Essays on Collaboration, Innovation, and Network Change in Organizations

by

Russell James Funk

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Sociology)
in The University of Michigan
2014

Doctoral Committee:

Professor Jason D. Owen-Smith, Chair
Professor Gautam Ahuja
Professor Mark S. Mizruchi
Assistant Professor Maxim Vitalyevich Sytch

For Kylee,

whose love and encouragement made this dissertation possible.

ACKNOWLEDGEMENTS

A central thesis of this dissertation is that ideas are often stimulated and developed with the help of a good network, and this dissertation itself is no exception to that claim. Since coming to the University of Michigan, I have been incredibly fortunate to have the support of an outstanding network of mentors, collaborators, and friends who, beyond their contributions to this dissertation, helped me develop as a scholar and person.

I am especially grateful to Jason Owen-Smith, my dissertation committee chair, for his outstanding mentorship. I first ran into Jason's research while working on my undergraduate thesis. I have a distinct memory of sitting in the Regenstein library, reading some of his articles, and thinking, Wouldn't it be neat if I could do research like that someday? Little did I know that I'd have my chance just a few short years later. Jason made my decision to accept Michigan's offer to enroll in their sociology PhD program a no-brainer. After I moved to Ann Arbor in the summer of 2008, he graciously invited me to work on his U.S. Knowledge Economy (USKE) project, even before I had set foot inside a graduate school classroom. It is hard to imagine a better way to learn the craft of research than simply digging in and getting one's hands dirty. Although I'm sure it must have slowed his own progress, Jason encouraged this style of learning on the USKE project and all of our subsequent collaborations. I am deeply thankful to Jason for being so generous with his time, for his always-helpful feedback, and most of all, for pushing me to ask questions and pursue ideas far beyond what I had thought I was capable of when starting the PhD program. I hope that I can be

the same kind of mentor to my future students.

I am also incredibly appreciative of the support and encouragement of the other members of my dissertation committee, Mark Mizruchi, Maxim Sytch, and Gautam Ahuja. Mark's intellectual curiosity, breadth of knowledge, and methodological contributions have been deeply inspiring to me as a budding academic. I am also grateful to Mark for organizing Michigan's economic sociology community, which has served as my primary intellectual home throughout graduate school. Anyone who has had the privilege of sitting in a workshop with Mark will know that he has an uncanny ability to (constructively) sniff out weak assumptions and faulty logic. Whatever ability I have to make good arguments owes much to watching Mark ply this craft and to his generous feedback over the years on my own written work and presentations.

Maxim pushed me in many ways to expand my methodological horizons. Although his influence pervades each chapter of this dissertation, it is especially apparent in the last two, which build on insights from his research on network communities, network dynamics, and their respective connections to innovation. Over the past few years, I've also had the privilege of working with Maxim on several other projects. These collaborations have given me an incredibly valuable window into Maxim's systematic and rigorous approach to scholarship and, in so doing, improved this dissertation and my capabilities as a researcher.

Last but not least, I am indebted to Gautam for teaching me about the interconnections between sociology and strategy. I first began to see the exciting possibilities at the intersection of these two fields in Gautam's doctoral seminar on corporate strategy. That class is also where I completed the first drafts of Chapter II, which benefited greatly from his helpful guidance. Additionally, I am grateful to Gautam for always pushing me to look for bigger, higher impact theoretical contributions. Beyond research, I owe Gautam special thanks for opening the doors to Michigan Strategy for me, which in many ways came to feel like a second home department.

To Jason, Mark, Maxim, and Gautam—I hope the completion of this dissertation marks just the beginning of many exciting collaborative projects.

I benefited from the guidance and support of many other faculty members at the University of Michigan. Jerry Davis has in many ways been like an unofficial member of my committee, who not only provided feedback on early drafts of some chapters, but also took the lead in organizing the Interdisciplinary Committee on Organizational Studies (ICOS) and other related Michigan communities that played a major role in shaping my research. I am also thankful to Sue Ashford, Bob Axelrod, Sarah Burgard, Michael Heaney, Greta Krippner, Sandy Levitsky, Ned Smith, Sara Soderstrom, Brian Wu, Mayer Zald, and Minyuan Zhao for valuable comments on drafts of chapters and closely related projects. John Hollingsworth deserves special thanks for giving me the opportunity to explore many of the ideas developed in this dissertation in the health care domain. I am also grateful to Brian Noble and Sharon Broude Geva for introductions in the scientific computing community.

Graduate school would not have been the same without my amazing group of friends and fellow PhD students. I am especially grateful to Dan Hirschman, for outstanding friendship and many fun collaborations, and Helena Buhr, for paving institutional trails at Michigan that helped make this dissertation possible. For camaraderie and intellectual support, I also thank Jon Atwell, Johan Chu, Natalie Cotton-Nessler, Maria Farkas, Spencer Garrison, Mikell Hyman, Julian Katz-Samuels, Heeyon Kim, Yong Hyun Kim, Suntae Kim, Sun Park, Tristan Revells, Kelly Russell, Todd Schifeling, Lotus Seeley, and Matt Sullivan.

This dissertation was made possible by generous financial and technical support. I am particularly indebted to the National Science Foundation, the Rackham Graduate School, and Michigan's Department of Sociology for fellowships that gave me time to build new skills and the opportunity to take risks on data collection and analysis. Rick Smoke spent countless hours helping me with what must have appeared to

be an endless and seemingly random array of technical challenges, from interview transcription to web server administration. I also thank Mark Montague for his patience with my many questions about Michigan's Flux computing cluster.

Finally, I thank my family. My mother, Diane, has always supported and inspired me in whatever I do, and instilled in me the love of learning from a very young age. My father, Jim, together with my aunt, Linna Place, opened me to the idea of a career in academia and created big shoes that I am always seeking to fill. I'm also very grateful to my brothers, Todd and Chris, and members of my extended family, including John, Wyn, and Bill Hyzer, for encouragement and inspiring conversations. I dedicate this dissertation to my wife and partner, Kylee, for her endless support, for always inspiring me to do better, and for ensuring we enjoyed the journey through the ups and downs of graduate school. To Kylee—I look forward to sharing many more adventures in the next exciting chapter of our lives.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Essays on Collaboration, Innovation, and Network Change in Organizations

by

Russell James Funk

Chair: Jason D. Owen-Smith

This dissertation examines how internal communication and collaboration networks influence organizations' performance at innovation. Because some configurations may be better than others, I also consider strategies for changing networks. I structure my investigation around three studies.

The first study examines the effects of different networks in different geographic settings. Using data on 454 firms active in nanotechnology, I find that sparse networks of inventors help geographically isolated firms retain diverse knowledge and promote innovation. By contrast, firms located close to industry peers benefit from highly connected networks among their inventors that facilitate information processing.

In the second study, I examine the effects of network structure in an investigation of brokers. A broker is a person connected to people who are not tied to each other. Studies find that brokers have better performance on many metrics. However, little is known about how brokers affect their contacts. Using data on the networks of over 18,000 inventors at 37 pharmaceutical firms, I examine the effect of connection to a broker. To disentangle causality, I focus on changes among inventor's existing contacts, where the decision to connect was made before the contact became a broker

and therefore is exogenous to performance. I find that although becoming a broker positively affects performance, the opposite is true for having a connection to one.

After focusing on performance in the first two studies, the final study considers reshaping networks. Using data on 23 million exchanges among 1.3 million members of 25 technical communities, I examine how a common organizational feature—knowledge categorization systems—influences bridging. Bridging ties create and strengthen connections among otherwise distant people and therefore are powerful tools for adapting networks. Categorization systems facilitate bridging by helping people locate distant peers. However, they may also inhibit bridging. First, as a categorization system grows large, it becomes harder to use and people are less able to establish distant ties. Second, as a categorization system decouples from real expertise, its value for bridging diminishes. Finally, as norms of evaluation vary more widely in an organization, people make fewer exchanges with unfamiliar peers. All three ideas are supported.

CHAPTER I

Introduction

Innovation is a social activity. Although we sometimes imagine that breakthroughs emerge from people working on their own—the scientist toiling away in her lab, the budding entrepreneur tinkering in his garage—today, perhaps more than ever, revolutionary new discoveries, products, and ideas are the result of collaborations.

It is easy to see why—social relationships help people achieve better outcomes. Decades of research demonstrate that working with diverse collaborators helps enhance creativity by exposing people to different ideas and perspectives. Highly connected individuals work more efficiently because they have knowledge about the capabilities of others inside their organizations and know where to look for assistance. And cohesive networks, with many dense interconnections, promote creative risk taking by fostering supportive environments where people feel comfortable sharing unconventional ideas.

As evidence accumulates about the importance of relationships, many organizations are taking an active role in cultivating and managing internal communication and collaboration networks. For example, in November 2013, Microsoft announced the end of its infamous stack ranking system, whereby managers were required to rank the relative performance of their employees on a fixed distribution, regardless of whether everyone had met or even exceeded expectations. One reason for aban-

doning the system was to promote a more connected company, which was limited by employee's fears of working with others who may outshine them at ratings time. Illustrations may also be found outside the corporate sector. In May 2012, for instance, the University of Michigan launched MCubed, a funding initiative designed to stimulate projects involving researchers from different disciplines, who may otherwise not have the opportunity to work together.

Although there is substantial research showing how and why communication and collaboration ties help individuals perform better, surprisingly little is known about the effects of larger internal network configurations—the aggregate of individual members' interconnections—on broader organizational effectiveness. Existing work, for example, offers few theoretical tools to explain whether greater internal connectivity—like that sought by Microsoft and the University of Michigan—is likely to result in better outcomes. Will a connected Microsoft produce more breakthroughs? Scholars also have yet to consider the conditions under which deliberate efforts at network change are likely to succeed. Will the University of Michigan's bottom-up, researcher-driven approach to promoting integration lead to enduring ties across disciplines?

The purpose of this dissertation is to develop a conceptual framework and methodological approach and to present some preliminary evidence that help address these kinds of questions.

1.1 Background

Intraorganizational networks, like all social networks, are defined by a set of actors and the relations among them. Relevant actors may include divisions, teams, or people (Guler and Nerkar, 2012; Hansen, 1999; Hargadon and Sutton, 1997; Mizruchi et al., 2011). Unlike other networks, the membership and activities of intraorganizational networks are circumscribed by the boundaries of an organization.

Researchers have long understood the importance of internal networks for organi-

zational processes and outcomes. For example, pioneering work by March and Simon (1958) argued that patterns of communication are related to an organization's ability to manage uncertainty and to the distribution of power and influence among its units. Similarly, in an early study of organization–environment relations, Burns and Stalker (1961) found that flexible channels of internal communication (as opposed to rigid hierarchies) are beneficial for firms that operate in more dynamic industrial sectors.

More recently, scholars have begun to systematically examine the implications of intraorganizational networks for innovation (Allen, 1977; Burt, 2004; Reagans and McEvily, 2003). The general findings of this work reveal that the structure of relationships among divisions, teams, and people may enhance creative performance through two general mechanisms. First, intraorganizational networks facilitate information sharing, which in turn helps to expose people to the diverse ideas and perspectives needed for solving complex problems (Paruchuri, 2010; Reagans and McEvily, 2003). Although context matters, many studies suggest that acquiring diverse information is easier in networks that are less densely interconnected and have fewer redundant ties (Burt, 2004). Second, networks inside organizations help people locate colleagues that they can mobilize around ideas that require collaborative development (Hansen, 1999; Obstfeld, 2005; Uzzi, 1997). In contrast to acquiring information, networks with many overlapping and highly cohesive ties tend to be better for mobilization (Coleman, 1988; Obstfeld, 2005).

Existing research on intraorganizational networks has led to many valuable insights about the important role of intraorganizational networks for promoting innovation. However, work in this area is also hampered by several empirical and theoretical limitations. Empirically, prior investigations of networks in organizations have been carried out in the context of large, established enterprises, often multinational corporations (Burt, 2004; Hansen, 1999; Hargadon and Sutton, 1997; Nerkar and Paruchuri, 2005; Obstfeld, 2005; Singh et al., 2010; Tsai, 2001). Moreover, given the expense

of collecting detailed network data across multiple settings, existing studies generally examine only one or two organizations and adopt divisions, teams, or people as their primary units of analysis, and therefore are unable to consider organizational outcomes.

Networks are clearly important in larger, established organizations, where the scale of operations can make communication among relevant parties particularly challenging. However, internal networks also have fundamental consequences for newer organizations. For example, in a seminal essay, Stinchcombe (1965) argues that patterns of communication are important for understanding why recently established organizations tend to have high failure rates—what he terms the "liability of newness." "For some time until roles are defined," Stinchcombe writes, "people who need to know things are left to one side of communication channels. John thinks George is doing what George thinks John is doing" (1965, 148-9). Furthermore, because their members are often strangers to one another, young organizations lack the trusting and reliable bonds that are found among colleagues at more established entities. Put differently, the potential benefits of networks for innovation, including information sharing and supportive relationships, may be difficult for entrepreneurs to attain.

Changes in the U.S. economy also suggest the need for greater attention to networks in organizations of all sizes and the consequences of these social structures for innovation. Although large firms (i.e., those with more than 25,000 employees) remain dominant when it comes to research and development (R&D) spending (accounting for 33% of U.S. industrial R&D in 2008 relative to 22% for companies employing 5-499 people) the locus of commercial innovation has shifted in many sectors from sizable corporate facilities like Bell Labs to more flexible startups (Drucker, 1985; Piore and Sabel, 1984; Powell et al., 1996; Saxenian, 1994; Stuart et al., 2007).[1] Because of differences in resources and strategy, the R&D operations of small and large firms are

[1]Figures on R&D expenditures are drawn from the National Science Foundation and U.S. Census Bureau's Business R&D and Innovation Survey (Wolfe, 2010).

often very different from one another. And as result, the applicability of findings from existing intraorganizational network research to this emerging sector of the economy are not always clear (Katila and Shane, 2005).

In addition to these empirical matters, theoretical considerations also call for a fresh approach to the study of intraorganizational networks. Existing work focuses on ego networks, which are the portfolios of ties held by individual actors like people or divisions. Investigations into the effects of global intraorganizational network structures on innovation, by contrast, are rare (Granovetter, 1992; Ibarra et al., 2005; Phelps et al., 2012). However, the evidence that does exist suggests that global network functioning does not always align with what may be anticipated from years of research on ego networks. For example, using a computational model, Lazer and Friedman (2007) found that when networks are less globally connected—and therefore slower at diffusing information—people (or agents, in their model) generate more diverse solutions to complex problems. Put differently, global network inefficiencies may sometimes lead to better performance (Fang et al., 2010). Similarly, in a rare empirical study, Guler and Nerkar (2012) show that although the effect of cohesion on innovation is sometimes ambiguous at the ego network level (because it enhances mobilization but also decreases diversity) cohesion at the intraorganizational network level should have a uniformly negative influence. The authors reason that global cohesion is detrimental to innovation because it is costly to maintain and the benefits of cohesion are generally local. Finally, studies outside the innovation literature demonstrate that what may appear to be relatively marginal nodes from an ego network perspective (e.g., those with few connections) can, depending on their global position, have a major influence over the flow of information through a larger network (Dodds et al., 2003; Liu et al., 2011).

To summarize, existing research focuses on the relationships of individual actors within established organizations. Variation in global network structure across orga-

nizations has been overlooked. It is important to correct this oversight not only to develop more complete theories of networks and organizations, but also to improve the ability of those theories to inform practice. Managers oversee collections of divisions, teams, or people, and therefore the ties of these actors individually may not be as relevant to managers as the broader patterns of exchange among them.

1.2 Summary of Chapters

Over the course of three studies, this dissertation examines how internal communication and collaboration networks influence organizations' performance at innovation, along with possibilities for reshaping those networks. Below, I provide a brief overview of each study.

1.2.1 Geography, Networks, and Innovation

The first study (Chapter II) explores the contingent effects of different intraorganizational collaboration and communication network configurations by examining a two complementary questions of economic geography. First, given the importance of spatial proximity for enhancing knowledge flows across organizations and the related benefits of access to diverse knowledge for innovation, how do geographically isolated firms develop novel products? And second, how do firms located in close proximity to industry peers generate distinctive ideas relative to nearby competitors, who have access to similar local knowledge?

I propose that intraorganizational network structure offers one answer. Using novel data on 454 U.S. firms active in nanotechnology, I find that sparse networks of inventors help geographically isolated firms retain diverse knowledge and promote innovation. By contrast, firms located in close proximity to many industry peers benefit from highly connected, cohesive networks among their inventors that facilitate information processing. These findings establish the importance but contingent

benefits of intraorganizational network structure for innovation.

1.2.2 Brokerage and Innovation

Building on the findings of Study 1, the next study (Chapter III) explores the effects of internal communication and collaboration networks in an investigation of brokers. A broker is a person who has disconnected contacts. Many studies find that brokers tend to have better performance, at least in part because of the benefits they gain from their unique network positions. Far less is known, however, about the implications of brokerage for those other than the person in the broker role.

The absence of research in this area is surprising because by definition, brokerage involves the broker and at least two other people. Do these other individuals benefit from their mediated connection? And if they do, then how do their returns compare to the broker's?

I address these questions with data on the intraorganizational networks of over 18,000 inventors at 37 pharmaceutical firms. To help disentangle causality, I use a novel estimation technique based on propensity score weighting that is unbiased if the specification is correct for either the exposure or outcome equations. Furthermore, in my tests of the effects of having a connection to a broker, I focus on changes among existing contacts, where the decision to connect is exogenous to performance because it was was made prior to when the contact came to occupy a brokerage position. Consistent with the idea that there are negative spillovers to brokerage, I find that becoming a broker has a positive effect on performance, but the opposite is true for having a connection to one.

1.2.3 Network Change

Finally, Study 3 (Chapter IV) examines network change. As the examples of Microsoft and the University of Michigan illustrate, organizations often attempt to

adapt internal communication and collaboration networks to changing external environments and evolving strategies. Yet surprisingly little research addresses how leaders might reshape internal networks.

Bridging ties create and strengthen connections among otherwise distant groups of people in an organization and therefore are powerful tools for adapting networks. In general, the systems that organizations use for categorizing and mapping their knowledge—what I call "knowledge categorization systems"—should facilitate bridging by making it easier for people to connect with peers in their organization who have relevant expertise. However, using data on millions of exchanges among members of 25 online technical communities, I find that in some cases, knowledge categorization systems may inhibit bridging. First, when a categorization system grows large, cognitive limitations make it difficult to use and therefore people are less able to establish ties with distant partners. Second, when a categorization system decouples from the actual distribution of expertise within an organization, its value for promoting bridging diminishes. Finally, when the norms used to evaluate the quality of contributions vary widely across an organization's categorization system, people are less likely to risk sharing knowledge outside their comfort zones and will make fewer exchanges with unfamiliar peers in distant parts of their organizations. These findings suggest that organizations' formal structures may serve as levers for guiding network change.

CHAPTER II

Making the Most of Where You Are: Geography, Networks, and Innovation in Organizations

2.1 Introduction

Social scientists have long recognized the importance of geography for innovation (Allen, 1977; Florida, 2002; Marshall, 1890; Pouder and St. John, 1996; Whittington et al., 2009). Regions like Silicon Valley and Boston's Route 128, home to concentrations of technology companies, catalyze innovation by facilitating face-to-face interaction, increasing the likelihood of chance encounters, and allowing firms to monitor competitors, all of which promote the local diffusion of ideas (Audretsch and Feldman, 1996; Bell and Zaheer, 2007; Fleming et al., 2007). In dynamic, innovation-intensive industries, access to these local knowledge sources helps firms develop competitive advantages (Porter and Stern, 2001; Saxenian, 1994).

More recently, researchers have extended theories of geography and innovation by showing that firms differ in their ability to reap the benefits of their locations. Knowledge spillovers from proximate organizations help firms develop ideas that are novel relative to those of distant rivals but do little to differentiate them from local competitors (Hervas-Oliver and Albors-Garrigos, 2009; McEvily and Zaheer, 1999; Tallman et al., 2004). Consequently, studies emphasize that firms benefit most from proximity

to other organizations if such firms can also access more exclusive sources of knowledge (Bathelt et al., 2004; Bell, 2005; Malmberg and Maskell, 2006). For instance, alliances with local collaborators let firms access complex or proprietary knowledge while limiting its diffusion among neighbors (Laursen et al., 2012; Owen-Smith and Powell, 2004). Ties to distant collaborators let firms acquire knowledge that is unavailable locally (Bell and Zaheer, 2007; Whittington et al., 2009). Embeddedness in scientific communities (Gittelman, 2007; Owen-Smith and Powell, 2004) and recruiting skilled employees (Almeida and Kogut, 1999; Zucker et al., 1998) also provide resources that, when leveraged with locally and informally acquired knowledge, help firms innovate.

Despite these advances, theoretical explanations of geography and innovation in organizations remain limited. Notably, although researchers have made progress in identifying how firms acquire information from external sources, little is known about how organizations internalize, adapt, and use the knowledge that diffuses to them geographically. Accounting for how firms process information from nearby sources, however, is essential for specifying the conceptual link between geography and innovation. Proximity is helpful because it offers access to diverse knowledge that firms can recombine in novel ways to make discoveries (Schumpeter, 1934). But as this access increases, so too do the number of potential recombinations and the possibility of cognitive overload. How, then, do firms that are proximate to many other organizations filter and process the potentially vast amounts of knowledge available to them locally?

Firms in geographically isolated locales face a very different set of challenges that have received only limited attention in existing theory. As proximity to other organizations decreases, so too does access to local, informal knowledge spillovers. Given the importance of exposure to diversity for innovation, current approaches suggest that isolated firms are severely disadvantaged. Although on average these

firms might fall short of those with greater proximity to peer organizations, many companies located far from peers do produce important innovations.[1] Forging ties to geographically distant partners can help isolated firms supply their employees with fresh knowledge (Alnuaimi et al., 2012; Rosenkopf and Almeida, 2003). However, long distance collaborations are also challenging to manage and often not conducive to complex knowledge transfer (Alnuaimi et al., 2012; Bathelt et al., 2004; Sorenson and Stuart, 2001). Recruiting skilled employees from distant areas offers another possibility, but research shows that among such workers, mobility is often localized (Almeida and Kogut, 1999; Breschi and Lissoni, 2009; Zucker et al., 1998). Even when isolated firms are able to recruit over long distances, knowledge transfer can be limited by legal and coordination barriers (Agarwal et al., 2009; Singh and Agrawal, 2011). Accounting for how geographically remote firms produce innovations despite these barriers is important for building a more complete theory of geography and innovation.

