. http://doi.org/10.1016/j.respol.2010.11.001

Velden, T., Haque, A., & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks - mesoscopic analysis and interpretation. *Scientometrics*, *85*(1), 1–37.

Wagner, C. S., & Leydesdorff, L. (2005). Mapping the network of global science: comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalisation*, *1*(2), 185–208.

Wagner, C. S., & Leydesdorff, L. (2009). Measuring the Globalization of Knowledge Networks. *arXiv:0911.3646 [physics]*. Retrieved from http://arxiv.org/abs/0911.3646

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442. http://doi.org/10.1038/30918

Weaver, W. (1948). Science and complexity. *American Scientist*, *36*, 536–544.

Whitfield, J. (2008). Collaboration: Group theory. *Nature News*, *455*(7214), 720–723. http://doi.org/10.1038/455720a

Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences*. Oxford University Press.

Wong, P. K., & Singh, A. (2013). Do co-publications with industry lead to higher levels of university technology commercialization activity? *Scientometrics*, *97*(2), 245–265. http://doi.org/10.1007/s11192-013-1029-1

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *316*(5827), 1036–1039. http://doi.org/10.1126/science.1136099

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, *60*(10), 2107–2118. http://doi.org/10.1002/asi.21128

Ziman, J. M. (1994). *Prometheus bound*. Cambridge University Press.

Zuckerman, H. (1967). Nobel laureates in science: Patterns of productivity, collaboration and authorship. *American Sociological Review*, *32*(3), 391–403.

Mark R. Costa
Curriculum Vitae

1710 N Lake Road, Cazenovia, New York 13035 | (315) 214-1712 (C) | mark.r.costa@gmail.com | www.linkedin.com/in/markrcosta/

EDUCATION

PhD, Information Science & Technology. School of Information Studies; Syracuse University (2016).
Dissertation: *The Interdependence of Scientists in the Era of Team Science: An Exploratory Study Using Temporal Analysis of Networks*
Committee: Prof. Jian Qin, Prof. Kevin Crowsten, Dr. Theresa Velden, Dr. Jun Wang

MLS Library and Information Science. School of Informatics; University of Buffalo, SUNY (2003)

B.A., History. University of Buffalo, SUNY (2002)

PEER REVIEWED PUBLICATIONS AND CONFERENCE PAPERS

Costa, M. R., Kim, S. Y., & Biocca, F. (2013). Embodiment and Embodied Cognition. In Virtual Augmented and Mixed Reality. Designing and Developing Augmented and Virtual Environments (pp. 333–342). Springer Berlin Heidelberg.

Costa, M. R., Qin, J., & Bratt, S. (In Press). Emergence of collaboration networks around large scale data repositories: A study of the genomics community using GenBank.*Scientometrics*.

Costa, M. R., Qin, J., & Wang, J. (2014). Research Networks in Data Repositories. Presented at the Digital Libraries 2014: ACM/IEEE Joint Conference on Digital Libraries, London: ACM.

Escalante, J., Butcher, S., Costa, M. R., & Hirshfield, L. M. (2013). Using the EEG Error Potential to Identify Interface Design Flaws. In Foundations of Augmented Cognition (pp. 289–298). Springer Berlin Heidelberg.

Hirshfield, L., Bobko, P., Barelka, A., Costa, M. R., Funke, G., Knott, B., & Mancuso, V. (2015). The Role of Human Operators' Suspicion in the Detection of Cyber Attacks. Accepted in: International Journal of Cyber Warfare and Terrorism.

Hirshfield, L., Costa, M. R., Paverman, D., Murray, E., & Hirshfield, S. (2013). Measuring the Trustworthiness of Ebay Seller Profiles with Functional Near-Infrared Spectroscopy. In Proceedings of the International Conference on Human Computer Interaction. Las Vegas.

Jones, J., Costa, M. R., Marlino, M., Qin, J., & Kelly, K. (2011). Value and impact metrics for open repositories. In Open Repositories 2011. Austin. Retrieved from http://eslib.ischool.syr.edu/pubs/RepositoryMetrics_OR2011_24x7.pdf

Serwadda, A., Phoha, V., Poudel, S., Hirshfield, L., Bandarra, D., Bratt, S., & Costa, M. R. (2015). fNIRS: A New Modality for Brain Activity-Based Biometric Authentication. In Biometrics (IJCB), 2015 IEEE International Joint Conference on Biometrics: Theory, Applications and Systems. Arlington, Virginia: IEEE.