In this chapter, in an effort to overcome some of these limitations of existing theory, I develop a new approach to explaining how firms' innovative performance relates to the makeup of their local environments. I build on prior studies of proximity and innovation, but differ in that I shift attention to a lower level of analysis by focusing on patterns of collaborations *within* firms. I connect insights from macro research that emphasizes the external determinants of innovation (Whittington et al., 2009; Zucker et al., 1998) with micro studies that point to the influence of more internal social network structures (Guler and Nerkar, 2012; Obstfeld, 2005). Informal

[1]Mayer (2011) details numerous cases of innovative firms that are geographically isolated from industry peers. For instance, Micron Technology, an electronics manufacturer, has been located in Boise, Idaho, since its 1978 founding—far outside established semiconductor manufacturing hubs in California, Oregon, Texas, and Arizona. Micron has introduced major innovations in electronics production and memory chips. Burleigh Instruments offers another illustrative example. This company—an early developer of the scanning tunneling microscope, atomic force microscope, and other instruments used in nanotechnology R&D—was located in the small community of Fishers, New York, until its 2000 acquisition (Mody, 2011, 140ff). For examples of companies that have struggled because of their geographic isolation, see Saxenian (1994, Chap. 3).

employee networks are important because they facilitate knowledge sharing—which in turn helps with information processing, project coordination, and ensuring efficient resource use (Hansen, 1999). Communication across units also leads to knowledge creation as members of one division adapt others' expertise to novel uses (Hargadon and Sutton, 1997).

I argue that firms' innovative performance can be enhanced by their local environments, but these geographic benefits are contingent on the structure of collaboration networks among their employees. The logic of the argument is as follows: Proximity allows firms to capture large volumes of knowledge through spillovers from nearby organizations, but as the volume of local knowledge increases, so too does the difficulty of internalizing, adapting, and using that knowledge. In these environments, cohesive networks that promote flexibility, communication, and collaboration are beneficial (Burns and Stalker, 1961). Firms in more remote areas, by contrast, should have less difficulty sifting through the smaller volumes of information they encounter locally. However, because these firms also have diminished access to spillovers, their employees are less likely to acquire knowledge outside their workplace. Here, less connected networks help to alleviate the challenges of isolation. Because such networks are slower at diffusing information, they preserve diverse ideas internally (Granovetter, 1992; Lazer and Friedman, 2007) and increase possibilities for recombination. In sum, external, geographically defined environments present opportunities and constraints, but internal factors moderate the degree to which a firm can make use of or is held back by them.

Examining the interdependencies between firms' local environments and the pattern of collaborations among their employees leads to new insights. Most significantly, the results of this study demonstrate that although firms in places like Silicon Valley and Route 128 derive benefits from proximity, performance gains are moderated by the structure of collaboration networks that connect their employees. Moreover, de-

pending on the degree of fit between a local environment and an inventor network, firms in isolated locales can in some cases have better performance than similar organizations that are proximate to industry peers. These results hold even in analyses that control for firms' access to alternative sources of local and distant knowledge, including collaborative ties, labor mobility, and embeddedness in scientific communities.

Beyond geography and innovation, this chapter also contributes to network theory. Studies have shown that the functioning of social network structures is contingent on a variety of actor (Fleming et al., 2007; Mehra et al., 2001), relationship (Reagans and McEvily, 2003; Tortoriello and Krackhardt, 2010), and task characteristics (Hansen, 1999). However, this work focuses on the structure of individual-level (or ego) networks. Little is known about the importance of such contingencies in the overall structure of relations at the global network level (Granovetter, 1992; Ibarra et al., 2005; Phelps et al., 2012). My results suggest that context matters tremendously for explaining how global network structure affects innovation. However, the findings also offer evidence that some well-documented ego network contingencies do not translate clearly to the global level. For instance, theories developed at the ego level emphasize the importance of increasing connectivity—either through bridging structural holes or building dense ties among collaborators—for actors who seek to transfer knowledge and produce innovations. By contrast, I show that at the global level, sometimes less connectivity is advantageous, particularly for groups that seek to generate diverse ideas but have limited exposure to external knowledge. This finding also runs contrary to much research on community social capital, which heavily emphasizes the benefits of dense ties and cohesive relationships (Coleman, 1988; Putnam, 2000).[2]

Moreover, this study contributes to ecological and institutional research that examines the effect of local community characteristics on organizational behavior (Mar-

[2]Lin (1999, 33ff) documents this tendency to emphasize cohesion in the community social capital literature.

quis, 2003; Romanelli and Schoonhoven, 2001). Research in this tradition emphasizes that organizations are heavily influenced by their neighbors. In an influential review, Freeman and Audia (2006, 156) pointed to the need for more attention to the potentially heterogeneous effects of geographically defined environments, noting, "it is possible... that some organizational forms are strongly affected by the local context, whereas others are largely insulated from it." The present study moves toward this goal by showing how the extent to which firms garner benefits or incur costs from their local environments depends in part on the collaboration patterns of their inventors.

Below, I develop hypotheses to explore the argument that geography and intraorganizational network structure have interdependent effects on innovative performance. I test these hypotheses using data on U.S. firms involved with nanotechnology R&D.

2.2 Theory and Hypotheses

2.2.1 Geographic Proximity and Knowledge Spillovers

Geographic concentration is an important feature of many industries (Florida, 2002; Saxenian, 1994; Sorenson and Audia, 2000). Proximity offers benefits such as lower transportation costs and convenient access to skilled labor (Porter and Stern, 2001). When it comes to innovation, however, often the greatest advantages of being located near other organizations are those resulting from the increased access to knowledge via spillovers. Geographically localized knowledge spillovers are a type of positive externality characterized by the transfer of knowledge between parties as a result of their proximity (Audretsch and Feldman, 1996). Such spillovers are invaluable for firms operating in dynamic industries because they help ensure that employees are frequently exposed to new knowledge and ideas. To the extent that innovation emerges through arranging existing ideas and materials into novel recombinations, inventors should benefit from having a diverse and frequently changing knowledge

base at their disposal (Schumpeter, 1934).

Several factors help to account for why proximity is likely to increase knowledge spillovers. Most broadly, geographic concentrations of firms operating in the same industry are often characterized by what Bathelt et al. (2004, 38) call "buzz": "the idea that a certain milieu can be vibrant in the sense that there are lots of piquant and useful things going on simultaneously and therefore lots of inspiration and information to receive for the perceptive local actors." Others highlight four more specific mechanisms that promote knowledge flows among proximate organizations. First, proximity enables firms to stay informed of technological frontiers by allowing them to monitor competitors' activities (Porter and Stern, 2001; Sorenson and Stuart, 2001). This information lets firms quickly meet the needs of customers and prioritize their R&D on promising areas. Knowledge of technological developments also helps firms create novel products by integrating competitors' discoveries into their own offerings (Christensen, 1997). Second, geographic concentration increases the likelihood of chance encounters. For instance, participation in community clubs, children's activities, and other local events increases opportunities for employees of different firms to interact (Marquis, 2003; Putnam, 2000); the probability of spillovers rises as work-related topics enter their conversations. Third, proximity helps to create and sustain informal social and professional networks that are not beholden to any particular organization (Saxenian, 1994). Such networks help channel diverse knowledge among local actors (Owen-Smith and Powell, 2004). Finally, over time, interaction among employees of neighboring organizations can result in common conventions, interpretive schemata, and other local institutions that improve the ease and efficiency of absorbing knowledge from proximate sources (Malmberg and Maskell, 2006). This discussion leads to a baseline hypothesis:

Hypothesis 1. *Increases in proximity to other companies that perform related R&D are positively associated with a firm's innovative performance.*

2.2.2 Networks and Innovation in Organizations

The prior section theorizes the importance of geographic proximity for innovation. However, arguments about proximity and innovation lead to puzzling observations that are not well accounted for by existing theory. For instance, how are firms in more remote places able to maintain diverse knowledge bases? As proximity increases, how do companies sift through the volumes of information available locally to find what is valuable? If, as emphasized by prior research, a firm's ability to access exclusive sources of external knowledge is important, then theories of geography and innovation should also be able to account for how firms that are proximate to many other organizations filter, process, and make sense of such knowledge—in addition to the knowledge these firms acquire informally from neighbors. Among organizations that are more geographically isolated, the ability to process high volumes of information from local sources should be less necessary. However, knowledge transfer over long distances—through collaborative ties, labor mobility, or other sources—is often difficult (Bathelt et al., 2004; Singh and Agrawal, 2011); thus it remains to be explained how such firms can keep their employees supplied with the diverse information necessary for stimulating innovation. Insights into these questions about the relationship between geography and innovation can be found by recognizing that the benefits firms obtain from their local environments are moderated by the structure of collaborations among their employees. Different degrees of proximity may be more or less advantageous, depending on structure of these networks.

Consider the importance of intraorganizational networks—defined as the set of relationships among members of a firm created by their common participation on projects—for firm-level innovation. First, intraorganizational networks are helpful for stimulating ideas and ultimately creating new knowledge. Interacting with colleagues aids inventors in their search for novel combinations by allowing them to access their organization's existing knowledge base and by exposing them to diverse

problem-solving perspectives (Hansen, 1999; Hargadon and Sutton, 1997). Second, these networks serve broadly as tools for information processing and idea development. Frequent interaction builds trust and fosters environments in which inventors feel comfortable seeking help when they encounter roadblocks and turning to others for assistance in identifying resources (Borgatti and Cross, 2003; Obstfeld, 2005; Tushman and Nadler, 1978).

Different network structures are better at providing some of these benefits than others. At the individual (or ego) level, one prominent position emphasizes the value of networks for innovation as deriving from the ability of actors to serve as brokers who combine the knowledge of diverse groups in new ways (Burt, 2004; Hargadon and Sutton, 1997). By sitting in the interstices between communities, brokers attain broader understandings of problems facing different groups and untried solutions. A second perspective emphasizes the benefits of cohesive networks for innovation (Obstfeld, 2005; Reagans and McEvily, 2003). Cohesive networks are highly connected with many redundant ties. Analysts who emphasize cohesion cite trust, coordination, and improved complex knowledge transfer as the main benefits of such network structures.

Researchers also stress that both network structures have limitations. Though brokerage is useful for obtaining diverse information, it is less valuable for transmitting complex or confidential knowledge, which requires frequent interaction and willingness of a sender to share knowledge with a receiver (Hansen, 1999; Tortoriello and Krackhardt, 2010; Uzzi, 1997). Further, brokers often have trouble mobilizing support for their ideas because the communities they bridge may have little interest in working together (Obstfeld, 2005). But cohesion too has drawbacks. Dense ties create the risk that individuals will rely too heavily on the knowledge of their immediate peers, with whom they feel most comfortable (Perry-Smith and Shalley, 2003). As group knowledge ages, it becomes harder to find untried combinations.

My approach extends theories of geography and innovation by building on these insights from ego-level research on the contingent benefits of different network structures. I depart from this work by focusing on the global structure of collaborative ties within firms—i.e., intraorganizational networks—rather than the portfolios of ties of individuals—i.e., ego networks. I differentiate between two structures to characterize intraorganizational collaboration patterns.

Inefficient networks have low connectivity and diffuse information slowly (Fang et al., 2010; Lazer and Friedman, 2007). They foster innovation by creating opportunities for brokers to join disconnected network areas. Moreover, when viewed from the firm level, inefficient networks also facilitate parallel problem solving, in which different individuals or groups work independently on the same task (O'Reilly and Tushman, 2004). Because "fragmentation of network structure... [reduces] homogeneity of behavior" (Granovetter, 1992, 36), disconnected parties are more likely to produce novel solutions, one of which might be superior to the others. Note that low connectivity and slow diffusion are distinct features of inefficient networks, as I use the term. Some networks, such as "small-world" structures, have sparsely connected regions, with dense pockets of highly cohesive ties. Such networks are good at rapidly diffusing information, but as a result, they are less likely to preserve diversity (Lazer and Friedman, 2007).

For developing new knowledge, inefficient networks are particularly strong because they stimulate and preserve diverse ideas, perspectives, and approaches. By contrast, global network structures that are more cohesive are advantageous for processing information and refining ideas. As Tushman and Nadler (1978, 618) explain, "because highly connected networks are relatively independent of any one individual, they are less sensitive to information overload or saturation than more limited networks." In these networks, information diffuses quickly (Aral and Van Alstyne, 2011). Parallel problem solving is less likely to take place than are joint approaches, but as a result,

individuals are also less prone to wasting time hunting down bad leads that have been pursued by others in their organization and can quickly move to other investigations.

2.2.3 Geographic Proximity and Intraorganizational Collaboration Networks

Intraorganizational network structures are likely to moderate the effects of proximity on innovation. Areas with dense concentrations of organizations provide their member firms with frequent opportunities to access new knowledge from spillovers. However, firms in these locations are not without challenges. Instead of having to manage problems resulting from a lack of access to external knowledge from local sources, firms in highly concentrated areas like Silicon Valley may find it challenging to process all potentially relevant spillovers. In those settings, inefficient networks are likely to be harmful while networks consisting of more cohesive relationships should offer the biggest payoffs.

Prior work suggests that because they promote frequent communication, joint problem solving, and heightened group focus, cohesive networks are useful for processing large amounts of complex information from external sources such as spillovers (Hansen, 1999; Reagans and McEvily, 2003; Tushman and Nadler, 1978). Although these features may offer advantages to firms in any region, they are likely to be particularly valuable in heavily concentrated areas, where employees have opportunities to interact with diverse parties not affiliated with their firm—and thus problems of knowledge stagnation from cohesion are less problematic. Through local interactions, employees acquire new knowledge. Cohesive networks make it easier for individuals to identify and share that knowledge with their most relevant colleagues. Moreover, through their encounters outside their firm, individuals might acquire valuable knowledge, but because they lack the relevant expertise, do not recognize its importance. Because cohesive ties entail frequent, high-volume information exchange (Aral and

Van Alstyne, 2011), this knowledge is more likely to be discovered in conversations with colleagues who have the necessary background and training. Together, these observations suggest a second hypothesis:

Hypothesis 2. *As proximity to companies that perform related R&D increases, a firm has greater innovative performance if the cohesiveness of its intraorganizational collaboration network also increases.*

The arguments of the previous sections suggest that unlike firms in concentrated locales, those in a more isolated environment with few neighboring industry peers will be at a disadvantage in terms of their ability to capture local spillovers, and consequently their employees will likely be less successful innovators. Though it may be impossible for these firms to match the innovative capability of those located closer to industry peers, they may attenuate some of the problems that stem from the absence of local sources of new knowledge, depending on the configuration of their intraorganizational networks. Recall that inefficient networks are ideal for generating and sustaining diversity. Because they are less connected and are slow at diffusing information, these networks promote parallel problem solving and create opportunities for brokerage, both of which are likely—at the firm level—to lead to novel discoveries (Lazer and Friedman, 2007).

Firms in more isolated areas may perform poorly if their intraorganizational networks are characterized by overly cohesive relationships. Such companies already have few sources of new knowledge to draw from in their immediate external environment to stimulate innovation. Because they enhance information diffusion, dense ties among inventors within these firms should also lead to more homogeneous ideas (Granovetter, 1992), which makes generating the kind of novel recombinations that underpin new discoveries challenging. This discussion leads to a final hypothesis:

Hypothesis 3. *As proximity to companies that perform related R&D decreases, a*

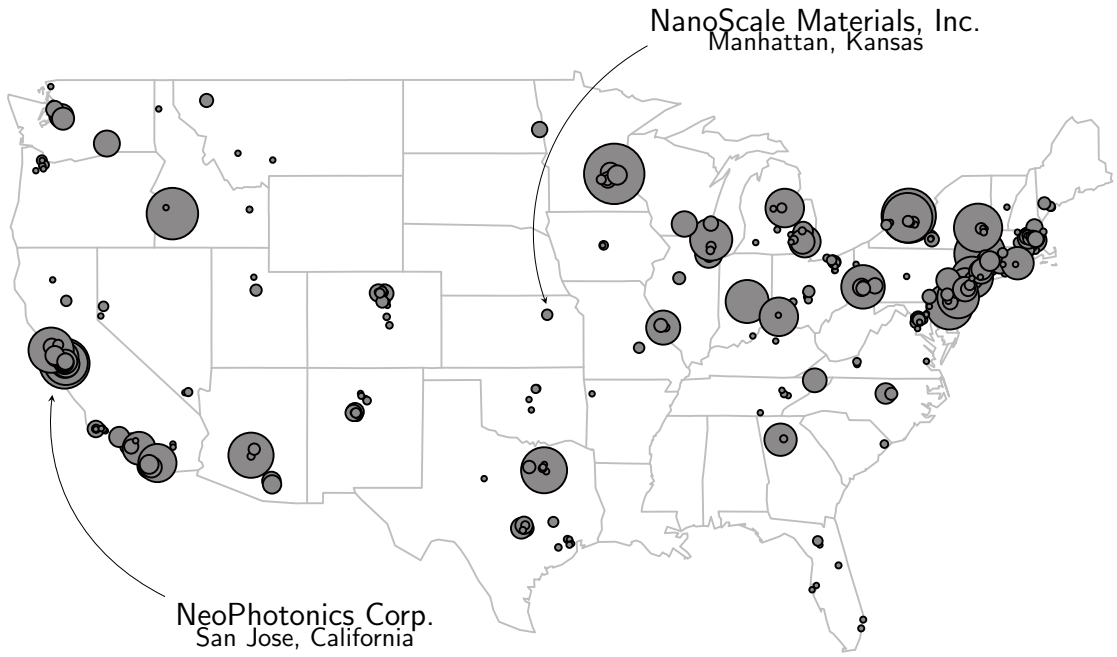

Figure 2.1: Map of sample nanotechnology firms, 2003. The plotting characters are weighted according to the cumulative number of nanotechnology patents held by a firm. Patent counts were log-transformed for display purposes.

firm has greater innovative performance if the inefficiency of its intraorganizational collaboration network increases.

Hypotheses 2 and 3 make predictions about the influence of intraorganizational network structure on innovation for firms with varying proximity to industry peers. They suggest that performance will be greatest when the characteristics of the network and geographic milieu fit in ways that play to their respective strengths and overcome their weaknesses. Although the hypotheses do not address the relative *magnitude* of the predicted effects in different regions, the theoretical considerations above suggest that firms operating in close proximity to industry peers may derive the greatest benefits from "fit." To the extent that firms in more isolated locales rely on inefficient networks, their employees should have access to more diverse knowledge from their colleagues. However, these firms also forego the benefits of cohesive social structures such as information processing and joint problem solving. (Though some of these benefits may be less critical, because external spillovers are less extensive as

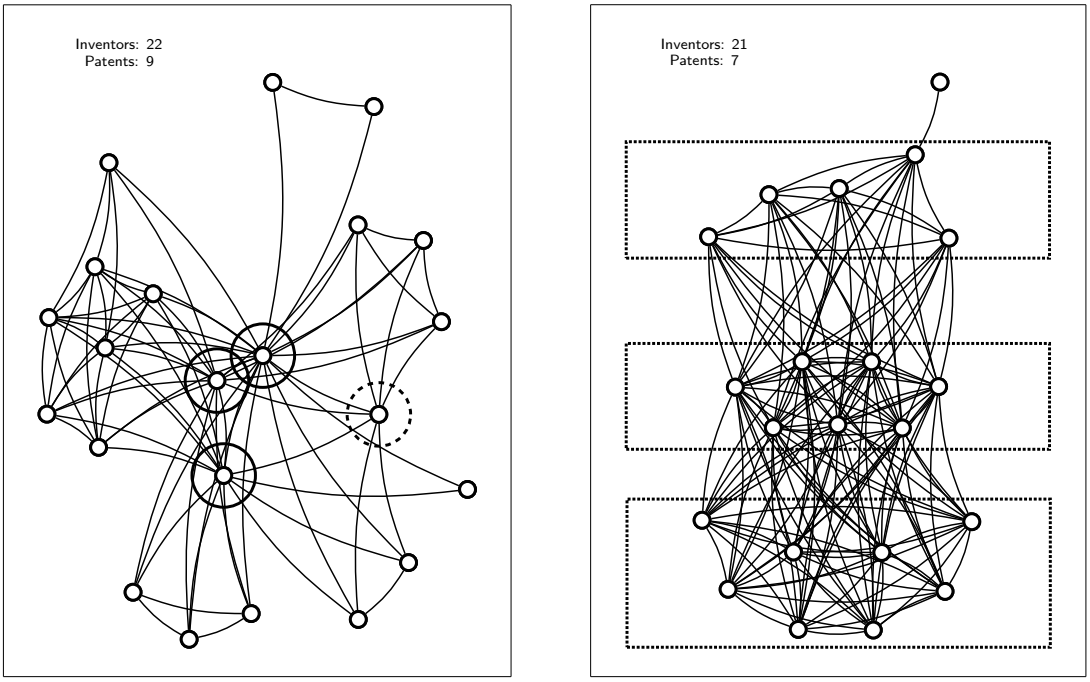

Figure 2.2: Networks of nanotechnology inventors at two firms, 2003. Nodes circumscribed by solid black lines are brokers, while those surrounded by dashed lines occupy more peripheral brokerage positions. The densely dotted boxes delineate communities within which collaboration may be more frequent. Note that the differences in structure do not result from differences in the number of inventors or patents, which are comparable for the two networks. Only main components are shown.

proximity decreases.) Firms in more concentrated locales, by contrast, should benefit from cohesion while also exposing their employees to diverse knowledge from external spillovers.

To put the hypotheses in perspective, in Figure 2.1 I show the geographic distribution of U.S. nanotechnology firms alongside the plots of collaborations (on patents) among inventors at NeoPhotonics Corp. and NanoScale Materials, Inc. displayed in Figure 2.2. NeoPhotonics is located in San Jose, California, the heart of Silicon Valley. Hypothesis 2 predicts that in this environment, a firm will perform better if it has more cohesive intraorganizational networks. As shown in the right panel of Figure 2.2, NeoPhotonics fits this profile. A total of 21 inventors produced seven patents. Collaboration is common. NanoScale Materials, in Manhattan, Kansas, provides a useful example of Hypothesis 3, which predicts that in more isolated regions firms will perform better if their networks are less connected. Comparable in size to

NeoPhotonics, the left panel of Figure 2.2 shows 22 inventors who collaborated on nine patents.

2.3 Research Setting

I tested my hypotheses in the context of commercial nanotechnology R&D. Nanotechnology is a scientific and technological field focused on the manipulation of matter at the atomic scale. The development of nanotechnology began in the early 1980s with the invention of the scanning tunneling microscope (STM). The STM, and later the atomic force microscope (AFM), gave researchers the remarkable ability to image and move individual atoms on material surfaces. As a result, the technology allows materials, drugs, electronics, and a virtually limitless array of other structures to be designed and built atom-by-atom. Much interest in nanoscale research stems from the belief that the technology represents the "invention of a method of inventing" (Darby and Zucker, 2003, 2) or an "enabling technology" (Rothaermel and Thursby, 2007, 834). Many argue that just as biotechnology revolutionized drug discovery, nanotechnology will reshape material design.

Nanotechnology is a strategic area in which to examine the hypotheses laid out above. Substantively, nanotechnology is a scientific and technologically diverse area that requires expert knowledge of domains spanning physics, molecular biology, and electrical engineering, among others (Porter and Youtie, 2009). This interdisciplinary character is notable for two reasons. First, because few individuals can assemble the expertise in all necessary fields, collaboration is important. Second, spillovers from proximate sources are relevant in nanotechnology because they enable access to expertise that firms lack internally.

Nanotechnology is also a valuable research setting because, unlike in fields that rely on other forms of intellectual property protection, in nanotechnology patenting is commonplace (Bawa, 2007; Lemley, 2005; Zucker et al., 2007). In reviewing evi-

dence showing the strong propensity among firms to seek patent protection for their discoveries, Bawa (2007, 719) explains, "Because development of nanotech-related technologies is extremely research intensive, without the market exclusivity offered by a patent, development of these products and their commercial viability in the marketplace would be significantly hampered." Even small start-ups "are willing to risk a larger part of their budgets to acquire, exercise, and defend patents" (Bawa et al., 2006, 29-5). This widespread use of patenting suggests that firms' patent portfolios should offer reliable insights into their innovative activities. Nanotechnology is also attractive because participants in the field are located in diverse geographic contexts, a condition that is necessary for testing the hypotheses.

2.4 Data and Methods

I collected longitudinal data on 454 firms that were engaged in nanotechnology R&D between 1990 and 2004. Data on collaboration networks come from the co-inventorship ties that are formed when two or more of a firm's employees work on a patent together. These data have a bipartite structure, in that they contain two types of nodes: actors (inventors) and events (patents). For the purposes of analysis, I created a unipartite projection of the bipartite network so that inventors were directly connected.

Patents have been widely used as an indicator of innovation and reflection of interpersonal ties (Fleming et al., 2007; Guler and Nerkar, 2012; Lahiri, 2010; Nerkar and Paruchuri, 2005; Wuchty et al., 2007). By law, a patent can only be granted if it describes an invention that (a) is *useful*, meaning that it could be commercially valuable, (b) is *novel*, in that it was unknown before its invention by the applicant for the patent, and (c) would be *nonobvious* to an individual with relevant expertise. Patent applications are evaluated by examiners who ensure these criteria are met.

Unlike scientific publications, for which individuals can be listed (or excluded)

as authors relatively independently of their contributions (Katz and Martin, 1997), patents that inaccurately list inventors may be rendered unenforceable (Sheiness and Canady, 2006). Moreover, patents are valuable because they provide longitudinal data on collaboration patterns and allow for the construction of firm-level structural measures of these networks, which are necessary for testing the hypotheses. Collecting similar data using survey methods would be extremely challenging. Of course, archival data like patents do have limitations. Importantly, they do not capture collaborations that leave no paper trail, nor do they account for informal contributions like feedback from colleagues. Despite these drawbacks, interviews with inventors suggest that patents provide a good reflection of their technological collaborations (Fleming et al., 2007).

I utilized a variety of directories and news sources to identify firms involved with nanotechnology. Major sources include company directories found in the *Lux Nanotech Report* (2001, 2004, 2006, 2008), the *Nanotechnology Opportunity Report* (2002), the *NanoVIP Database* (2005), *Understanding Nano*, and *BioScan*. I relied on trade publications to identify firms active early in the study period. U.S. subsidiaries of foreign firms were excluded because much of the knowledge used by these organizations comes from their parents (Gomes-Casseres et al., 2006). I also excluded wholesale suppliers and firms that did not perform R&D. Following prior work, I treated parents and subsidiaries as single units (Lahiri, 2010).

The hypotheses predict that innovation is in part a function of the interdependent relationship between the structure of collaborations among a firm's inventors and the geographic context in which those collaborations take place. Firms that perform R&D in multiple locations complicate this framework. For example, suppose a firm does most of its R&D in San Jose but also has a small satellite facility in a more remote setting like Boise. How should the geographic context of collaborations be measured for this firm? One possibility is to use an establishment level of analysis,

wherein satellites of multilocal firms are treated as distinct analytic units. Although this approach simplifies the measurement of geographic context, it is problematic for theoretical reasons. R&D subsidiaries serve diverse functions. In some cases they act as full-fledged research facilities, but often they are designed be listening posts that monitor developments in distant locales (Gassmann and von Zedtwitz, 1999). Either way, their performance may be driven by factors that differ from those driving a firm's core R&D facility.

Given these considerations, I excluded satellite R&D facilities and selected a "main research facility" for use in the analysis if a firm performed R&D in multiple locations. Note that the distinction between main research facilities and satellite locations is not artificial and is used by many firms in descriptions of their operations. For example, MEMC Electronic Materials' website explains that the company's "St. Peters [MO] plant serves as the corporate world headquarters for MEMC. In addition, it serves as the research and development headquarters because of its skilled workforce."[3] Similarly, in a 2005 10-K filing from Cabot Microelectronics states "our principal U.S. facilities that we own consist of: a global headquarters and research and development facility in Aurora, Illinois, comprising approximately 200,000 square feet."[4]

For most firms, a main nanotechnology research location could be identified using narrative descriptions of their operations in publicly available data.[5] The main sources included annual Securities and Exchange Commission (SEC) filings for publicly traded corporations (particularly Item 2, "Properties," found in 10-Ks and lease agreements in filing appendixes), archives of company websites, historical press releases, articles from trade journals and local newspapers, and in some cases, direct contact.[6] This

[3] `http://www.memc.com/index.php?view=st-peters`, accessed April 18, 2012

[4] `http://www.sec.gov/Archives/edgar/data/1102934/000110293405000054/body.htm`

[5] In a small number of instances, a firm's main research and main nanotechnology research facility did not coincide, in which case I took the main nanotechnology research facility to be the unit of analysis.

[6] I was able to access webpages for 93 percent of sample companies—even many that long ago ceased operations—using the Internet Archive Wayback Machine (`http://www.archive.org/web/web.php`), which has records back to 1996.

data collection effort was an exhaustive, yearlong process involving a team consisting of the author and two experienced undergraduate research assistants. The team met regularly to discuss coding. I corroborated the final data using author addresses from scientific publications and government grants awarded to individuals associated with each firm. When descriptions were not available, I labeled the location with the greatest quantity of nanotechnology R&D outputs—measured in terms of patents and scientific publications—as the primary facility. To account for relocations, I updated the location of each main research facility annually.

After identifying a main research facility for each firm, I followed prior research and used inventor addresses to exclude patents that were not associated with the main research facility (Fleming et al., 2007; Jaffe and Trajtenberg, 2002). Patents that do not list at least one inventor with an address in the same region as a firm's main research facility are excluded from the calculation of the core network variables.[7] Unless noted, for purposes of assigning inventors to establishments and constructing measures, I operationalized regions as U.S. Core Based Statistical Areas (CBSAs). Defined by the U.S. Office of Management and Budget, CBSAs are "area[s] containing a recognized population nucleus and adjacent communities that have a high degree of [social and economic] integration with that nucleus" (Spotila, 2000, 82228).

To account for the possibility that geographic context influences multilocal firms in ways that differ from how it influences single-establishment organizations, I collected annually updated data on the U.S. locations of active satellite R&D facilities for each firm using an approach similar to that employed for the main research facilities described above. Note that these data differ from data in much of the prior work on geography and innovation in that they are not dependent on the availability of

[7]This approach to assigning inventors to intrafirm networks is directly analogous to that used in an array of prior research on inventor networks at the regional level (Fleming et al., 2007; Graf, 2011). Using this procedure, 75 percent of sample firm nanotechnology patents could be associated with a main research facility. Following prior work, the inferential models introduced below include controls for the possibility of nonlocal collaborators.

inventor addresses listed on patents for location information.

I selected the time period for the study, 1990 to 2004, for several reasons. First, although the advances that made nanotechnology possible occurred in the early and mid 1980s, instruments like the STM and AFM were expensive, and nanotechnology was not commercially viable for most businesses. This changed in the late 1980s and early 1990s as equipment prices fell and more firms entered the sector (Darby and Zucker, 2003). Second, though some nanotechnology patents were granted before the late 1980s, these can be hard to identify using keyword-based methods (discussed below) since in earlier years terminology was being developed. I chose 2004 to end the analysis since the version of the Patent Network Dataverse database (Lai et al., 2011) from which I draw my data on patents is current through 2008, and a three-year lag between patent application and issue dates is common (Jaffe and Trajtenberg, 2002, 409–410).

The following subsections present the variables. The dependent, independent, and most control variables were calculated using only data from the main research facility, according to the logic outlined above; however a handful of controls are firm-level constructs. For clarity, I note explicitly the level at which each is measured.

2.4.1 Dependent Variables

I modeled two dependent variables to assess a firm's innovation performance. First, I measured *impact* as the citation-weighted sum of nanotechnology patents applied for by firm i at times $t + 1$ and $t + 2$ (Kotha et al., 2011). To ensure that the outcome accurately matches the hypotheses, I counted only patents that were produced by an inventor associated with a firm's main research facility. As discussed above, patents are in general a good proxy for innovative activity; however, they can vary widely in their quality, and consequently raw patent counts may be a misleading indicator of performance. To account for heterogeneity in quality, I weighted each

patent by the number of citations it received from future patents in the first five years after issue—the window during which annual citations to most patents reach their peak (Jaffe and Trajtenberg, 2002).[8] Concretely, I define the citation-weighted patent count for firm i at time t as

$$CWP_{it} = \sum_{j \in n_{t+1}} \left(1 + \sum_{\tau=t+1}^{t+5} c_{j\tau}\right) + \sum_{j \in n_{t+2}} \left(1 + \sum_{\tau=t+2}^{t+6} c_{j\tau}\right), \tag{2.1}$$

where n_{t+1} and n_{t+2} are the sets of patents applied for by firm i at times $t+1$ and $t+2$, respectively, and $c_{j\tau}$ is a count of citations to patent j at time τ (Trajtenberg, 1990). The number of citations a patent receives is correlated with its economic value (Griliches, 1990; Hall et al., 2005).[9]

The second dependent variable measures *new combinations* as the sum of nanotechnology patents applied for by firm i at $t+1$ and $t+2$ that bridge previously uncombined technological domains. Incremental innovations build on existing combinations by offering minor improvements. Breakthroughs are characterized by novel combinations and have the potential to create new fields (Christensen, 1997). The U.S. Patent and Trademark Office (USPTO) organizes all inventions using a fine-grained system of roughly 100,000 subclasses (Fleming et al., 2007). This system is updated regularly as science and technology evolve. With each update, all patents dating to the USPTO's (1790) founding are revised to ensure uniform classification. To identify new combinations, I counted the times a particular set of subclasses used by the USPTO to classify a given patent had been used previously between its issue date and 1790. Patents that were first to fall into a particular combination of subclasses were given a score of 1, while those that were classified using an existing combination received a score of 0.[10] Note that impact is determined *ex post* as the

[8] The results are robust to alternative windows and to the use of lifetime citation counts.

[9] Although the general patent data set used in this study ends in 2008, I have citation data through 2011, which ensures that all patents have a full window within which to accumulate citations.

[10] For more detail on this type of measure, see Fleming et al. (2007, 474–475). To ob-

patent is used, while new combinations are determined *ex ante* during the application process.

To identify nanotechnology patents, I searched the full-text of all patents granted by the USPTO. I identified nanotechnology patents as those containing at least one of 29 keyword and wild-card terms (e.g., "atomic force microscope," "biomotor," "quantum dot") identified by subject specialists as reliable indicators of the domain. I dropped patents containing only noise terms that could have been picked up in the keyword search but did not involve nanotechnology (e.g., patents containing "nanosecond" were excluded if no additional "nano-keyword" warranted inclusion). An array of prior work on nanotechnology has employed this methodology (Rothaermel and Thursby, 2007; Zucker et al., 2007).

2.4.2 Independent Variables

Firm proximity. The first hypothesis predicts that nanotechnology firms will be more innovative if they are near other companies that perform related R&D. I computed proximity as

$$FP_{it} = \sum_{j \neq i} \frac{x_j}{(1 + d_{ij})}, \tag{2.2}$$

where x_j is a weight, d_{ij} is the distance between firm i and firm j, t is an index for time, and j is an index for all firms other than i (Sorenson and Audia, 2000). The measure effectively represents the average distance between firm i and all other firms at time t. Proximity may be less important if nearby firms have little knowledge to share. To control for this possibility, I set the weighting parameter, x_j, equal to the logged number of nanotechnology patents awarded to firm j at time t. I computed the distance parameter, d_{ij}, by obtaining the latitude and longitude of each firm at time t based on the center of the zip code in which they were located at time t. I

tain patent classifications back to 1790, I relied on the 2011 U.S. Patent Grant Master Classification File, available at `http://commondatastorage.googleapis.com/patents/patent_classification_information/mcfpat.zip`.

then calculated d_{ij} in Euclidean distance, following Sorenson and Audia (2000), as

$$d_{ij} = \alpha \left\{ \arccos \left[\sin \left(lat_i \right) \sin \left(lat_j \right) + \cos \left(lat_i \right) \cos \left(lat_j \right) \cos \left(\left| long_i - long_j \right| \right) \right] \right\}, \quad (2.3)$$

where d_{ij} is the distance between points i and j, α is a constant, set to 343.78, which gives the result in units of 10 miles, and latitude (lat) and longitude ($long$) are measured in radians.

Inventor networks—Cohesion. I measured cohesion as the overall level of clustering in a firm's network of inventors. Clustering captures the extent to which inventors' collaborators also collaborate with one another—a hallmark of cohesive groups (Coleman, 1988; Newman et al., 2001). The measure, known as the clustering coefficient, was calculated for each firm i at time t as

$$CC_{it} = \frac{3N_{\triangle}}{N_{\vee}} = \frac{3 \times (\text{number of triangles})}{(\text{number of connected triples})}, \quad (2.4)$$

where a triangle is a closed triad and a triple is an open triad. The coefficient ranges from 0 to 1. Larger values signal higher levels of cohesion.

Recall that the collaborative ties analyzed here are derived from a unipartite projection of a bipartite network. As a result, some of the observed clustering may be artificial: all inventors listed on a patent are automatically clustered through the process of projection. To account for this artificial clustering, I followed the approach suggested by Newman et al. (2001) and scale the clustering coefficient in the observed network (CC_{it}^{O}) by clustering coefficient for the unipartite projection of a simulated random network (CC_{it}^{R}) with an identical bipartite degree distribution. Values at or below 1 imply that any clustering in an observed network is largely an artifact of the projection process. Values greater than 1 indicate that clustering in the observed network results from collaborations that span teams (i.e., the set of inventors who work on a patent together) and consequently a higher level of cohesiveness than

expected by chance (Uzzi and Spiro, 2005). Following prior work, I dropped ties after five years. Only collaborations involving an inventor from a firm's main research facility were included in the measure.

Inventor networks—Inefficiency. Hypothesis 3 predicts that more inefficient inventor networks will lead to better performance for firms that are less proximate to others in their field. I measured inefficiency as the average length of the paths in each firm i's network of nanotechnology inventors at time t. Networks with longer characteristic path lengths have more intermediaries separating nodes. Information should spread more slowly and less accurately as path length increases. Because paths across disconnected components are undefined, I considered only those within connected components when computing the measure (Gulati et al., 2012, 458).[11] As described below, all models include controls for number of components. I also performed robustness tests using an alternative proxy for inefficiency; these provided similar results. Because the unipartite projection process can result in artificially short path lengths, I scaled this measure relative to the expected path length of a random network (Newman et al., 2001). I dropped ties after five years.

2.4.3 Control Variables

Publicly traded. I controlled for whether a firm was publicly or privately held at time t. This measure varies as firms make initial public offerings (IPOs) or transition to private holding. Public firms may have more resources than private ones, which could influence performance.

Research sites. Some firms perform a portion of their R&D in satellite facilities, which may enable inventors to source knowledge from multiple locales (Lahiri, 2010). To account for the possibility that firms with distributed R&D are influenced differ-

[11] I also tested an alternative measure based on closeness centrality that sets path lengths between nodes in different components to 0 (Opsahl et al., 2010). The results were similar. I thank an anonymous reviewer for this suggestion.

ently by the region of their main research facility, I controlled for the total number of U.S. research sites operated by each firm at time t. The number varies as firms open and close satellites.

Global geographic distribution. Firms also differ in terms of the global geographic distribution of their R&D. I was unable to find reliable data on international R&D satellites for many private companies. Given these constraints, I created for each firm a measure of international geographic R&D dispersion using the inventor addresses listed on all of the firm's patents, in all technology areas. I constructed a Herfindahl index based on the distribution of inventors across countries, calculated as $\sum c_i^2$, where c_i is the proportion of inventors in country i. I subtracted the result from 1 so that firms with more globally distributed R&D operations have higher values on this measure. Inventors were dropped after five years of no activity.

Technological diversity. Some sample firms also perform R&D in areas other than nanotechnology. Firms active in multiple fields could have greater exposure to diverse knowledge, which may enhance their innovation. To account for this possibility, I collected data on all patents (not just those in nanotechnology) granted to each sample firm. Using a five-year window, I measured technological diversity as the Herfindahl index of the primary classes of the firm's patents, subtracting the result from 1 so that firms with diverse portfolios have higher values.

Nonnanotechnology patent stock. A related alternative explanation for predicted outcomes is that neither networks nor proximity are important if a firm has a large technological base outside of nanotechnology on which inventors can draw for new ideas. To control for this possibility, I included a measure of the size of each firm's nonnanotechnology patent portfolio. Given the pace of change in nanotechnology, I constructed this variable using a depreciated stock model in which older patents are

worth less than recent discoveries. Formally,

$$NS_{it} = \sum_{\tau=0}^{t} \Big(1 - \delta\Big)^{t-\tau} K_{i\tau}, \tag{2.5}$$

where K is the set of patents applied for by firm i at time τ and δ is a constant, set to 0.15, that imposes a 15 percent annual depreciation (Hall et al., 2005). The measure includes all patents, regardless of geographic origin. I logged the variable to account for diminishing returns to large stocks.

Main R&D facility and HQ separate. Using the data on establishment locations and their movements described above, I controlled for whether a firm's main research facility and corporate headquarters were located in the same CBSA at time t. The separation of management and R&D (sometimes called "skunk works") may lead to greater exploration in product development.

Recent relocation. Some sample firms relocated their main research facilities during the study period. Although these moves are reflected through annual changes in the value of the firm proximity variable, I also include an indicator for whether a firm relocated its main research facility at time t. Long-distance moves in particular could disrupt projects, collaborative relationships, and ultimately innovation.

California. Researchers have long noted the unique success of high-technology clusters in California's Silicon Valley and San Diego regions (Saxenian, 1994). California also generally invalidates noncompete agreements, which helps enhance knowledge flows between firms. To account for these unique state characteristics, I controlled for whether or not a firm's main research facility was located in California at time t.

Local university ties. Existing theories point to the importance of local embeddedness for explaining how geography affects innovation. Local connections offer one source from which firms can acquire relatively exclusive new knowledge. Such connectivity also signals legitimacy to neighboring organizations and in so doing creates

opportunities for exchange (Owen-Smith and Powell, 2004). Given the importance of science for nanotechnology R&D, I controlled for the number of collaborations a firm had with local universities at time t. I defined collaborations as coauthored nanotechnology publications whose authors' affiliations include a focal firm's main research facility and a university in the same CBSA. Nanotechnology publication data were obtained from Scopus using a search strategy analogous to the one used for patents.

Local inventor hires. I also sought to capture each firm's embeddedness in local labor markets. To do so, I controlled for the number of new inventors at time t that joined a focal firm's main research facility after having previously patented with another organization in the same CBSA. Patenting serves as a proxy for employment. Because the measure only captures inventors with prior patents, it is a conservative estimate of local hiring.

Inventor geographic distribution. I assigned all inventors listed on nanotechnology patents applied for by employees at a firm's main research facility to a CBSA, using a five-year window, and then constructed a Herfindahl index to capture the spread of inventors across U.S. communities. After the result is subtracted from 1, higher values indicate more dispersed R&D activities. This control helps account for the alternative explanation that performance differences result from the presence or absence of distant ties.

Distant inventors. This control counts the number of inventors who are included in a main research facility's collaboration network but who resided in an external CBSA at time t. The variable helps to account for the geographic composition of a network while also offering an additional way of assessing the possibility that distant, external knowledge sources explain the varying benefits of proximity to performance, as emphasized in prior theory.

Inventor career experience. I controlled for the total number of distinct orga-

nizations, as of time t, with which active inventors at a firm's main research facility patented before joining a focal firm. This variable helps to control for spillovers that result from labor mobility (possibly over long distances) or strategic recruitment (Agrawal et al., 2006; Rosenkopf and Almeida, 2003).

High-mobility inventors. What may matter more than the aggregate career experience of a firm's R&D team is whether it can attract even a small number of talented individuals. To account for this, I controlled for the number of high-mobility inventors at a firm's main research facility at time t. High-mobility inventors are those with a value two standard deviations above the mean on the number of prior organizations patented with.

Inventor technological experience. I also controlled for the median number of career nanotechnology patents granted to inventors at a firm's main research facility as of time t. This variable helps account for the possibility that the performance of some firms is simply due to their ability to recruit successful inventors.