OTHER PUBLICATIONS

Costa, M. R. (2010). Impact factor inflation: Measuring the gatekeeper effect in scientific literature. Proceedings of the American Society for Information Science and Technology, 47(1), 1–2. http://doi.org/10.1002/meet.14504701316

Costa, M. R. (2014). The dynamics of social capital in scientific collaboration networks. Proceedings of the American Society for Information Science and Technology, 51(1), 1–4. http://doi.org/10.1002/meet.2014.14505101137

Costa, M. R., & Qin, J. (2012). Analysis of networks in cyberinfrastructure-enabled research communities: A pilot study. Proceedings of the American Society for Information Science and Technology, 49(1), 1–4. http://doi.org/10.1002/meet.14504901244

Qin, J., Costa, M., & Wang, J. (2014). Attributions from data authors to publications: Implications for data curation. Presented at the 9th International Digital Curation Conference, San Francisco. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/IDCC14/230Qin_idcc_14.pdf

Rieks, A. R., Greene, D. T., Costa, M. R., Flaherty, M. G., & Solinger, C. (2011). Bridging theory and practice: connecting coursework to internships in LIS programs. In Proceedings of the 2011 iConference (pp. 763–764). ACM.

Small, R. V., Costa, M. R., & Rothwell, S. L. (2011). The role of information and motivation in the process of innovation. In *Academic entrepreneurship and community engagement: Scholarship in action and the Syracuse miracle* (pp 126-135). Bruce Kingma (Ed.). Northhampton: Edward Elgar.

PROFESSIONAL EXPERIENCE

RESEARCHER, Syracuse University, School of Information Studies, Syracuse New York – 2007 to present

- Conducted statistical analysis using R, Gephi, SPSS, Python, and Excel.
- Used data mining techniques to study collaboration trends in a large social network.
- Employed clustering algorithms and Markov chain modeling to study social network data.
- Designed and developed algorithms to process and analyze social network and time series data.
- Re-engineered algorithms to improve efficiency, resulting in 90% reduction in execution time or greater on datasets 10x larger than the datasets the algorithms were originally written for.
- Designed algorithms to leverage the parallel processing capabilities of the doParallel and foreach packages in R.
- Configured and managed VMs that handled the storage, retrieval, and analysis of several hundred million rows of data.
- Planned, developed, and trained teammates on database queries for analyzing large data sets.
- Oversaw the design, development, and implementation of a Neo4j graph database to store 200 million rows of social network data.
- Use machine learning algorithms on neurophysiological data to develop predictive models of mental states.
- Co-authored a $200,000 NSF grant to study the collaboration patterns of scientists.
- Designed and conducted studies as an intern in the Applied Neuroscience Division of the Air Force Human Performance Wing.

OWNER, M.R. Costa Consulting Group, LLC – 2013 to present

- Assisted in the design and population of a critical asset inventory database.
- Co-authored a $6 million Statewide Interoperability Communications grant.

RESEARCH LIBRARIAN, United States Army War College; Carlisle, Pennsylvania - 2006 to 2007

- Provided consultation to faculty and staff on the design of complex database queries to meet research needs.
- Planned and developed personal information platforms for senior DoD personnel using off-the-shelf web products.
- Developed custom search engines in order to provide dynamic information resource lists for various distance education courses.
- Provided in-depth research consultation to senior Department of Defense, Department of State, and military officials.
- Provided in-depth research consultation to senior officials from various international military organizations.
- Consulted with faculty and staff on the selection, acquisition, deployment, training, and troubleshooting of database and software systems.

CONSULTANT, Black Line Group, February 2006 to August 2007

- Analyzed technical projects for relevancy to IRS Research & Development Tax Credit.
- Composed narratives explaining project activities and relevancy to IRS Research.

OFF-CAMPUS LIBRARIAN, EAST REGION, Central Michigan University; Richmond, Virginia - 2005 to 2006

- Planned, developed, and implemented inter- and intra- departmental IT project to improve service quality and the department's efficiency.
- Consultation to faculty & staff on design of complex database queries to meet research needs.

RESIDENT LIBRARIAN AND VISITING INSTRUCTOR, University of Illinois at Chicago; Illinois - 2004 to 2005

- Web Page Consultant to the Electronic Resources Quadrant of the University Library.
- Conducted personal in-depth research consultation sessions.

COMBAT ENGINEER, U.S. ARMY - 1996-2001

- Conducted critical infrastructure inventory and analysis.
- Managed the personnel records for a 100 person company.
- Conducted peacekeeping operations in Bosnia-Herzegovina in accordance with Dayton Accords.
- Held secret clearance.

TECHNICAL PROFICIENCIES

R, RStudio Server, Amazon EC2, Ubuntu, MySQL, Neo4j, Gephi, SPSS, Excel, Network Workbench, Python, VBA