Median team size. I controlled for team size, measured as the median number of inventors listed on each nanotechnology patent awarded to a firm's main research facility over the past five years. Team size has been shown to be related to performance in knowledge production (Wuchty et al., 2007). This variable also helps to control for the alternative explanation that team size or a culture of participation (and not structure) drives innovation.

Components. As discussed above, path lengths are undefined for inventors who cannot be connected through any intermediaries. To account for differences among networks with varying number of undefined paths, I controlled for the number of components present in each firm's network, using a five-year window.

Inventors. The number of active inventors in a firm is likely related to the firm's innovative potential—the more employees in R&D, the more chances for discovery. Inventors were dropped after five years if they had not been awarded any new patents

at their firm's main research facility.

Nanotechnology patent stock. This variable controls for nanotechnology patents awarded to a main research facility. As with nonnanotechnology patents, I used a depreciated stock model to account for the fact that older patents are likely less valuable than newer ones. The control captures unobserved heterogeneity in innovative capabilities (Blundell et al., 1995). Patent stocks help control for R&D spending (Griliches, 1990).

Nanotechnology portfolio complexity. Prior research suggests that proximity is advantageous for sharing complex knowledge. Proximity may not be important for firms with simpler technologies, because they can acquire knowledge from sources like technical publications. I therefore controlled for the median complexity of each firm's nanotechnology patent portfolio using Sorenson et al.'s (2006) measure of interdependence. This metric defines complexity as a function of "the historical difficulty of recombining the elements that constitute" a technology (Sorenson et al., 2006, 1002). Higher values indicate more complex portfolios.

Table 2.1: Variable Names and Definitions

Name	Definition	Panel Structure
Dependent Variables		
Patent impact$_{(t+1 \text{ and } t+2)}$	Count of citations to nanotechnology patents applied for by inventors at the main research facility at times $t+1$ and $t+2$ for first 5 years after issue	Updated annually as inventors apply for patents
New combinations$_{(t+1 \text{ and } t+2)}$	Count of nanotechnology patents applied for by inventors at the main research at times $t+1$ and $t+2$ that brought together a previously uncombined set of technology subclasses	Updated annually as inventors apply for patents
Independent Variables		
Firm proximity	Local density of nanotechnology firms at time t, weighted by logged nanotechnology patent counts; region based on location of main research facility	Updated annually as facilities relocate and as other sample firms open, close, relocate, and apply for patents
Cohesion	Extent to which inventors' collaborators collaborate with one another; emphasizes cohesion	Updated annually as inventors collaborate on new projects and ties older than 5 years are dropped
Inefficiency	Harmonic mean path length separating inventors; emphasizes inefficient networks	Updated annually as inventors collaborate on new projects and ties older than 5 years are dropped
Cohesion × firm proximity	Interaction of cohesion and firm proximity (weighted)	
Inefficiency × firm proximity	Interaction of inefficiency and firm proximity (weighted)	
Covariates—Firm		
Publicly traded	Dummy variable; 1 = public company	Updated annually if the firm makes an IPO or transitions to private holding
Research sites	Count of active U.S. R&D facilities in any technology area	Updated annually if the firm opens or closes research facilities
Global geographic distribution	Herfindahl of inventor countries for all firm patents over past 5 years	Updated annually as inventors apply for patents and those older than 5 years are dropped
Technological diversity	Herfindahl of primary classes for all patents filed by firm over past 5 years	Updated annually as inventors apply for patents and those older than 5 years are dropped
Non-nanotechnology patent stock (log)	Cumulative non-nanotechnology patents awarded to the firm, depreciated annually by 15%	Updated annually as inventors apply for non-nanotechnology patents
Covariates—Main Research Facility		
Main research and HQ separate	Dummy variable; 1 = research facility and corporate headquarters are not co-located	Updated annually as facilities relocate

Table 2.1 (Continued)

Recent relocation	Dummy variable; 1 = research facility relocated at time t	Updated annually as facilities relocate
California	Dummy variable; 1 = research facility is located in California	Updated annually as facilities relocate
Local university ties	Count of ties to universities in local CBSA based on nanotechnology paper collaborations	Updated annually as new collaborations form and ties older than 1 year are dropped
Local inventor hires		Updated annually as new inventors join the firm
Inventor geographic distribution	Herfindahl of inventor CBSAs for research facility nanotechnology patents over past 5 years	Updated annually as inventors apply for patents and those older than 5 years are dropped
Distant inventors	Count of inventors residing outside local CBSA	Updated annually as inventors apply for patents and those older than 5 years are dropped
Inventor career experience	Count of distinct organizations with which inventors patented before joining the firm	Update annually as new inventors join the firm
High mobility inventors	Count of inventors 2 SD or more above the mean on career experience	Update annually as new inventors join the firm inventors join the firm
Inventor technological experience	Median number of career nanotechnology patents per inventor at main research facility	Updated annually as inventors apply for patents and new
Median team size	Median number of inventors per nanotechnology patent over past five years	Updated annually as inventors apply for patents and those older than 5 years are dropped
Inventors (log)	Number of inventors in the network	Updated annually as inventors apply for patents and those older than 5 years are dropped
Components	Count of discrete subsets of collaborating inventors who are disconnected from all other subsets	Updated annually as inventors apply for patents and those older than 5 years are dropped
Nanotechnology patent stock	Cumulative nanotechnology patents, depreciated annually by 15%	Updated annually as inventors apply for nanotechnology patents
Nanotechnology portfolio complexity	Median historical difficulty of combining subclasses represented among patents in portfolio	Updated annually as inventors apply for nanotechnology patents
Nanotechnology paper stock	Cumulative nanotechnology papers, depreciated annually by 15%	Updated annually as employees publish nanotechnology papers
Other Variables		
Constraint	Alternative proxy for inefficient networks, higher constraint implies fewer structural holes and less inefficient networks; used in robustness checks	Updated annually as inventors collaborate on new projects and ties older than 5 years are dropped

Nanotechnology paper stock. Finally, firms might differ in the extent to which they are embedded in larger scientific communities (Gittelman, 2007). Firms that are active in science might depend less on local sources for new knowledge, because scientific networks are geographically dispersed. Embeddedness might also serve as a source of relatively more exclusive knowledge, because not all firms have such access. To account for different levels of engagement, I controlled for the stock of scientific nanotechnology publications produced by individuals at each firm's main research facility. As with patents, I used a depreciated stock model in which the contribution of a publication to the measure decreases 15 percent annually.

Table 2.2: Descriptive Statistics and Correlations[†]

	Variable	Mean	SD			1	2	3	4	5	6	7	8	9	10	11	12	13
			Overall	Between	Within													
1.	Patent impact$_{(t+1 \text{ and } t+2)}$	62.92	157.18	98.78	96.62	1.00												
2.	New combinations$_{(t+1 \text{ and } t+2)}$	9.37	21.15	13.52	12.13	0.90	1.00											
3.	Publicly traded	0.64	0.48	0.48	0.13	0.15	0.21	1.00										
4.	Research sites	3.02	3.91	3.26	0.91	0.28	0.37	0.32	1.00									
5.	Global geographic distribution	0.09	0.11	0.10	0.05	0.09	0.16	0.26	0.27	1.00								
6.	Technological diversity	0.72	0.24	0.23	0.10	0.21	0.26	0.35	0.38	0.21	1.00							
7.	Non-nanotechnology patent stock (log)	3.71	2.58	2.43	0.49	0.39	0.47	0.61	0.64	0.37	0.59	1.00						
8.	Main research and HQ separate	0.12	0.32	0.27	0.10	0.05	0.10	0.17	0.43	0.21	0.19	0.30	1.00					
9.	Recent relocation	0.03	0.16	0.09	0.15	−0.03	−0.05	−0.09	−0.04	−0.06	−0.06	−0.11	−0.01	1.00				
10.	California	0.33	0.47	0.47	0.06	0.01	−0.03	−0.02	−0.14	−0.07	−0.13	−0.11	−0.17	0.06	1.00			
11.	Local university ties	0.37	1.40	0.87	0.99	0.18	0.26	0.09	0.29	0.15	0.16	0.26	0.07	−0.03	−0.08	1.00		
12.	Inventor geographic distribution	0.31	0.22	0.21	0.09	−0.04	−0.04	0.02	0.10	0.07	0.13	0.04	0.11	0.05	0.23	−0.04	1.00	
13.	Distant inventors	8.63	21.51	13.38	12.41	0.34	0.48	0.20	0.53	0.26	0.23	0.44	0.26	−0.05	−0.05	0.28	0.21	1.00
14.	Local inventor hires	1.00	2.23	1.75	1.65	0.39	0.41	0.06	0.22	0.09	0.14	0.22	0.05	−0.03	0.07	0.22	0.04	0.25
15.	Inventor career experience	53.51	98.18	62.90	71.66	0.27	0.37	0.12	0.21	0.13	0.16	0.29	0.07	−0.05	0.07	0.32	0.11	0.54
16.	High mobility inventors	1.84	3.14	2.33	2.13	0.18	0.27	0.12	0.23	0.15	0.17	0.26	0.11	−0.04	0.11	0.31	0.21	0.54
17.	Inventor technological experience	2.43	1.94	1.62	1.16	−0.06	−0.07	−0.17	−0.16	−0.10	−0.12	−0.23	−0.01	0.04	−0.01	−0.05	−0.01	−0.08
18.	Median team size	3.16	1.36	1.14	0.83	−0.10	−0.10	−0.07	−0.11	−0.02	−0.19	−0.15	−0.01	0.02	0.03	−0.03	0.19	0.07
19.	Inventors (log)	2.97	1.16	0.93	0.51	0.46	0.59	0.36	0.51	0.32	0.45	0.69	0.16	−0.09	−0.06	0.38	0.12	0.61
20.	Components	5.36	8.23	6.27	2.84	0.55	0.65	0.32	0.63	0.28	0.37	0.66	0.17	−0.07	−0.14	0.49	0.01	0.60
21.	Nanotechnology patent stock	26.93	55.08	35.02	31.06	0.56	0.73	0.24	0.39	0.22	0.26	0.52	0.10	−0.06	−0.06	0.41	−0.05	0.64
22.	Nanotechnology portfolio complexity	2.32	1.08	1.04	0.58	−0.09	−0.10	−0.05	0.01	−0.01	0.00	0.00	0.04	0.03	0.04	0.05	0.01	−0.04
23.	Nanotechnology paper stock	11.77	36.48	27.27	19.97	0.18	0.28	0.15	0.36	0.20	0.19	0.34	0.08	−0.04	−0.07	0.74	0.02	0.45
24.	Firm proximity	17.15	16.44	16.08	7.72	0.09	0.07	−0.04	−0.04	0.02	−0.01	−0.04	−0.11	0.07	0.59	−0.01	0.29	0.08
25.	Cohesion	1.32	0.39	0.28	0.26	0.29	0.37	0.27	0.35	0.22	0.29	0.50	0.14	−0.05	−0.06	0.25	0.07	0.37
26.	Inefficiency	0.87	0.29	0.20	0.20	0.15	0.22	−0.17	−0.15	−0.07	−0.11	−0.17	−0.04	0.02	0.03	0.14	0.00	0.26
27.	Constraint	0.70	0.19	0.16	0.12	−0.11	−0.14	0.00	−0.02	−0.07	−0.14	−0.06	−0.02	0.01	−0.01	−0.09	−0.16	−0.22

	Variable	14	15	16	17	18	19	20	21	22	23	24	25	26	27
14.	Local inventor hires	1.00													
15.	Inventor career experience	0.30	1.00												
16.	High mobility inventors	0.37	0.78	1.00											
17.	Inventor technological experience	−0.04	0.13	0.03	1.00										
18.	Median team size	0.00	0.10	0.10	0.11	1.00									
19.	Inventors (log)	0.34	0.57	0.55	−0.15	0.02	1.00								
20.	Components	0.35	0.39	0.39	−0.18	−0.10	0.79	1.00							
21.	Nanotechnology patent stock	0.28	0.64	0.48	−0.03	−0.05	0.73	0.71	1.00						
22.	Nanotechnology portfolio complexity	−0.02	0.00	0.05	0.01	−0.02	−0.02	−0.03	−0.04	1.00					
23.	Nanotechnology paper stock	0.22	0.45	0.42	−0.07	0.00	0.48	0.56	0.47	0.05	1.00				
24.	Firm proximity	0.20	0.28	0.33	0.07	0.03	0.10	−0.03	0.07	0.09	0.02	1.00			
25.	Cohesion	0.21	0.30	0.29	−0.14	−0.11	0.57	0.55	0.45	−0.02	0.30	0.03	1.00		
26.	Inefficiency	0.07	0.44	0.34	0.19	0.28	0.23	0.01	0.36	−0.02	0.16	0.16	−0.06	1.00	
27.	Constraint	−0.10	−0.39	−0.32	−0.12	−0.36	−0.47	−0.07	−0.25	−0.02	−0.13	−0.15	−0.07	−0.55	1.00

[†] $N = 2,760$

Inverse Mills ratio (λ). Some measures can only be constructed for networks of a minimal size, which raises the possibility of selection bias. To address this possibility, I tested the hypotheses using firms with networks consisting of at least three nodes and two ties, as is required to measure clustering.[12] This criterion helps shift selection to the independent variables. Some firms had more than 15 patents, but all were filed by lone inventors—these firms did not use networks. In short, both innovative and noninnovative firms can be excluded.

To examine any remaining bias, I estimated a probit model predicting membership in the sample, from which I calculated the inverse Mills ratio to include as a covariate. The first-stage equation includes the main (nonnetwork) controls listed above but adds *de novo* entry and firm age as exclusion restrictions. Founded explicitly to develop new nanotechnologies, *de novo* firms may be less likely to organize research around teams because they are often small start-ups that spend their first years building and diversifying their R&D operations. Age may also predict the use of R&D teams. Recent studies document dramatic growth in team-based knowledge production (Wuchty et al., 2007). To the extent that the organization of R&D is path dependent, much older firms may be slow to adopt team-based approaches in nanotechnology (Porter and Youtie, 2009). The proportion of correct predictions for the probit model was 0.76.

Tables 2.1 and 2.2 provide variable summaries and descriptive statistics, respectively. After entry and exit of firms over the observation period are accounted for, the panel consists of 2,760 firm-year observations. Variance inflation factors were well within acceptable ranges (i.e., none larger than 10).

[12] In unreported analyses, I set the clustering coefficient to 0 if the measure was undefined, which eliminated the need for a selection model. The results were similar. I thank an anonymous reviewer for this suggestion.

2.5 Model Estimation

Both dependent variables are counts and take on only nonnegative integer values. Given this distribution, I estimated conditional fixed-effects quasi-maximum-likelihood Poisson models with conditional means of the form

$$\mathbb{E}[y_{it}|\alpha_i, \mathbf{x}_{it}] = \alpha_i \exp(\mathbf{x}'_{it}\beta),\ i = 1, \ldots, N,\ t = 1, \ldots, T_i, \tag{2.6}$$

where y_{it} is the dependent variable for firm i at time t, $\mathbf{x}_{it}$ are the independent and control variables, β are the coefficients to be estimated, and α_i are time-invariant, unit (firm) specific effects. Fixed-effects models are advantageous in that they control for all time-invariant unobserved heterogeneity by relying on within-unit variation. The models allow for correlation between the unit specific (time-invariant) intercepts (α_i) and the independent variables (x_{it}), but not for correlation between the independent variables and the (time-varying) idiosyncratic error term (ϵ_{it}). Thus, they relax the assumption of random effects, but maintain the assumption of strict exogeneity, i.e. $\mathbb{E}(\epsilon_{it}|x_{i1}, \ldots, x_{iT}, \alpha_i) = 0$, $t = 1, 2, \ldots, T$. Note that the fixed-effects approach models changes within units over time and does not exploit cross-sectional variability.

Although negative binomial models are common in research on innovation, a Poisson approach has several features that make it attractive for the analysis of panel count data. Most importantly, Poisson models rely on weaker distributional assumptions than negative binomial methods and provide consistent estimates as long as the conditional mean is correctly specified (Gourieroux et al., 1984). Further, quasi-maximum-likelihood Poisson standard errors are robust to overdispersion, which occurs when the conditional variance of an outcome variable is greater than the conditional mean.

Despite their attractions, fixed-effects models' exclusive reliance on within-unit variance has drawbacks. For example, the models may produce incorrect point esti-

mates and inflated standard errors for variables that exhibit relatively little change within units. I therefore also present random-effects models for each dependent variable. Random effects was attractive because it takes advantage of between-unit variation but allows for different intercepts, provisions more realistic here than what pooled models would allow, given the diversity of sample firms.

2.6 Results

Tables 2.3 and 2.4 present models of nanotechnology patent impact and new combinations, respectively. Estimates in Models 1-7 (Table 2.3) and 9-15 (Table 2.4) are derived from a conditional fixed-effects quasi-maximum-likelihood specification. Models 8 (Table 2.3) and 16 (Table 2.4) present estimates derived from a random-effects Poisson specification with bootstrap estimates of the standard errors to account for overdispersion and serial correlation. The results offer substantive support for the findings obtained with the fixed-effects models.

Table 2.3: Models of Impact†

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Controls—Firm								
Publicly traded	−0.5671**	−0.6091**	−0.5864**	−0.5450**	−0.6074**	−0.5657**	−0.5554**	−0.5747**
	(0.2526)	(0.2463)	(0.2414)	(0.2319)	(0.2381)	(0.2278)	(0.2366)	(0.2262)
Research sites	−0.0192	−0.0060	−0.0094	0.0061	−0.0170	−0.0004	−0.0007	−0.0010
	(0.0402)	(0.0330)	(0.0322)	(0.0304)	(0.0323)	(0.0304)	(0.0313)	(0.0376)
Global geographic distribution	−2.1978***	−2.1260**	−1.9839**	−2.0006**	−2.0970**	−2.1262***	−1.9490**	−2.1214**
	(0.8220)	(0.8485)	(0.8431)	(0.8257)	(0.8420)	(0.8192)	(0.8662)	(0.9079)
Technological diversity	−0.4148	−0.3211	−0.3266	−0.2452	−0.3036	−0.2169	−0.5145	−0.2046
	(0.4853)	(0.4662)	(0.4549)	(0.4539)	(0.4522)	(0.4502)	(0.4241)	(0.4584)
Non-nanotechnology patent stock (log)	0.3642***	0.3197***	0.3105***	0.3112***	0.3097***	0.3111***	0.2862***	0.3072**
	(0.1187)	(0.1133)	(0.1107)	(0.1092)	(0.1069)	(0.1049)	(0.1086)	(0.1210)
Controls—Main Research Facility								
Main research and HQ separate	0.2241	0.2560	0.2750	0.2411	0.3315	0.2980	0.3427	0.2651
	(0.4438)	(0.4341)	(0.4348)	(0.4166)	(0.4347)	(0.4149)	(0.3982)	(0.5255)
Recent relocation	−0.0992	−0.0756	−0.0818	−0.1047	−0.0629	−0.0860	0.0686	−0.0833
	(0.2153)	(0.2260)	(0.2221)	(0.2176)	(0.2245)	(0.2196)	(0.1906)	(0.2163)
California	−0.1114	−0.1908	−0.2107	−0.2504	−0.1938	−0.2374	−0.0649	−0.1895
	(0.2561)	(0.2615)	(0.2502)	(0.2515)	(0.2549)	(0.2574)	(0.2436)	(0.3392)
Local university ties	−0.0104	−0.0098	−0.0111	−0.0133	−0.0080	−0.0097	−0.0019	−0.0097
	(0.0156)	(0.0153)	(0.0151)	(0.0148)	(0.0147)	(0.0143)	(0.0133)	(0.0151)
Local inventor hires	0.0129*	0.0051	0.0035	0.0011	0.0038	0.0014	0.0017	0.0016
	(0.0072)	(0.0084)	(0.0083)	(0.0085)	(0.0077)	(0.0078)	(0.0084)	(0.0099)
Inventor geographic distribution	0.4159	0.5091	0.6056	0.6882*	0.5457	0.6345*	0.5646	0.6244
	(0.3838)	(0.3858)	(0.3944)	(0.3930)	(0.3675)	(0.3648)	(0.3588)	(0.3829)
Distant inventors	0.0006	0.0010	0.0006	0.0005	0.0004	0.0004	0.0007	0.0004
	(0.0016)	(0.0017)	(0.0016)	(0.0017)	(0.0016)	(0.0016)	(0.0018)	(0.0052)
Inventor career experience	−0.0011	−0.0016*	−0.0016*	−0.0018**	−0.0013	−0.0015*	−0.0015	−0.0015
	(0.0008)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0010)	(0.0011)
High mobility inventors	−0.0012	−0.0049	−0.0059	−0.0070	−0.0042	−0.0057	−0.0016	−0.0060
	(0.0242)	(0.0248)	(0.0237)	(0.0243)	(0.0236)	(0.0242)	(0.0269)	(0.0296)
Inventor technological experience	−0.0248	−0.0391	−0.0411	−0.0367	−0.0460	−0.0421	−0.0849***	−0.0405
	(0.0289)	(0.0287)	(0.0284)	(0.0291)	(0.0283)	(0.0291)	(0.0289)	(0.0328)
Median team size	−0.1335**	−0.1045*	−0.1158*	−0.1049*	−0.1194**	−0.1076*	−0.0760	−0.1093*
	(0.0599)	(0.0587)	(0.0597)	(0.0576)	(0.0589)	(0.0562)	(0.0567)	(0.0564)
Inventors (log)	0.3925**	0.3040*	0.2715	0.3064*	0.2657	0.3009*	−0.0194	0.3062
	(0.1939)	(0.1758)	(0.1729)	(0.1705)	(0.1732)	(0.1694)	(0.1427)	(0.2031)

Table 2.3 (Continued)

Components	0.0081	0.0053	0.0050	0.0027	0.0048	0.0022	0.0094	0.0021
	(0.0066)	(0.0072)	(0.0076)	(0.0081)	(0.0069)	(0.0073)	(0.0073)	(0.0088)
Nanotechnology patent stock	0.0018	0.0024**	0.0023*	0.0025**	0.0022*	0.0025**	0.0036***	0.0024
	(0.0012)	(0.0011)	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0013)	(0.0018)
Nanotechnology portfolio complexity	−0.0066	−0.0186	−0.0143	−0.0123	−0.0127	−0.0097	−0.0134	−0.0097
	(0.0531)	(0.0519)	(0.0530)	(0.0528)	(0.0517)	(0.0515)	(0.0528)	(0.0585)
Nanotechnology paper stock	−0.0017	−0.0011	−0.0014	−0.0013	−0.0016	−0.0016	−0.0015	−0.0015
	(0.0011)	(0.0012)	(0.0012)	(0.0012)	(0.0011)	(0.0011)	(0.0013)	(0.0027)
Independent Variables								
Firm proximity$_{(centered)}$		0.0133*	0.0134**	0.0094	0.0171***	0.0134**	0.0167***	0.0134
		(0.0068)	(0.0066)	(0.0070)	(0.0065)	(0.0067)	(0.0061)	(0.0094)
Cohesion$_{(centered)}$			0.0856	0.0452	0.0850	0.0400	0.0075	0.0402
			(0.1034)	(0.1022)	(0.0997)	(0.0981)	(0.1002)	(0.0966)
Inefficiency$_{(centered)}$			0.2033*	0.1636	0.2510**	0.2138*	0.2444**	0.2149*
			(0.1181)	(0.1195)	(0.1181)	(0.1155)	(0.1188)	(0.1253)
Cohesion × firm proximity				0.0117***		0.0126***	0.0118***	0.0125***
				(0.0041)		(0.0038)	(0.0038)	(0.0046)
Inefficiency × firm proximity					−0.0107**	−0.0118**	−0.0133***	−0.0119*
					(0.0053)	(0.0051)	(0.0051)	(0.0068)
Other								
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mills ratio (λ)	1.5722***	1.3988***	1.3913***	1.4354***	1.3356***	1.3726***		1.3710***
	(0.3928)	(0.4185)	(0.4130)	(0.4082)	(0.4140)	(0.4093)		(0.4253)
Constant								0.8424***
								(0.1054)
N	2569	2569	2569	2569	2569	2569	2569	2760
Firms in model	317	317	317	317	317	317	317	376
Log likelihood	−27170.10	−26744.87	−26593.71	−26339.41	−26403.72	−26108.13	−26724.61	−28256.12
Model d.f.	37	38	40	41	41	42	41	42

* $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$; two tailed tests.

† The estimates presented in Models 1-7 are derived from a conditional fixed effects quasi-maximum likelihood Poisson specification with robust standard errors (in parentheses). Model 8 reports results from a random effects Poisson specification with bootstrap estimates of the standard errors to account for overdispersion and serial correlation.

The controls are highly consistent in sign and significance across models and dependent variables. Although most coefficients are in the expected directions, a few surprising deviations offer insights into the nature of nanotechnology innovation. For example, in the models of patent impact, publicly traded firms have lower predicted innovative performance than those that are privately held. One explanation for this negative association is that as firms grow and seek out investment from public equity markets, they may shift their focus from exploratory R&D to an emphasis on exploiting discoveries made in their earlier years. Similarly, contrary to expectations, inventor career experience exhibits a negative association with innovative performance. As inventors gain experience, they may rely on knowledge and routines acquired earlier in their careers. Given the pace of change in nanotechnology, this reliance could make innovation harder. Finally, inefficiency has a positive and significant association with both measures of innovation, a finding that accords with work done using computational modeling (Lazer and Friedman, 2007).

Table 2.4: Models of New Combinations†

	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Controls—Firm								
Publicly traded	−0.2915	−0.3408	−0.3049	−0.2498	−0.3372	−0.2808	−0.2844	−0.3254
	(0.2816)	(0.2770)	(0.2661)	(0.2613)	(0.2652)	(0.2608)	(0.2534)	(0.2136)
Research sites	−0.0210	−0.0098	−0.0127	0.0006	−0.0204	−0.0060	−0.0033	−0.0055
	(0.0347)	(0.0286)	(0.0259)	(0.0244)	(0.0259)	(0.0240)	(0.0243)	(0.0282)
Global geographic distribution	−1.1624	−1.1098	−0.9545	−0.9084	−1.0860	−1.0495	−0.9172	−0.9278
	(0.8370)	(0.8359)	(0.7970)	(0.7806)	(0.7871)	(0.7660)	(0.8079)	(0.7414)
Technological diversity	0.0711	0.1199	0.1095	0.1226	0.1448	0.1626	−0.0237	0.2708
	(0.3801)	(0.3716)	(0.3637)	(0.3540)	(0.3700)	(0.3592)	(0.3426)	(0.3560)
Non-nanotechnology patent stock (log)	0.2751**	0.2259**	0.2177**	0.2183**	0.2223**	0.2236**	0.2013**	0.2250***
	(0.1134)	(0.1028)	(0.0954)	(0.0934)	(0.0936)	(0.0910)	(0.0912)	(0.0649)
Controls—Main Research Facility								
Main research and HQ separate	0.1126	0.1515	0.1664	0.1521	0.2140	0.2007	0.2031	0.1259
	(0.3162)	(0.3139)	(0.3123)	(0.3010)	(0.3126)	(0.2997)	(0.2891)	(0.2951)
Recent relocation	−0.1005	−0.0877	−0.0926	−0.1119	−0.0811	−0.1009	0.0055	−0.0832
	(0.1909)	(0.1967)	(0.1921)	(0.1855)	(0.1977)	(0.1908)	(0.1744)	(0.1764)
California	0.0492	−0.0603	−0.0635	−0.0912	−0.0522	−0.0845	0.0242	−0.0738
	(0.2236)	(0.2408)	(0.2224)	(0.2269)	(0.2318)	(0.2379)	(0.2144)	(0.2030)
Local university ties	−0.0147	−0.0130	−0.0139	−0.0158	−0.0116	−0.0131	−0.0078	−0.0138
	(0.0143)	(0.0138)	(0.0136)	(0.0131)	(0.0133)	(0.0128)	(0.0107)	(0.0137)
Local inventor hires	0.0129	0.0062	0.0043	0.0030	0.0040	0.0026	0.0026	0.0054
	(0.0080)	(0.0074)	(0.0074)	(0.0074)	(0.0065)	(0.0064)	(0.0065)	(0.0080)
Inventor geographic distribution	−0.3431	−0.2457	−0.1113	−0.0554	−0.1569	−0.0933	−0.1241	−0.1368
	(0.2888)	(0.3020)	(0.3277)	(0.3319)	(0.3053)	(0.3074)	(0.2875)	(0.2992)
Distant inventors	0.0023*	0.0027*	0.0024*	0.0024*	0.0022*	0.0022*	0.0024*	0.0021
	(0.0012)	(0.0014)	(0.0013)	(0.0014)	(0.0013)	(0.0013)	(0.0014)	(0.0038)
Inventor career experience	−0.0008	−0.0011	−0.0011*	−0.0013**	−0.0009	−0.0011*	−0.0012*	−0.0010
	(0.0008)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0008)
High mobility inventors	0.0094	0.0056	0.0028	0.0009	0.0047	0.0025	0.0054	0.0006
	(0.0224)	(0.0229)	(0.0202)	(0.0205)	(0.0205)	(0.0207)	(0.0225)	(0.0240)
Inventor technological experience	−0.0089	−0.0210	−0.0235	−0.0153	−0.0311	−0.0235	−0.0422	−0.0102
	(0.0297)	(0.0298)	(0.0290)	(0.0286)	(0.0284)	(0.0280)	(0.0269)	(0.0273)
Median team size	−0.0975**	−0.0720	−0.0855*	−0.0766*	−0.0896**	−0.0802*	−0.0634	−0.0888**
	(0.0486)	(0.0456)	(0.0455)	(0.0445)	(0.0454)	(0.0438)	(0.0433)	(0.0416)
Inventors (log)	0.4260***	0.3458**	0.2910**	0.3229**	0.2828**	0.3154**	0.1200	0.3898***
	(0.1461)	(0.1416)	(0.1377)	(0.1301)	(0.1368)	(0.1278)	(0.1077)	(0.1354)

Table 2.4 (Continued)

Components	0.0018	0.0001	0.0000	−0.0021	−0.0005	−0.0029	0.0014	−0.0048
	(0.0050)	(0.0053)	(0.0054)	(0.0057)	(0.0051)	(0.0053)	(0.0054)	(0.0066)
Nanotechnology patent stock	0.0010	0.0014	0.0014	0.0016	0.0013	0.0016	0.0023**	0.0014
	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0012)	(0.0011)	(0.0012)	(0.0015)
Nanotechnology portfolio complexity	−0.0220	−0.0296	−0.0242	−0.0241	−0.0214	−0.0205	−0.0224	−0.0244
	(0.0546)	(0.0558)	(0.0557)	(0.0558)	(0.0551)	(0.0552)	(0.0547)	(0.0602)
Nanotechnology paper stock	−0.0012	−0.0007	−0.0010	−0.0009	−0.0012	−0.0011	−0.0010	−0.0009
	(0.0013)	(0.0013)	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0013)	(0.0025)
Independent Variables								
Firm proximity$_{(centered)}$		0.0124*	0.0129**	0.0086	0.0171***	0.0130**	0.0154***	0.0119*
		(0.0066)	(0.0061)	(0.0060)	(0.0059)	(0.0058)	(0.0054)	(0.0067)
Cohesion$_{(centered)}$			0.1049	0.0719	0.1063	0.0685	0.0544	0.0689
			(0.0940)	(0.0868)	(0.0950)	(0.0854)	(0.0869)	(0.0852)
Inefficiency$_{(centered)}$			0.2813***	0.2484**	0.3060***	0.2725**	0.2930***	0.2694**
			(0.1071)	(0.1034)	(0.1137)	(0.1060)	(0.1093)	(0.1219)
Cohesion × firm proximity				0.0120***		0.0130***	0.0130***	0.0128***
				(0.0026)		(0.0023)	(0.0024)	(0.0036)
Inefficiency × firm proximity					−0.0116***	−0.0130***	−0.0139***	−0.0131***
					(0.0040)	(0.0036)	(0.0038)	(0.0045)
Other								
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mills ratio (λ)	1.1348***	0.9384**	0.8946**	0.9060**	0.8400**	0.8435**		0.9063***
	(0.3834)	(0.4022)	(0.3986)	(0.3936)	(0.3925)	(0.3881)		(0.3259)
Constant								0.2093*
								(0.1187)
N	2491	2491	2491	2491	2491	2491	2491	2760
Firms in model	302	302	302	302	302	302	302	376
Log likelihood	−5831.26	−5779.05	−5733.18	−5694.37	−5702.61	−5656.67	−5691.43	−7121.68
Model d.f.	37	38	40	41	41	42	41	42

* $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$; two tailed tests.

† The estimates presented in Models 9-16 are derived from a conditional fixed effects quasi-maximum likelihood Poisson specification with robust standard errors (in parentheses). Model 16 reports results from a random effects Poisson specification with bootstrap estimates of the standard errors to account for overdispersion and serial correlation.

Hypothesis 1 predicts that proximity will positively affect a firm's innovation. Models 2 and 10 test this hypothesis by introducing the proximity variable. The results show that proximity to other firms (weighted by their annual nanotechnology patents) has a positive and significant effect on impact and new combinations. These findings support Hypothesis 1.

The next hypotheses address the interdependent effects of inventor network structure and firm proximity on innovation. Hypothesis 2 predicts a positive interaction between proximity and intraorganizational network cohesion. When firms have many neighbors, cohesive networks benefit innovation because they facilitate information processing and focused collaboration. In addition, given the frequency of spillovers in more concentrated areas, firms that have cohesive inventor networks in these settings are at less risk for the knowledge redundancies that plague such social structures. Hypothesis 3 predicts that inefficient networks will be ideal when firms are situated in more isolated environments with fewer peers that perform related R&D. In these contexts, inventors may be better able to compensate for reduced access to spillovers if they have opportunities to bridge disconnected colleagues who serve as nonredundant information sources. These types of networks should also promote parallel problem solving.

Models 6 and 14 test these hypotheses by introducing interactions between the two measures of network structure—cohesion and inefficiency—and firm proximity. The models show strong support for both hypotheses. The interaction between cohesion and proximity is positive and significant in both Models 6 (impact) and 14 (new combinations); as proximity to other nanotechnology firms increases, more cohesive inventor networks lead to greater innovative performance, as predicted by Hypothesis 2. Both models reveal, in support of Hypothesis 3, a significant negative association between proximity and inefficiency. These coefficients imply that as proximity decreases, inefficient networks are beneficial.

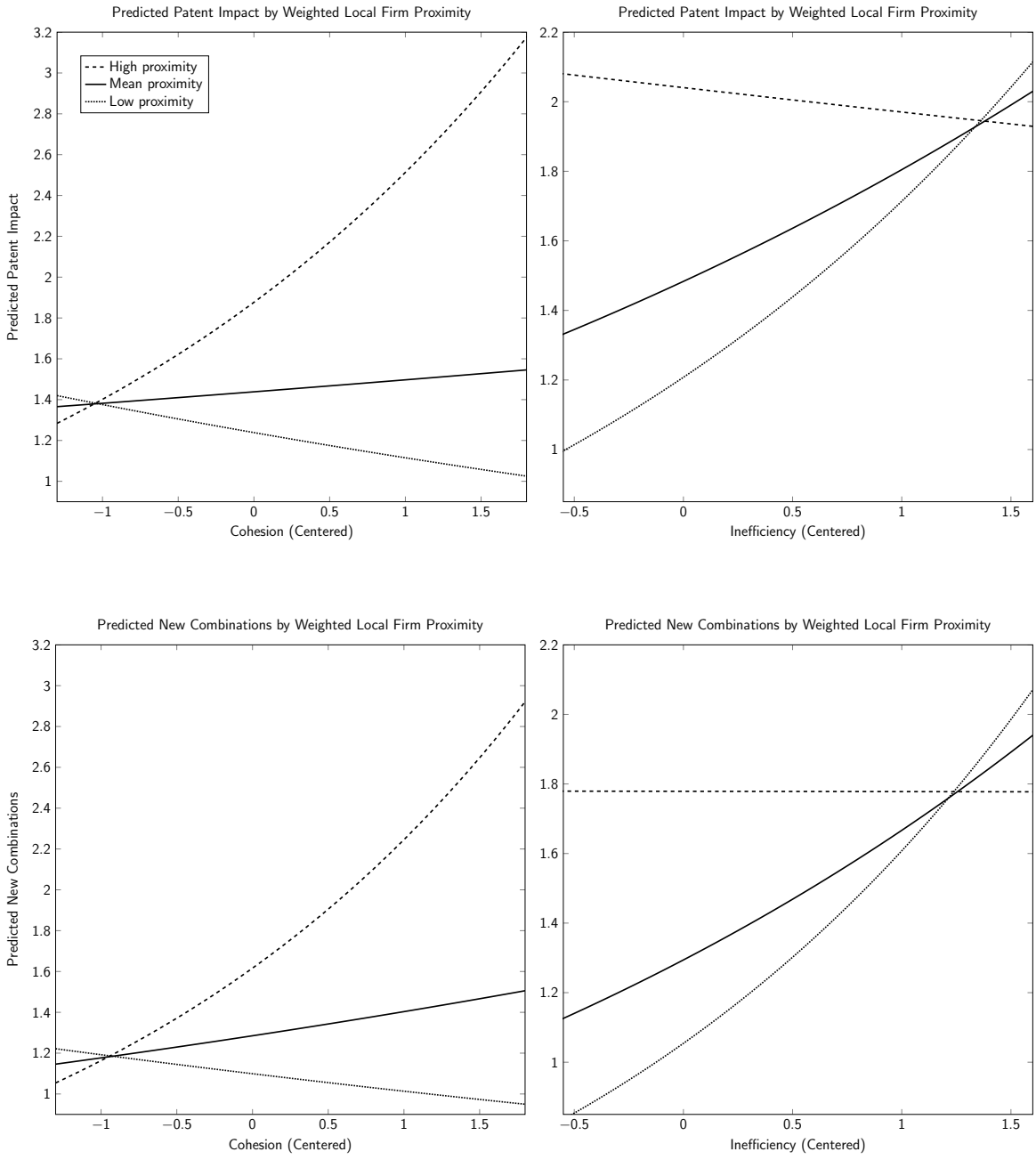

Figure 2.3: Predicted patent impact and new combinations. Calculations for impact are based on Model 6 and those predicting new combinations come from Model 14. In both cases, all control variables are held at their mean. Measures of high and low density are taken from the 90th and 10th percentiles of the regional density variable, respectively.

The plots displayed in Figure 2.3 offer illustrations of the interdependent effects of inventor network structure and firm proximity on innovation. The top two panels show predicted patent impact as a function of network cohesion (left side) and inefficiency (right side) for sample firms at the 10th, mean, and 90th percentile of

firm proximity using estimates derived from Model 6, while the bottom two panels present analogous plots for predicted new combinations using estimates from Model 14. To put the illustrated levels of firm proximity in perspective, note that the value of this measure for companies in San Jose, California—also the home of NeoPhotonics Corp.—is only slightly above the 90th percentile (in the 91st). By contrast, NanoScale Materials, Inc., in Manhattan, Kansas, has a firm proximity at almost exactly the 10th percentile. The values displayed for cohesion and inefficiency all lie within the observed data.

The figure reveals interesting nuances about the relationship between proximity and network structure. First, consider the two plots on the left-hand side, which illustrate the predicted effects of cohesion. Both graphs show that the most advantaged firms are those that are located near many other organizations and that also have highly cohesive networks. In fact, even at the mean level of cohesion (which is 0 after centering), firms with high proximity to industry peers have a predicted patent impact of 1.88 and a predicted 1.62 new combinations—respectively a 31 and 29 percent increase over those at the mean level of proximity. The two plots also suggest that cohesion leads to important performance variations within levels of proximity. As Hypothesis 2 predicts, firms with high proximity to other organizations see the most improvement with increases in cohesion. For example, a two standard deviation increase above mean cohesion (i.e., over the mean for cohesion) leads to a predicted 2.35 impact-weighted patents and 2.08 new combinations, increases of 25 and 28 percent, respectively. By contrast, among firms with less proximity, inventor network cohesion leads to a *decrease* in performance. Here, a two standard deviation increase above mean cohesion results in predictions of 1.14 impact-weighted patents and 1.03 new combinations, or respective performance *declines* of 8 and 6 percent.

The two panels displayed on the right side of Figure 2.3 elaborate the effects of inefficient networks on innovation. Findings are similar to those for network cohe-

sion: highly proximate firms have the best expected performance on both dependent variables when inefficiency is at its mean (which is also zero after centering). These predictions start to change with a two standard deviation increase above mean inefficiency (0.58 on the x-axis), where the gap separating firms at all three levels narrows considerably. This change stems from the different implications of inefficient networks—and brokerage opportunities and parallel problem solving—for firms at varying levels of proximity. For highly proximate firms, a two standard deviation increase above mean inefficiency leads to a 2.9 percent *decrease* in impact, while the predicted number of new combinations remains effectively unchanged. By contrast, firms with much lower proximity should see a 22 percent *increase* in impact and a 29 percent *increase* in new combinations for a corresponding two standard deviation move up on inefficiency.

Although the effects are most pronounced for firms with either very many or very few neighbors, fit between proximity and intraorganizational network structure is consequential for companies in many geographic settings. To see this, consider once again the 2003 map of U.S. nanotechnology firms displayed in Figure 2.1. Models 6 and 14 suggest the presence of an inflection point for determining whether more cohesion is likely to be beneficial or harmful at values of firm proximity that correspond to locations such as the Chicago suburbs (e.g., Hoffman Estates, Willowbrook). In 2003, 62 percent of sample firms were above this inflection point; 38 percent were below. If a firm is more isolated than those in such locations, then increases in intraorganizational network cohesion will likely harm performance. If a firm is less isolated, then increases in cohesion will generally be beneficial. For inefficiency, Models 6 and 14 reveal an inflection point at values of firm proximity corresponding to locations such as Livermore, California. The average proximity of a firm located in Livermore is approximately one standard deviation higher than the average proximity of a firm located in the Chicago suburbs. In 2003, 17 percent of sample firms were

above this inflection point; 83 percent were below. The position of this inflection points suggests that, for instance, firms moving from Livermore to the even more concentrated Menlo Park, California, should try to minimize network inefficiencies. By contrast, firms moving from Livermore to the Chicago suburbs will likely have better performance if inefficiency increases.[13]

In sum, all three hypotheses receive strong support. Location matters for innovation in nanotechnology. However, the magnitudes of locational benefits—and constraints—are moderated by intraorganizational network structure. If a firm performs R&D in proximity to industry peers, cohesive networks can promote innovation by making it easier for inventors to process information and enroll the support of their colleagues. By contrast, if a firm has fewer neighboring organizations, an inefficient network can be beneficial by creating brokerage opportunities that allow inventors access to nonredundant information and by promoting parallel problem solving. Finally, note that the results do not suggest that network structure substitutes for location. Although firms with high proximity do not perform better in all cases, at mean levels of cohesion and inefficiency, these firms have more favorable outcomes than others on both dependent variables. The results do imply, however, that performance advantages accrue to firms that have the right fit between their intraorganizational network structure and geographic context.

2.7 Robustness Checks

I performed a variety of analyses to examine the robustness of the findings. One concern was the potential endogeneity of location choice. If firms choose to locate in particular areas that offer unmeasured benefits for nanotechnology R&D, then the

[13] Although pairing these inflection points with locations on a map is informative for purposes of illustration, it is important to remember that the design of the proximity measure suggests that absolute counts of neighbors are less important than relative distances between proximate firms, weighted by their patent productivity.

observed benefits of proximity for innovation may actually result from the concentration of astute entrepreneurs. The fixed-effects specification, lag structure of the panel, and controls for patent stocks and inventor experience should alleviate many of these concerns (Blundell et al., 1995). However, knowledge about the growth of nanotechnology could be used to provide an additional robustness check. As noted earlier, equipment prices limited the commercial pursuit of nanotechnology until the late 1980s and early 1990s. To the extent that factors particular to nanotechnology drive location choices (e.g., proximity to customers, state tax credits), consideration of these factors should have been much less influential for firms that selected locations in the distant past. Demonstrating the results hold for this subsample would weaken the plausibility of endogeneity-based explanations. Models 21 and 22 in Table 2.5 provide estimates for firms that established their nanotechnology R&D in facilities that were in existence before 1990 and did not relocate during the study. Because these facilities never moved, the coefficients for recent relocation and California are not identified. Despite losing 60 percent of firms and 50 percent of firm-year observations, the results support the main findings.

Table 2.5: Robustness Checks†

	Model 17 (Impact)	Model 18 (NC)	Model 19 (Impact)	Model 20 (NC)	Model 21 (Impact)	Model 22 (NC)
Controls—Firm						
Publicly traded	−0.5531∗∗∗	−0.2684	−0.6624∗∗∗	−0.3599	−0.7771∗∗∗	−1.2231∗∗∗
	(0.2118)	(0.2602)	(0.2526)	(0.2331)	(0.2912)	(0.2641)
Research sites	0.0016	−0.0015	0.0025	−0.0161	0.0063	0.0048
	(0.0287)	(0.0243)	(0.0313)	(0.0239)	(0.0292)	(0.0223)
Global geographic distribution	−2.0803∗∗∗	−1.0444	−1.8595∗∗	−1.0111	−1.7495∗	−0.6640
	(0.8069)	(0.7831)	(0.8272)	(0.7391)	(0.9222)	(0.8610)
Technological diversity	−0.0396	0.2866	−0.1641	0.1825	0.1785	0.3664
	(0.4887)	(0.3913)	(0.4586)	(0.3466)	(0.5053)	(0.3552)
Non-nanotechnology patent stock (log)	0.3059∗∗∗	0.2051∗∗	0.2744∗∗∗	0.1863∗∗	0.5074∗∗∗	0.3934∗∗∗
	(0.1034)	(0.0932)	(0.1030)	(0.0853)	(0.1289)	(0.1112)
Controls—Main Research Facility						
Main research and HQ separate	0.2899	0.1848	0.3776	0.2388	0.8559∗∗∗	0.5673∗∗∗
	(0.4300)	(0.3109)	(0.4752)	(0.3175)	(0.1686)	(0.1393)
Recent relocation	−0.1110	−0.1580	−0.1850	−0.2819∗		
	(0.2345)	(0.2115)	(0.2418)	(0.1704)		
California	−0.1978	−0.0712	−0.2903	−0.1386		
	(0.2794)	(0.2600)	(0.3228)	(0.3146)		
Local university ties	−0.0108	−0.0149	−0.0189	−0.0131	−0.0167	−0.0119
	(0.0147)	(0.0132)	(0.0187)	(0.0117)	(0.0199)	(0.0186)
Local inventor hires	0.0053	0.0068	−0.0055	0.0028	−0.0031	−0.0038
	(0.0076)	(0.0066)	(0.0098)	(0.0072)	(0.0103)	(0.0083)
Inventor geographic distribution	0.6239∗	−0.1512	0.4497	−0.2961	1.1682∗∗∗	0.2986
	(0.3616)	(0.2927)	(0.3948)	(0.3065)	(0.4337)	(0.3445)
Distant inventors	0.0006	0.0024∗	0.0002	0.0020∗	0.0003	0.0024
	(0.0016)	(0.0013)	(0.0015)	(0.0012)	(0.0021)	(0.0016)
Inventor career experience	−0.0012	−0.0009	−0.0014∗	−0.0009	−0.0017∗	−0.0013∗
	(0.0008)	(0.0006)	(0.0008)	(0.0006)	(0.0009)	(0.0007)
High mobility inventors	−0.0041	0.0060	−0.0044	0.0042	−0.0057	0.0022
	(0.0248)	(0.0231)	(0.0236)	(0.0204)	(0.0234)	(0.0211)
Inventor technological experience	−0.0429	−0.0200	−0.0570∗	−0.0274	−0.0294	−0.0490
	(0.0318)	(0.0298)	(0.0317)	(0.0254)	(0.0323)	(0.0381)
Median team size	−0.0859∗	−0.0536	−0.0970	−0.0651	−0.1028∗	−0.0696
	(0.0510)	(0.0402)	(0.0613)	(0.0449)	(0.0571)	(0.0511)
Inventors (log)	0.3417∗	0.3823∗∗	0.2538	0.3109∗∗	0.3773∗∗	0.4515∗∗∗

Table 2.5 (Continued)

	(0.1845)	(0.1484)	(0.1653)	(0.1252)	(0.1630)	(0.1220)
Components	0.0013	−0.0035	0.0040	0.0013	−0.0073	−0.0121**
	(0.0086)	(0.0068)	(0.0071)	(0.0050)	(0.0077)	(0.0055)
Nanotechnology patent stock	0.0025**	0.0017	0.0032***	0.0021**	0.0013	0.0004
	(0.0011)	(0.0010)	(0.0012)	(0.0010)	(0.0010)	(0.0010)
Nanotechnology portfolio complexity	0.0020	−0.0163	−0.0416	−0.0397	0.0683	0.0091
	(0.0496)	(0.0563)	(0.0540)	(0.0568)	(0.0661)	(0.0684)
Nanotechnology paper stock	−0.0011	−0.0006	−0.0010	−0.0011	−0.0022	−0.0032
	(0.0012)	(0.0013)	(0.0012)	(0.0011)	(0.0027)	(0.0024)
Independent Variables						
Firm proximity$_{(centered)}$	0.0187**	0.0164**	0.0183***	0.0147***	0.0232***	0.0206***
	(0.0073)	(0.0067)	(0.0069)	(0.0052)	(0.0057)	(0.0056)
Cohesion$_{(centered)}$	−0.0033	0.0222	0.0890	0.1016	0.0917	0.1512*
	(0.0997)	(0.0901)	(0.1086)	(0.0956)	(0.1068)	(0.0916)
Inefficiency$_{(centered)}$			0.3041**	0.3145***	0.1496	0.2275**
			(0.1302)	(0.1195)	(0.1145)	(0.0969)
Constraint$_{(centered)}$	0.2144	0.2795				
	(0.4875)	(0.4107)				
Cohesion × firm proximity	0.0104***	0.0115***	0.0108***	0.0114***	0.0090*	0.0111***
	(0.0035)	(0.0026)	(0.0033)	(0.0024)	(0.0051)	(0.0033)
Inefficiency × firm proximity			−0.0131**	−0.0127***	−0.0123***	−0.0142***
			(0.0058)	(0.0037)	(0.0047)	(0.0031)
Constraint × firm proximity	0.0523***	0.0455***				
	(0.0160)	(0.0146)				
Other						
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mills ratio (λ)	1.2904***	0.8165**	1.1204***	0.6214*	2.1309***	1.7765***
	(0.4198)	(0.4161)	(0.3898)	(0.3642)	(0.4525)	(0.4875)
N	2569	2491	2561	2483	1374	1344
Firms in model	317	302	315	300	130	125
Log likelihood	−25956.54	−5675.59	−21554.77	−4355.76	−16444.15	−3646.90
Model d.f.	42	42	42	42	40	40

* $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$; two tailed tests.

† All estimates are derived from a conditional fixed effects quasi-maximum likelihood Poisson specification with robust standard errors (parentheses).

Next, I evaluated how sensitive the results were to a particular lag structure. Following research on technology lags (Kotha et al., 2011) I defined the dependent variables as, respectively, the number of impact-weighted nanotechnology patents and the number of nanotechnology patents introducing new combinations produced by each firm in two subsequent time periods (i.e., $t+1$ and $t+2$). However, because I relied on application dates for assigning patents to years—which usually correspond closely to the date of invention—and nanotechnology firms are generally quick to seek patent protection, I also considered a shorter lag, one year. Models 19 and 20 report these results, which are consistent with those relying on lags of two periods.

I also consider an alternative measure of inefficient networks. As discussed above, average path length is an attractive proxy because it captures networks that are less connected and in which information diffusion is slow (Lazer and Friedman, 2007). However, path lengths are undefined for nodes that reside in different components. To confirm that the findings are not dependent on this measure choice, I also estimated models that substituted Burt's (1992) measure of constraint averaged over all actors in the network. Lower values on this measure signal less connected networks. The results, shown in Models 17 and 18, support the core findings.

The hypotheses are based on the assumption that a key advantage of proximity is access to knowledge. Thus, I weighted firm proximity by the annual count of patents awarded to each sample firm. However, patents may be a poor proxy for knowledge access; furthermore, a few isolated firms with unusually high patenting rates might influence the measure unexpectedly. In unreported analyses I address these concerns using a version of firm proximity with the weight, x_j, set to 1. The results are supportive of the findings with the alternative measure.

Finally, some authors suggest that including the inverse Mills ratio in nonlinear models may lead to inconsistent estimates (Terza, 1998). In unreported analyses, I performed a Box-Cox transformation of the two count dependent variables and

reestimated the models with OLS. The results are similar to those discussed above.

2.8 Discussion and Conclusion

Organizational innovation is a complex phenomenon, one that in many cases appears largely driven by serendipity. Despite this complexity, researchers have uncovered many factors bearing on firms' ability to generate novel ideas, processes, and products. Since at least the time of Marshall's (1890, 271) observation that in geographically concentrated industries "the mysteries of the trade become no mysteries; but are as it were in the air," social scientists have emphasized the importance of proximity for innovation. To generate the novel recombinations that underpin important discoveries, actors need exposure to diverse knowledge. Moreover, in dynamic fields of technology, cutting-edge knowledge is often difficult to codify and transmit over distances. Location near other organizations helps firms acquire both sorts of knowledge by increasing chance encounters—and thus the broad diffusion of diverse information—and by facilitating face-to-face interaction (Bathelt et al., 2004; Malmberg and Maskell, 2006).

More recently, organizational theorists have extended these insights by showing that not all firms benefit equally from proximity (McEvily and Zaheer, 1999; Tallman et al., 2004). Researchers emphasize that the effects of proximity on innovation are moderated by whether a firm can pair information acquired locally and informally from neighbors with other more exclusive sources of new knowledge. Firms acquire such knowledge by establishing ties to distant collaborators (Bathelt et al., 2004; Whittington et al., 2009), recruiting employees (Saxenian, 1994; Zucker et al., 1998), embedding themselves in scientific communities (Gittelman, 2007; Owen-Smith and Powell, 2004), or forging other connections outside their boundaries.

Despite these advances, theories of geography and innovation in organizations remain limited in several respects. Importantly, because they have been designed

primarily to explain the process of knowledge acquisition, existing perspectives are not equipped to account for how firms that are located near many other organizations internalize, process, and use the potentially enormous volumes of information available to them locally. Contemporary theories also do not address how, despite lacking the advantages of proximity, some firms that are relatively isolated geographically are able to produce important innovations.

The approach developed in this chapter is an effort to advance theories of geography and innovation by integrating insights from research on networks in organizations to demonstrate the importance of considering firms' local external environments and their internal patterns of collaboration in tandem. The findings show that firms can be successful innovators whether they are located in the heart of Silicon Valley or in the more remote areas of the American Midwest, but doing especially well in either environment requires making the most of where they are. Intraorganizational networks are a key source of support for inventors, but the structure of these networks has consequences for their value. If a firm is proximate to many industry peers, then a cohesive network helps inventors process knowledge spillovers and mobilize support from colleagues for developing their ideas. By contrast, when a firm has few proximate organizations from which to capture spillovers, inefficient networks that are slow at diffusing information are beneficial. Such networks provide opportunities for brokerage and parallel problem solving, which together help create and sustain the diverse ideas and perspectives needed for innovation.

While firms can succeed in both higher- and lower-proximity environments, they do not necessarily succeed equally. An important though unhypothesized finding of this study is that firms with high proximity to industry peers and very cohesive networks should have the greatest absolute performance. One explanation for this finding is that these firms have the best of both worlds: from their external environment, their employees obtain frequent exposure to new knowledge, which stimulates

ideas and provides raw material for novel recombinations. Further, the cohesive structure of their networks suggests that employees of these firms will find support from their colleagues in developing ideas. Despite these benefits of proximity, location near many neighboring organizations can pose challenges. As illustrated in Figure 2.3, for those located in the highest-proximity areas, sparsely connected, inefficient networks lead to worse performance than will be demonstrated by comparable firms in more isolated environments. Further, the findings show that when firms in isolated locales have highly inefficient networks, they are able to perform better than some peers in higher-proximity areas. In sum, although appropriate intraorganizational network structures should not be viewed as a substitute for proximity, network structure does moderate the benefits of proximity for innovation in important ways.

To arrive at these conclusions, in this study I employed a novel research design that made use of patent data to approximate the structure of collaborations among inventors at several hundred high-technology firms over a 14-year period. Though existing studies examine intraorganizational networks, they tend to focus on single organizations and examine the relationships among individuals or units. While this work has led to valuable insights, research comparing global network structures among a broad sample of organizations is necessary for understanding how aggregate patterns of relations and the structural embeddedness of actors influence performance (Granovetter, 1992; Phelps et al., 2012). Moreover, this study also used detailed, time-varying data on R&D locations to situate each intraorganizational network in physical space. Care was taken to control for explanations that emphasize knowledge acquisition from external sources such as collaboration, mobility, and science. Together, these data enabled a rare investigation that relates insights from macro research on the *external* determinants of innovation with those from micro studies that point to the importance of *internal* network structures.

Before discussing some broader implications, I note several limitations. First,

the models presented here use a measure of proximity based on distance to other firms. However, organizations including universities, nonprofit research institutes, and government laboratories are active in nanotechnology and may provide spillover benefits. Moreover, for some firms, distance to a key partner could be more important than overall proximity. Future research should explore more nuanced measures of regional composition to offer a better understanding of the effects of geography and intraorganizational ties on outcomes of interest. Second, like the models in all studies that derive network and innovation data from archival sources like patents, the models used in this chapter miss many of the collaborative ties and innovations that leave no paper trail. Though I have made efforts to mitigate this problem by studying a field where patenting is common, future research could benefit from alternative data on intraorganizational ties and performance, such as scientific publications or surveys. These data would also be valuable for better documenting the interpersonal connections of a firm's employees to colleagues outside the organization—such as those to collaborators from earlier career stages—and how various human resource practices might facilitate or constrain the persistence of these ties (Stern, 2004).

Despite these limitations, the theoretical approach and empirical findings have implications for research on geography and innovation, collaboration networks, and social capital.

Social structure and the geography of innovation. A major contribution of this study is that it shows how internal social structures moderate the effects of geography on organizations. I focused specifically on the implications of this insight for firms' performance at innovation. Beyond helping to account for innovation among firms in very different local environments, knowledge of collaborative structures within organizations is likely to have broader value for research on the geographic diffusion of knowledge and ultimately the vitality of regional economies. An array of recent work argues that the character of different regions results in large part from the net-

works connecting organizations within them (Bell, 2005; Whittington et al., 2009). Surprisingly, even among comparatively successful regions, these network structures vary widely in form (Buhr and Owen-Smith, 2010; Saxenian, 1994). Moreover, some research suggests that well-known models of network diffusion do poorly at explaining regional knowledge flows (Fleming et al., 2007). Future research on how networks affect the geographic diffusion of knowledge might benefit from attending to patterns of collaboration within the organizations that constitute the nodes of such networks. To the extent that they vary among organizations, within and across regions, intraorganizational networks should have consequences for broader knowledge flows because of differences in their capacity to absorb, transmit, and alter the information that diffuses to (and through) them geographically.

The findings of this study also point more broadly to the need for systematic research on geographically isolated firms. Existing work on geography and innovation largely focuses on explaining the conditions under which proximity leads to maximum performance gains. Yet many innovative companies, even in knowledge-intensive sectors, are located far from peer organizations. Future analyses should seek to further explain the success of such companies given the disadvantages of isolation. Additional research in this area will both help to clarify the conceptual relationship between geography and innovation and also lead to valuable insights for practicing managers of firms in locations less proximate to industry peers.

Collaboration networks and performance. This study also makes theoretical contributions to the understanding of networks in organizations. Researchers have made progress in identifying the relative advantages and disadvantages of different network structures for a variety of outcomes ranging from creativity and innovation to career advancement (Burt, 2004; Fleming et al., 2007). A consistent finding in this literature echoes an early contingency theory observation that "there is no best way to organize" (Galbraith, 1973, 2). Prior research at the individual level has

focused extensively on how factors like personality, task requirements, and tie strength moderate the effects of networks (Hansen, 1999; Mehra et al., 2001; Uzzi, 1997).

My approach builds on these insights from ego network research but departs by considering contingencies in the overall structure of relations—at the global network level. The findings accord with ego research in supporting the value of a contingency approach. However, the results of this study also suggest that theoretical insights drawn from ego network research about the effects of different contingencies may not easily translate to the global level. Moreover, the functioning of global network structures can appear counterintuitive if interpreted through an ego network lens. For example, a key finding of this study is that organizations with less access to new knowledge from geographically proximate sources have better performance if their employees are less connected to one another. In such cases, decreases in connectivity are beneficial because they create and sustain diversity through parallel problem solving and by opening brokerage opportunities. This diversity prevents organizations from settling prematurely on suboptimal solutions to problems (by generating many to pick from) and allows employees to bridge diverse pockets of knowledge in search of novel recombinations (Burt, 2004; Lazer and Friedman, 2007).

Note that my prediction of greater performance is at a collective, not an individual, level. Lower connectivity does not imply that all actors in a network will see performance gains; some individuals will likely do worse because they either devote too much time to following unpromising leads or because they cannot identify a necessary piece of information, both contingencies that might be averted with better communication. Likewise, few theories of networks suggest that at an ego level, becoming less connected or occupying a network position that is less efficient for collecting information fosters innovation (Singh and Fleming, 2010; Wuchty et al., 2007). Thus, the effects of network contingencies can diverge across ego and global levels of analysis; future research should work to identify the applicability of different

contingencies across levels.

Structure of community social capital. The findings of this study also have implications for community social capital, defined as "the benefits that accrue to the collectivity as a result of the maintenance of positive relations between different groups, organization units, or hierarchical levels" (Ibarra et al., 2005, 360). From the time of a few early theoretical statements (Bourdieu, 1986; Coleman, 1988), community social capital theorists have generally emphasized the importance of dense, cohesive ties among network members for producing collective benefits (Lin, 1999). Network fragmentation, moreover, is typically expected to create a host of negative consequences for members of a community (Burt and Ronchi, 1990; Putnam, 2000). The results presented in this chapter, particularly the finding that less connected networks can improve collective innovation, suggest that a more nuanced, contingency-based perspective might prove useful in future research on community social capital. Although dense, cohesive ties provide community members with certain benefits, such as trust and monitoring (Coleman, 1988), they can also preclude other advantages, such as diverse ideas and perspectives. Depending in the goals of those embedded in the community, greater or lesser connectivity could be beneficial.

Team design and venture location choice. This study also has a number of managerial implications. Although it may seem counterintuitive in light of beliefs about the value of connectivity, managers seeking to stimulate innovation should consider structuring teams in ways that limit overall ties among their R&D employees, especially when those employees have little exposure to new knowledge from peers at proximate organizations. Decreasing connectivity might lead to slower information sharing, but such inefficiencies can also help preserve the diverse perspectives necessary for tackling complex problems. To decrease connectivity, managers could create formal units to house separate project teams. Skunk-works models that introduce rigid, physical separations between new product development groups might also

be appropriate (Fang et al., 2010). By contrast, for managers operating in settings proximate to industry peers, increasing cohesion among R&D employees should foster innovation. Managers might achieve greater cohesion by holding frequent meetings at which employees who work on diverse kinds of projects can come together and engage in collective problem solving (Hargadon and Sutton, 1997).

The results might also be useful for entrepreneurs or managers seeking to choose a location for a new venture or relocate an existing one. Dense concentrations of firms in places like Silicon Valley or Boston are attractive because they facilitate local knowledge transfer and offer other well-documented benefits (Owen-Smith and Powell, 2004). Managers, however, might have reasons for favoring other locations that are less proximate to rivals; in that case, the findings of this study suggest, closely monitoring patterns of collaboration among employees might attenuate some of the innovation disadvantages of isolation. Put differently, the findings of this research should help managers evaluate the trade-offs of different kinds of locations.

CHAPTER III

The Dark Side of Brokerage: Conflicts Between Individual and Collective Pursuits of Innovation

f

3.1 Introduction

In the vocabulary of social network analysis, a broker is a person who is connected to people who are not directly connected to one another (Burt, 1992).[1] Although interest in this general phenomenon can be traced back at least to the writings of Simmel (1950), research on the causes and consequences of brokerage has increased dramatically over the past two decades (Ahuja, 2000; Fernandez and Gould, 1994; Fleming et al., 2007). One common finding among studies in this area is that brokers tend to have better outcomes. For instance, people with less densely interconnected contacts are likely to get more compensation than their colleagues (Burt, 1997; Mizruchi et al., 2011; Podolny and Baron, 1997). Individuals who occupy brokerage positions in a network also tend to have more timely access to information about job openings and other resources that help them advance their careers. And many studies report

[1]More precisely, people have the *opportunity* to broker if they have disconnected contacts. They may or may not act on that opportunity. Nevertheless, for smoother exposition and to follow the convention of earlier writing, I use the terms "broker" and "brokerage" to imply these opportunities.

positive associations between network brokerage, creativity, and other dimensions of workplace performance (Aral and Van Alstyne, 2011; Lingo and O'Mahony, 2010; Obstfeld, 2005). The mechanism by which brokerage may lead to these sorts of benefits depends on the context, but in many cases it emanates from the broker's ability to access and control the flow of information among his or her disconnected contacts.

As these findings illustrate, the private benefits of brokerage are well documented in prior research. However, surprisingly little is known about the implications of brokerage for people other than the person in the broker role. This knowledge gap is surprising because by definition, brokerage involves the broker and two other people. Do these other individuals get any returns to performance by virtue of their mediated connection? If so, how do these returns compare to those of the broker? More broadly, do the effects of brokerage matter for the organizations in which the broker and the people he or she connects are embedded?

There are undoubtedly many cases where having a tie to a broker is a good thing. After all, by virtue of their position, brokers have access to unique information, contacts, and other resources that they may share with others.

However, there is also a potentially darker side of brokerage. A large part of what may make a broker especially helpful is his or her ability to access and share information. But there are many barriers to sharing. For instance, brokers may deliberately refrain from passing along information if doing so would allow them to exploit potentially valuable opportunities on their own. Similarly, because the value of their position often depends on the absence of ties between their contacts, brokers may choose not to facilitate otherwise useful connections. And even in cases where they do not behave opportunistically, brokers may withhold potentially valuable information because sharing would require they devote too much time and energy to translation. Finally, having disconnected contacts may place more demands on brokers, and therefore leave them with less time and attention that they can devote to

each of their contacts individually. Given these considerations, it is unclear whether the advantages of a broker's position spill over in a positive, negative, or neutral way to influence other people's performance.

In this chapter, I explore the potentially negative effects of connecting to brokers for innovation using data on the intraorganizational collaboration networks of inventors in 37 pharmaceutical firms that were active in research and development (R&D) between 1997 and 2001. To help disentangle the causal effects of brokerage, I use propensity score weighting in a pretest–posttest framework with a double robust estimator that is unbiased if either the model for exposure or outcome (or both) are correctly specified. Additionally, when testing the effects of connection to a broker, I am able to limit my investigation to changes among existing contacts, where the decision to connect was made prior to the contact becoming a broker and is therefore exogenous to the focal inventor's performance. My findings are strongly consistent with the notion of a dark side of brokerage. Using multiple proxies for innovation, I find that becoming a broker has a positive and significant effect on performance, but the opposite is true for having a connection to one.

The remainder of this chapter is organized as follows. First, I review the existing theoretical and empirical work on brokerage and innovation to develop my hypotheses. Next, I offer a sketch of the research setting, which I follow with an overview of my statistical approach and strategy for identifying causal effects. I then turn to the findings, and close with a discussion of their implications for theories of social capital, innovation, and future research.

3.2 Brokerage and Innovation

In this section, I offer theoretical arguments in support of two predictions alluded to above. First, net of other factors, becoming a broker should have a positive effect on an inventor's performance at innovation. Although not without exceptions (e.g.,

Ahuja, 2000; Lee, 2010), prior research offers support for this idea and therefore the role of the prediction in this study is to help establish a baseline. The second prediction is that connecting to a broker should negatively affect an inventor's performance, again holding other factors constant. Earlier writings also anticipate this prediction (e.g., Brass, 2009), but relative to the direct effects of brokerage it has received far less attention, and I am unaware of any empirical investigations that examine the consequences of connection to a broker for innovation. Below, I discuss each of these hypotheses in greater detail.

One prominent view of innovation suggests that novelty emerges from the rearrangement of existing physical and conceptual materials into new configurations (Arthur, 2009; Fleming, 2001; Schumpeter, 1934). For people in the business of creating new things, access to colleagues with diverse ideas, perspectives, and experiences is essential because it expands the amount of raw material available to them for recombination and offers insight into possible applications. Given the importance of diversity for creativity, it seems reasonable to anticipate that an inventor's position in a social network will influence his or her performance at innovation. But what kinds of network positions are most helpful?

Evidence from prior research suggests that having ties to people who are not directly connected to one another is useful for gaining exposure to diverse information. One reason for this is that there is generally greater variation in the information that people have between groups than within them. People tend to pass information along more quickly to friends, colleagues, and acquaintances with whom they are closely connected; after those friends, colleagues, and acquaintances receive that information, they may in turn pass it along to their own friends, colleagues, and acquaintances. The information among people with many dense interconnections tends to be more homogeneous because of this sharing behavior; if a person does not get a particular piece of information from one contact, he or she may still get it from another mutual

acquaintance (Granovetter, 1973). The opposite is true for disconnected contacts. Because they travel in different circles, an inventor's disconnected contacts will likely offer him or her exposure to more diverse information. Therefore, occupying a brokerage position should enhance inventors' performance at innovation, not only by increasing the stock of physical and conceptual material that they may recombine, but also by exposing them to more general information about different processes, methods, and tools to support their workflow.

Connection to otherwise disconnected partners may also enhance an inventor's performance in deeper ways, especially through its effects on learning. Burt (2004, 2005) touches on this idea in describing what he calls the "vision advantage" of brokerage. In his formulation, brokers have a vision advantage because their position helps them become "more familiar with alternative ways of thinking and behaving" and provides a window into "options otherwise unseen" (Burt, 2005, 59). The vision advantage of brokerage is related to the benefit of access to diverse information, but it also differs in important ways. Access to diverse information should be useful primarily for inventors because it allows them to import materials that they may use in novel combinations. By contrast, a vision advantage implies a deeper change in perspective that happens when a person is pushed to be more open-minded. To the extent that occupying a brokerage position makes an inventor more receptive to different ideas, perspectives, and experiences more generally, he or she may be better at developing innovations that build on physical and conceptual materials from a range of sources, even beyond his or her immediate contacts. In short, a vision advantage implies that more than just being useful for acquiring information and knowledge, occupying a brokerage position also leads to potentially much deeper benefits of learning.

Benoît Mandelbrot—the founding father of fractal geometry and one of the 20th century's most influential mathematicians—offers a useful illustration of the vision

advantage. In an interview given just a few months before his death, Mandelbrot remarks on the relationship between his unconventional career and his unique intellectual contributions.

> When people ask me what's my field? I say, on one hand, a fractalist. Perhaps the only one, the only full-time one. On the other hand, I've been a professor of mathematics at Harvard and at Yale... But I'm not a mathematician only. I'm a professor of physics, of economics, a long list. Each element of this list is normal. The combination of these elements is very rare at best. And so in a certain sense, it is not the fact that I was a professor of mathematics at these great universities, or professor of physics at other great universities, or that I received, among other doctorates, one in medicine, believe it or not. And one in civil engineering. It is the coexistence of these various aspects that in one lifetime it is possible, if one takes the kinds of risks which I took, which are colossal, but taking risks, I was rewarded by being able to contribute in a very substantial fashion to a variety of fields. I was able to reawaken and solve some very old problems. The problems are just so old that in a certain sense, they were no longer being pursued. And nobody... It was a hopeless subject. But I did it and there's a whole field by which has been created by that. (Mandelbrot, 2010)

Mandelbrot's quote does not suggest he was able to make great breakthroughs because he had access do lots of diverse information that he could piece together and make into something useful. Rather, his connection to so many different organizations and fields seems to have instilled in him a particular way of thinking—a vision—that allowed him to take on problems in ways that were fundamentally different from his peers who did not have those kinds of experiences.

Beyond information and vision, brokers also get control benefits from their positions. This dimension of brokerage has been explored most extensively in research on competitive economic contexts (Fernandez-Mateo, 2007). Fewer studies consider the potential control benefits of brokerage in creative settings, perhaps because these settings tend to be more collaborative in nature (Lingo and O'Mahony, 2010; Obstfeld, 2005). Nevertheless, the ability to serve as an intermediary and control the flow of information between two otherwise disconnected parties may be important for

understanding performance at innovation, for several reasons. First, to the extent that brokers are able to regulate access to essential information among their disconnected contacts, they may be in a better position to scoop those contacts when a breakthrough is imminent. A second, related possibility is that a broker may use his or her exclusive access to information to create a relationship of dependence among the disconnected groups. This may be especially likely to happen in creative settings when the broker has some special knowledge, skill, or credential that allows him or her to translate information or access resources across groups that the otherwise disconnected members could not do on their own. Both of these scenarios suggest the possibility that brokers may have better performance at innovation at least in part because of their ability to control the flow of information.

Despite these potential benefits, there are also some reasons to believe that occupying a brokerage position may be harmful for innovation. For instance, because they straddle otherwise disconnected groups, brokers may be seen as outsiders, and therefore have trouble getting the support of their colleagues. Burt (2004) offers some support for this idea in his finding that although supply chain managers who spanned disconnected sets of contacts at a large electronics company were more likely to have good ideas, they were unlikely to act on them. The value of occupying a brokerage position for innovation may also depend to a certain degree on whether the disconnected contacts have information that is complementary or potentially useful for recombination. Brokers who span groups that are too drastically different from one another may simply be draining their time.[2] Finally, to the extent that brokers use their position to control the flow of information for their own gain, they may lose the trust of their colleagues, thereby making innovation more challenging.

Despite these potential downsides, most theory and evidence points to some benefits of occupying a brokerage position for innovation. Put differently, the benefits

[2]However, some of the benefits cited above, especially those relating to vision and learning, may persist even if there are no complementarities across the groups.

appear to outweigh the costs. With these considerations in mind, I therefore propose the following, baseline hypothesis.

Hypothesis 4. *Entry into a brokerage position will have a positive effect on an inventor's performance at innovation.*

If becoming a broker is helpful for an inventor's performance, can the same also be said of having a connection to one? Are there spillover effects, such that a broker's contacts also benefit from his or her network position? On quick inspection, there seem to be some reasons to believe connection to a broker may entail positive spillovers, while perhaps even eliding several of the costs of being one. By virtue of their positions, brokers should have access to relatively more diverse information, which they may in turn pass onto their contacts; in so doing, brokers may help advance their contacts' performance. Moreover, brokers may filter, sort, and screen the information that they pass along to others—keeping only the most relevant pieces—which suggests that a connection to a broker may help a person economize on search costs. And unlike a broker, people who simply have a connection to someone who straddles disconnected group may have fewer problems getting their colleagues to devote time and energy to supporting them.

Despite these potential benefits, a closer look suggests a darker side of having a connection to a broker. Notably, several considerations cast doubt over the reliability of brokers as useful sources of novel information. First, there is a problem of incentives. When brokers pass along novel information among their contacts, they make it easier for others to compete with them by eroding the unique value of their position. To the extent that a person who straddles disconnected contacts fears such competition, he or she may choose not to share otherwise useful information. Of course, these concerns about disclosure may vary substantially with organizational policies. Many qualitative studies of brokerage examine highly collaborative organizations where there are policies that foster information sharing (e.g., Obstfeld, 2005;

Hargadon and Sutton, 1997; Lingo and O'Mahony, 2010). But there are also many organizations where there are few incentives for sharing across boundaries. Microsoft is an especially illustrative example (Eichenwald, 2012). Under its stack-ranking system (abandoned in late 2013), employees were evaluated once every six months. Managers were required to rank their employees in order to fit a predetermined distribution from top performers to poor, regardless of whether their employees were meeting expectations. Some observers suggest that the policy stifled collaboration, as employees sought to avoid working with people who might outshine them at evaluation time.

Second, even when brokers do have an incentive to share information, problems of transmission may prevent them from doing so effectively. Brokers are able to access diverse information at least in part because they have direct access to disconnected contacts. But any time a broker attempts to pass along information to others, he or she adds a step between the ultimate sender and receiver, thereby creating the possibility for noise and distortion, even if unintentionally. Closely related to this idea is the observation that people who have ties to many different areas are less likely to be experts in any of them. To the extent that brokers have a more shallow understanding of the information they receive from their disconnected contacts, they may be more likely to introduce errors when transmitting it to others, or to simply refrain from sharing because it is too hard for them to articulate. Finally, even wen brokers do wish to share information, their ability to do so effectively may suffer if they have a limited knowledge of what their contacts would actually find most useful.

Up to this point, I have focused on outlining why having a connection to a broker may not be especially helpful for getting novel information. But is it possible that brokers may facilitate learning among their partners? In short, are there positive spillovers of the vision advantage? Available evidence offers hints that the learning benefits of brokerage come through direct contact. For instance, Burt (2007) conducted three separate studies of performance among managers, bankers, and ana-

lysts. Across these three groups, he finds a positive association between direct access to non-redundant contacts and performance, but no evidence of benefits to indirect access. Because information should flow relatively easily among indirect contacts, Burt (2007) interprets these findings as evidence that brokerage creates value by facilitating learning, and moreover, that learning requires direct connection, and a deep need (or even requirement) to confront diverse ideas and perspective.[3]

Findings from studies in social psychology also suggest that the learning benefits of connection to a broker may be minimal. In a series of experiments, Stasser and Titus (1985, 1987) examined information sharing and group decision-making. They found that when people share more information relevant to a decision before beginning their deliberations, they tend to focus their discussions on the information they have in common relative to the unique pieces each brings to the table. Therefore, it seems plausible that when people talk with brokers, they may spend more time discussing already shared information, rather than the potentially novel insights and perspectives of the brokers, thereby minimizing the learning benefits of connection.

To the extent that connections to brokers do not offer information or learning benefits, it may be reasonable to predict that the implications of such ties for an inventor's performance are neutral. Consider the counterfactual. Had the inventor not been connected to a broker, he or she may not have acquired any novel information or learned anything anyway—nothing ventured, nothing gained. But there are a few reasons to think that connection to a broker may entail costs. Recall that one of the advantages of being a broker is control over information flows. By virtue of their position, brokers who are especially opportunistic may take advantage of their disconnected contacts, swooping in at the last minute to scoop them on a breakthrough or withholding a critical piece of information necessary for success. Even connection

[3]Burt's (2007) findings are, however, open to other interpretations. For example, following the logic of my earlier discussion about the incentive and transmission problems of sharing, it may be that brokerage does create value by providing access to novel information (rather than through learning and vision) but that brokers are bad at diffusing information among their contacts.

to a more benevolent broker entails risks. An inventor may depend on a broker for access to information, data, or other resources that are essential for his or her work. But what if the two have a falling out, or something happens to the broker? Losing any relationship can be bad for a person's performance, but losing one that offers irreplaceable access to a resource may be devastating.

One final consideration suggests that there may be negative (rather than neutral) spillovers from connection to a broker, and that is time. All relationships require some level of maintenance. At the very least, people need to have conversations, phone calls, exchange emails, or other forms of communication to share information. And gaining access to some information likely entails substantially greater investments of time, especially when that information is complex or sensitive (Hansen, 1999). Following this logic, it seems likely that brokers spend more time on maintaining relationships. One advantage of having contacts that know each other is that the amount of time and energy spent on maintenance should be less. As an example, with a single lunchtime meeting, an inventor who has connected contacts can catch up with all of his or her colleagues, while a broker needs to have at least two and maybe more lunches to do the same. An implication is that brokers will have less time and energy to devote to any given contact. From the perspective of a person with a connection to a broker, relative to other colleagues, the broker will likely have less ability to offer support, and in so doing, may negatively effect performance.[4]

By way of summary, the arguments above suggest that brokers are likely an unreliable source of novel information, and moreover, their ability to facilitate learning among their contacts also appears limited. Having a connection to a broker is also risky because it may make an inventor vulnerable to opportunism and dependence.

[4]Vedres and Stark (2010, 1174) offer some support for this idea at the inter-firm network level, and moreover, suggest that connection to a broker may ultimately harm the group, perhaps by increasing demands on the time of members. As they report on their findings, "the number of brokered ties to other groups is significantly correlated with decreased group cohesion, a finding that suggests that brokers adversely affect the structures they exploit. This finding is in line with the idea that the price of brokerage is borne by those who are connected by the broker."

And finally, connections may require investments of time, but it is unclear whether brokers can reciprocate, potentially making them less valuable than closely connected colleagues. These considerations lead to a second hypothesis.

Hypothesis 5. *A tie to a broker will have a negative effect on an inventor's performance at innovation.*

A few existing studies make predications that are related to mine, although they do not address innovation. For example, Burt (2007, 2010) reports on findings from three separate studies comparing the relative effects of brokerage and "secondhand" brokerage among managers, bankers, and analysts. Secondhand brokerage refers to indirect connections to otherwise disconnected groups. Although the results of Burt's investigation suggest a strong relationship between direct brokerage and performance, there is little evidence that the secondhand type is beneficial.[5] Galunic et al. (2012) conducted a related study that also looked at bankers. Similar to Burt's (2007, 2010) results, they do not find evidence of positive returns from indirect access to disconnected contacts, unless that access comes from someone of higher rank. Finally, Fernandez-Mateo (2007) examined a firm known as InterCo that specialized in finding temporary positions for information technology (IT) professionals. Because it acted as an intermediary between buyers and sellers, the firm was a broker in the true sense of the word. In line with observations about the control benefits of brokerage, Fernandez-Mateo reports that InterCo was able to offer discounts to its preferred buyers at the expense of the IT professionals, and therefore without lowering its own margins.

[5]However, Burt (2007) does find some evidence that secondhand brokerage is related to direct brokerage in a later period; watching brokers may be useful for learning how to become one.

3.3 Research Setting

I test the hypotheses laid out above using data on 37 pharmaceutical firms that were active in researching and developing novel human therapeutic compounds between 1997 and 2001. Several considerations make pharmaceuticals an attractive setting in which to study the relationships between brokerage and innovation. R&D is essential to the success of leading pharmaceutical firms, and according to National Science Board (2012) data, the industry consistently ranks among the highest in R&D spending and intensity. Moreover, pharmacologists, medicinal chemists, and other inventors in the field rely on highly specialized, complex, and often-tacit knowledge. Successful innovation therefore requires teamwork, which suggests the relevance of network factors in this setting.

Pharmaceuticals are also attractive because there are data readily available that allow me not only to measure the structure of inventors' personal collaboration networks over time, but also to identify each inventor's position in the broader, intraorganizational collaboration network that links together inventors across a particular firm. Companies in this sector have a high propensity to guard their intellectual property by seeking patent protection. As I elaborate below, patents are valuable for studying collaboration networks because they provide a written record of the people who worked together on a particular invention.

More broadly, in contrast to many other sectors, it is possible to link pharmaceutical patents directly to products and therefore to get some sense for the potential marketplace implications of collaboration network structure. And finally, several other recent studies of collaboration networks and innovation in organizations use data from the pharmaceuticals sector, which helps to facilitate comparisons between my findings and prior investigations (Grigoriou and Rothaermel, 2013; Guler and Nerkar, 2012).

3.4 Data and Methods

3.4.1 Sample

My hypotheses are targeted at the individual level of analysis. However, because organizational factors may have substantial influence over inventors' collaboration patterns and their performance at innovation, I begin my sampling strategy with firms. This approach helps to ensure that I only compare inventors who work at similar types of organizations and also allows me to carefully control for a range of other possible organizational confounders in my statistical models.

Following prior research on the pharmaceuticals industry, I relied on a variety of sources to identify appropriate firms, especially Wards Business Directory and the Security and Exchange Commission's EDGAR database. Using these sources, I created a list of companies that had 2834 ("Pharmaceutical Preparations") as their primary Standard Industrial Classification code, revenues of more than $100 million, and were publicly traded, at any point in time between 1975 and 2010 (Gerstner et al., 2013). With this list in hand, I then excluded companies that focused primarily on the manufacture of generic drugs or the creation of non-human products.

I then sought to link firms with data on patents and pharmaceutical products. There have been several merger waves in the pharmaceuticals industry over the past few decades, which makes it hard to track organizations over time and therefore complicates this matching process. For example, how should matching occur after the merger of two major firms? How can major, transformational acquisitions be distinguished from relatively minor ones that are less likely to influence collaboration patterns and innovation? Prior research offers a number of possibilities. To facilitate replication and maximize transparency, I adopted a fairly simple approach in which I treat a merger or acquisition among firms in the list I describe above as two deaths and one birth (i.e., the creation of a new, combined company). I ignore all other

mergers and acquisitions, as these are typically much smaller or are less relevant to pharmaceutical innovation. Using similar procedures, I also fold major subsidiaries into their respective parent organizations.

After implementing this procedure, I then matched companies with patent data taken from the U.S. Patent and Trademark Office (USPTO) and the Harvard Patent Dataverse (Lai et al., 2011). To identify pharmaceutical products, I linked the patents to drugs approved by the U.S. Food and Drug Administration (FDA) using the Approved Drug Products with Therapeutic Equivalence Evaluations reference, commonly known as the Orange Book (April 2014 edition).[6] First published in 1980, the Orange Book began listing patents after the 1984 passage of the Hatch-Waxman Act (Hemphill and Sampat, 2012).

Although the underlying data span several decades, I focus the statistical analyses on 1997 to 2001. Several factors led me to narrow the time period of the study. Many of the data sets I draw on are of higher quality in later years, and because this is not historical investigation, my goal is to characterize innovation in the pharmaceuticals industry as it is today, not as it once was. Both of these considerations suggest using more recent data. However, as I move closer to the present, I also increase the risk of censoring because I do not have data on patents and products that are still making their way through their respective approval processes. [7] An end year of 2001 achieves some level of balance between recency-censoring tradeoff. I chose 1997 as an appropriate start year because it is far enough from the beginning of my data that I am able to carefully control for several dimensions of inventor's past productivity, an important source of heterogeneity. Finally, I favor a relatively short study window because it helps minimize the effects of time trends, including industry and regulatory transformations.

[6] `http://www.fda.gov/Drugs/InformationOnDrugs/ucm129689.htm`, accessed May 16, 2014

[7] As a frame of reference, the mean duration of clinical and regulatory evaluation for FDA approved small molecule drugs between 1982 and 2001 was 7.6 years (Reichert, 2003).

After implementing all of the above criteria, my sample consists of 41,051 inventor-year observations among 18,668 inventors, across 37 pharmaceutical firms.[8]

3.4.2 Network Construction

I used a multistep procedure to map the collaboration networks of inventors and the broader firms in which they are embedded. As discussed above, patent collaborations serve as my proxy for network structure. One major challenge with patent data from the USPTO is that inventors are not given unique identifiers, and moreover, their names are often listed inconsistently across filings. To overcome this problem, I identify inventors and their collaborators using a unique record label for each inventor obtained with a probabilistic name matching algorithm, supplied by the Harvard Patent Dataverse.

Once the data are prepared, I then build the networks. Patent data have a bipartite structure, meaning that there are two types of nodes: actors (inventors) and events (patents). Inventors are not directly connected to one another—strictly speaking, they only have ties to patents. Interlocking corporate directorates are another commonly studied bipartite network, in which the actors are directors and the events are the boards on which they sit (Mizruchi, 1996). Because many measures are designed for unipartite networks, I follow prior work and project the collaboration network data such that inventors are directly connected to one another. After completing the projection process, I have a list of inventor-inventor dyads.

The final step in the process of building the networks is to situate each dyad in time. A collaboration between two inventors on a patent is an manifestation of a relationship that may persist into the future. To capture this idea, I approximate the structure of each inventor's network (and of broader collaboration networks that span their organizations) at any given point in time t using a sliding window that includes

[8]These analyses have been approved by the University of Michigan Institutional Review Board (study id: HUM00064545).

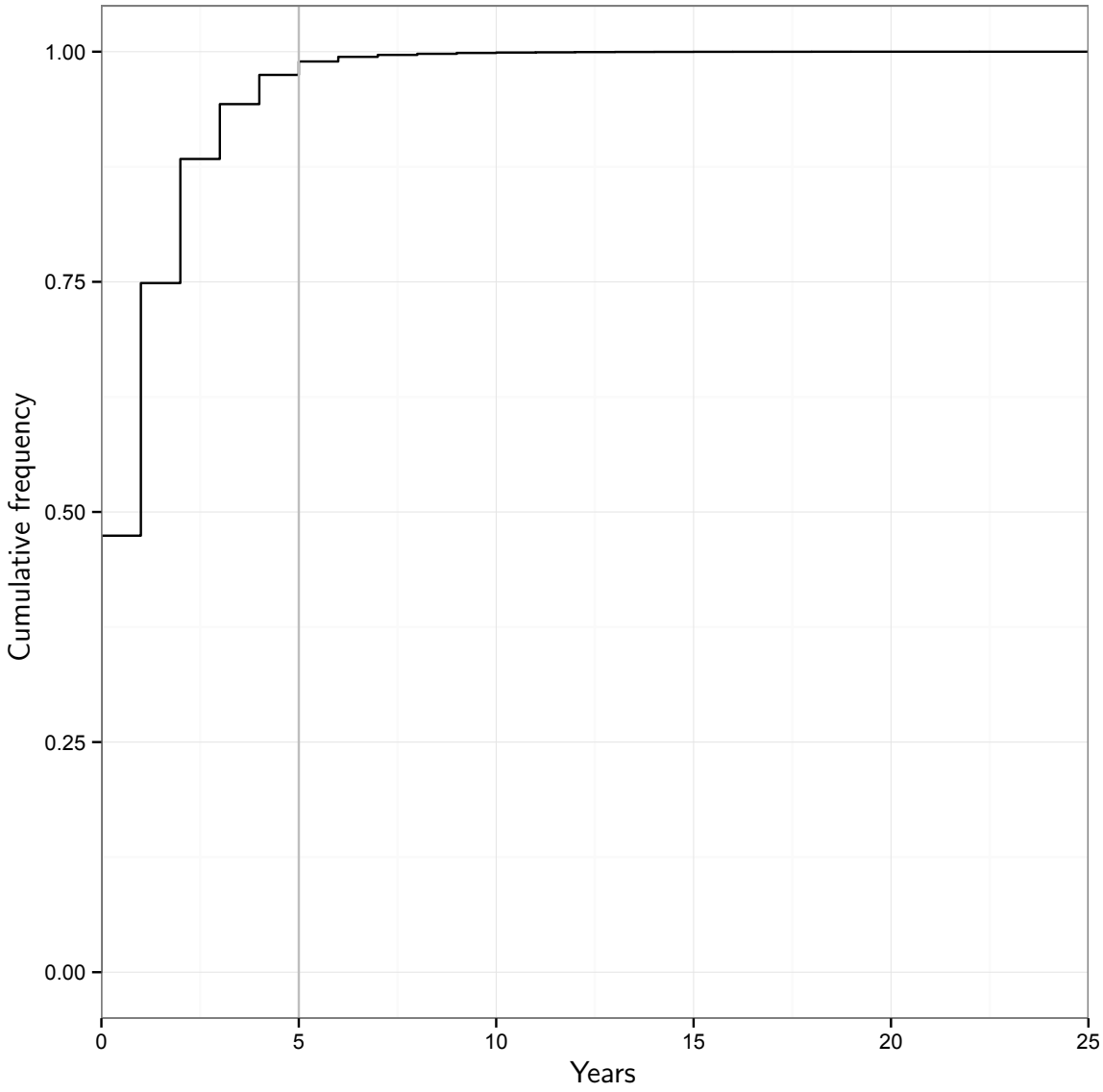

Figure 3.1: Years separating repeat collaborations. This plot helps to justify the use of a five-year sliding window filter for constructing the inventor networks, as almost all collaborations that will ever be repeated occur within that timeframe. To the extent that ongoing relationships among inventors manifest themselves in collaborations on patents, this window should be also relatively effective in pruning dead or dormant ties.

all collaborations that occurred over the past five years, in line with prior research. Ties that are older than five years are dropped from the network. Figure 3.1 helps to justify the use of a sliding window by showing the cumulative distribution of years separating repeat collaborations between pairs of inventors. As indicated by the dark vertical line, almost all collaborations that will ever be repeated occur within five years.

3.4.3 Study Design

A major challenge of linking a person's position in a social network to his or her performance is that position is rarely exogenous. Inventors are not randomly assigned to collaborators, and unobservable factors that guide who they choose to work with may also shape how well they do at creating innovations. By using a quasi-experimental design and a novel statistical modeling approach, I was able to gain some leverage on these challenges. I describe each in turn.

Sampling. Recall that Hypothesis 4 suggests that becoming a broker will have a positive effect on an inventor's performance. My strategy for testing this prediction is to find a subsample of inventors who became brokers during the study period and to compare their performance to similar inventors who did not make the transition. In the language of an experiment, the former inventors make up the treatment group, while the latter are controls. To identify appropriate cases for each group, I began by following prior work (e.g., Fleming et al., 2007) and breaking inventor's careers into units of three years, which become my observations in the statistical analyses described in greater detail below. I use data from the first year, $t-1$, to model an inventor's propensity to be exposed to the treatment (i.e., to become a broker). Data from the second year, t, then allow me to model an inventor's performance in the third year, $t+1$. After creating these three year career blocks, I built the treatment group by identifying cases where inventors were not brokers at $t-1$ but had become

one by t. Following this logic, inventors who were not brokers at $t-1$ and did not make the transition at t are eligible to be controls.

My approach to testing Hypothesis 5—which anticipates a negative effect of having a connection to a broker on performance—is analogous to my test of Hypothesis 4, but with the important difference that here, treatment is exogenous to performance. To fill the treatment group, I extracted career blocks where inventors had zero ties to brokers at $t-1$ and one tie at time t, but only where that tie came from an existing contact. Using this criterion, treatment is exogenous because the decision to connect was made prior to the contact occupying a broker role. To isolate potential controls, I identified inventors who had zero ties to brokers at both $t-1$ and t.

Figures 3.2 and 3.3 show hypothetical examples of treatment and control inventors for tests of Hypotheses 4 and 5, respectively.

Statistical modeling. A number of different statistical techniques are also available for aiding with the challenges of disentangling causality in observational data. I make use of a technique based on propensity score weighting (Rosenbaum and Rubin, 1983) that has several attractive features.

A propensity score is the probability of exposure to a treatment given a set of observable covariates, $\mathbf{X}$. Formally, this may be written as

$$e_i = P(T = 1|\mathbf{X}), \tag{3.1}$$

where T is a dummy variable that takes on a value of 1 for people exposed to treatment and 0 otherwise. The appropriateness of propensity scores for estimating treatment effects rests on several assumptions. Let Y_1 denote the potential outcome for those who for those who are exposed to the treatment (i.e., occupation of a particular network position) and Y_0 for those who are not exposed. The first assumption, then, is that

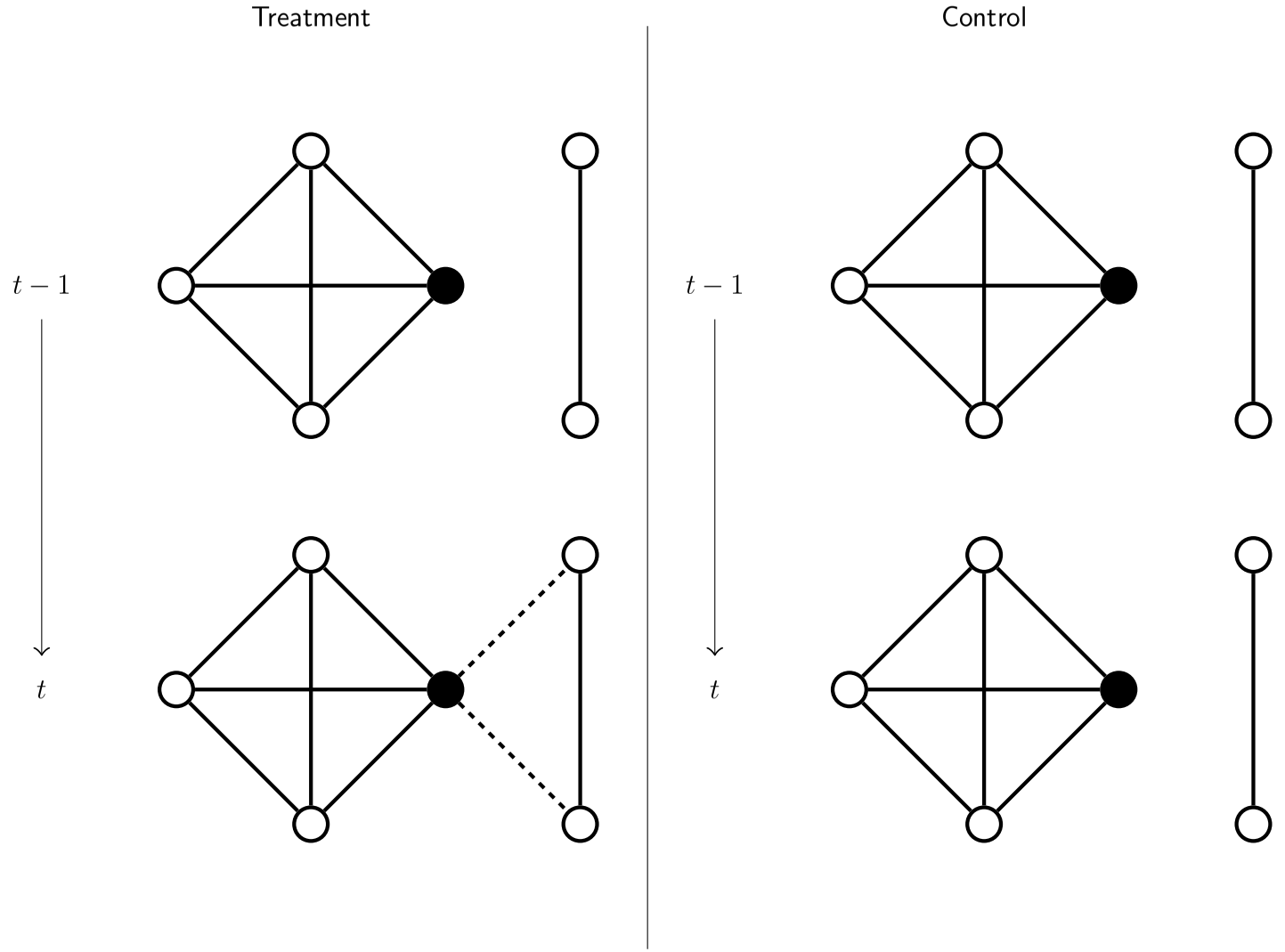

Figure 3.2: Estimating the effect of becoming a broker. Black nodes indicate the focal inventors for each experimental condition. In the transition from $t-1$ to t, the treated inventor becomes a broker by connecting to the pair of inventors that are not connected directly to any other members of the network.

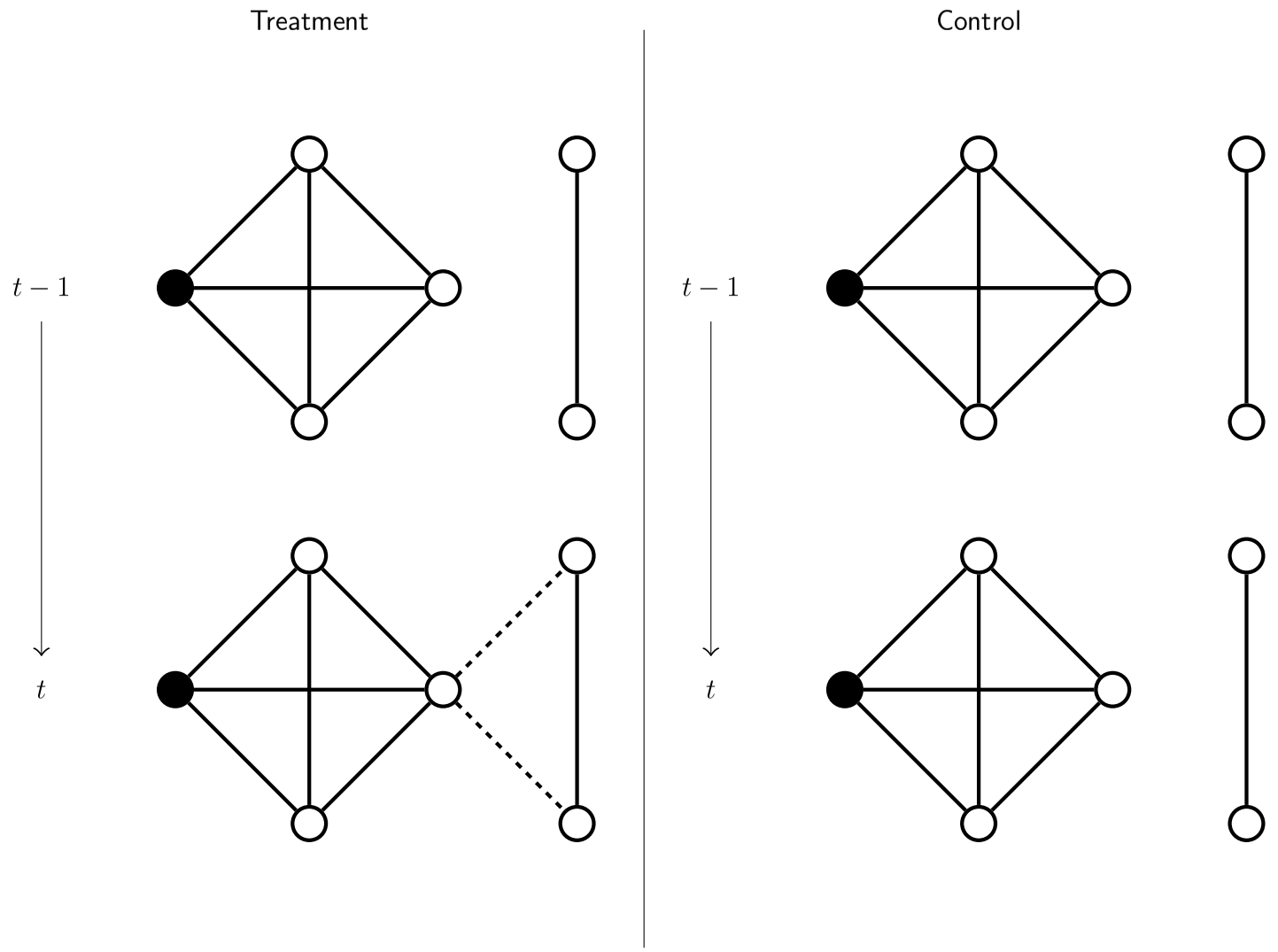

Figure 3.3: Estimating the effect of connection to a broker. Black nodes indicate the focal inventors for each experimental condition. In the transition from $t-1$ to t, the treated inventor becomes connected to a broker after an existing collaborator connects to the pair of inventors that are not connected directly to any other members of the network. Because the focal inventor only becomes connected to a broker through an existing collaborator, the treatment is exogenous to actions of the focal inventor.

$$(Y_1, Y_0) \perp\!\!\!\perp T|\mathbf{X}. \tag{3.2}$$

In words, potential outcomes Y_1 and Y_0 are independent of T, conditional on $\mathbf{X}$. The second assumption is that

$$0 < P(T = |\mathbf{X}) < 1, \tag{3.3}$$

such that all inventors have a nonzero probability of either treatment or non-treatment. When these two assumptions are met, treatment assignment is said to be strongly ignorable, and propensity scores may be used to obtain estimates of the treatment effect (Rosenbaum and Rubin, 1983).

Researchers typically estimate propensity scores with a logit or probit model that predicts treatment as a function of observable covariates. After obtaining these scores, they may be used to estimate the treatment effect by either stratifying and comparing observations or by weighting observations in a regression model of the outcome of interest (Morgan and Winship, 2007).

To obtain unbiased estimates of the treatment effects, the model for the propensity score must be correctly specified. Moreover, in the case of weighting, the quality of the estimates also hinges on getting the model for the outcome right. Recently, however, a new class of estimator has been proposed that eases some of this burden on the researcher by using information in both the exposure and the outcome models to estimate the effects of treatment. Known as "double robust," these estimators remain unbiased as long as either the model for the exposure or outcome (or both) are correctly specified. Following (Lunceford and Davidian, 2004), the estimator I use is

$$\hat{\Delta}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_i Y_i - (T_i - \hat{e_i})m_1(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_1)}{e_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-T_i)Y_i + (T_i - \hat{e_i})m_0(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_0)}{1-e_i}, \tag{3.4}$$

where $m_T(\mathbf{X}, \hat{\boldsymbol{\alpha}}_T)$ are the predicted values of different regressions for treated and control observations of $\mathbf{X}$ on the outcomes. This double robust estimator is similar in form to those that rely on inverse probability weights, but differs by augmenting with the regressions (Glynn and Quinn, 2010).

3.4.4 Dependent Variables

I consider three dependent variables, all of which serve as useful proxies for the oftentimes-elusive phenomenon of innovation. Consistent evidence across all three variables should add greater confidence to any findings.

Patents (weighted). The first variable is a count of patents applied for by inventor i at time $t+1$ that were assigned to his or her primary firm during the focal career period. I use patent application dates rather than grant dates because application dates usually correspond reasonably well to the time of invention, while grant dates are influenced by the duration of the USPTO evaluation process. Following Lee (2010), I weight the value of each patent by team size, which is simply the number of listed inventors, before summation. Because patents may vary widely in their quality, many authors recommend weighting raw patent counts by the number of future citations that each patent receives (Trajtenberg, 1990). I found broadly similar results when using citation weights, but those models were also less tractable due to excess zeros and extreme outliers.

Products (weighted). The second dependent variable helps to address potential concerns about patent quality. Like the first measure, it is a count of patents applied for by inventor i at time $t+1$. Once again, I include only patents that were assigned

to the inventor's primary firm during the focal career period, and I weight by team size. The unique feature of this measure is that I only include patents that ultimately contributed to an FDA approved drug. As noted above, I obtain these data from the April 2014 edition of the FDA's Orange Book, which lists patents associated with approved drug products under provisions of the 1984 Hatch-Waxman Act.

Products. Finally, because very few inventors apply for patents that ultimately contribute to an FDA approved drug product, especially in a single year, I also consider a version of the products measure without the weight for team size. This helps to account for the possibility that simply being involved with the invention of a successful drug product matters more than how many other inventors contributed to the product's creation.

3.4.5 Independent Variables

Broker. Prior research uses many different measures of brokerage. One of the most widely known is Burt's (1992) measure of network constraint, which captures, on a continuous scale, the extent to which a person's direct contacts are interconnected with one another and therefore are redundant. Although attractive for some purposes, I chose not to use this measure because it does not offer a clear and objective way of identifying when a person is a broker, which is essential for testing my hypotheses.

Instead, I created a dummy variable, broker, that I set to 1 if, at time t, an inventor's immediate network neighborhood (i.e., the other inventors to which he or she is directly connected and their connections among one another) breaks into two or more discrete, disconnected segments (or components) upon the inventor's removal from the network. This measure is attractive because it allows me to clearly and unambiguously identify "treated" inventors (i.e., those who become brokers) from potential controls, and in so doing, to apply statistical techniques that are helpful for disentangling causality in quasi-experimental, pretest–posttest study designs. And

more importantly, it corresponds exactly to the definition of brokerage offered at the beginning of this chapter—i.e., it identifies inventors who are connected to people who are not directly connected to one another.

Connection to a broker. After identifying brokers, I created the second independent variable, connection to a broker, which takes on a value of 1 if an inventor has a tie to a broker at time t, and 0 otherwise. Note that the value of this variable will be 1 regardless of the number of brokers to which the focal inventor is connected (as long as they are connected to at least 1). As I discuss later, when creating the matched treatment and control data sets, I focus on inventors who were treated by a connection to only one broker to ensure comparability of effects across cases. This measure offers a clearer indication of having a connection to a broker than those of related studies by Burt (2007, 2010) and Galunic et al. (2012), discussed earlier, who used the average constraint and network density of direct contacts, respectively.

3.4.6 Control Variables

I include a variety of control variables to help account for other factors that may influence an inventor's propensity to become a broker (or to have a connection to one) and his or her performance at innovation. Unless otherwise noted, all variables are included in both the exposure (i.e., propensity score) and outcome equations. Control variables that appear in the outcome equation are measured at time t; those appearing in the exposure model are measured at time $t - 1$.

Degree centrality. My hypotheses place a heavy emphasis on the structure of connections among collaborating inventors. A simpler explanation, however, is that what really matters is having connections and that the structure among them is irrelevant. With this possibility in mind, I include a control for each inventor's degree centrality, which is a count of the number of other inventors to which he or she is directly connected. This variable is also relevant for predicting exposure. People with

more connections may have better interpersonal skills that help them form strategic partnerships and build better network positions. Additionally, having more partners also increases the likelihood of having a connection to a broker.

Clustering (local). I also control for the level of clustering in each inventor's network neighborhood (i.e., among his or her direct contacts), defined as the ratio of closed triangles (i.e., sets of three inventors with three ties) to connected triples (i.e., sets of three inventors with two ties) running through each inventor. Clustering is often used as a measure of cohesion, which prior studies link to innovation (Fleming et al., 2007). The cohesiveness of an inventor's network neighborhood may also relate to his or her propensity to become a broker or to have a connection to one—to the extent that an inventor's contacts are already highly interconnected, it may be harder to establish ties to disconnected people (or to have a connection to someone who does).

Main component. Many of the intraorganizational networks in my sample of pharmaceutical firms contain multiple components, where a component is a discrete set of inventors such that there is some (potentially indirect) pathway connecting each pair of inventors. Typically, there was also one component that was substantially larger than all others, which I call the main component. Prior research suggests that location in the main component of a network may influence innovation (Owen-Smith and Powell, 2004). Membership may also relate to an inventor's propensity to become a broker or to have a connection to one, since there are likely more potential collaborators. Given these considerations, I include a dummy variable that takes on a value of 1 if the inventor is a member of the largest connected component and 0 otherwise.

Community size. Inventors may differ substantially in their level of access to potential collaborators, even within connected components (Sytch et al., 2012). As a way of accounting for these differences, I control for the size of each inventor's

network community. A network community is a set of nodes that have many dense interconnections among one another, but relatively few connections to other nodes in the network. There are many algorithms available for partitioning networks into communities. I use the Infomap algorithm (Rosvall and Bergstrom, 2008). Infomap partitions a network into communities by trying to minimize the description length of the path of a random walker traversing the network. After running the algorithm, I compute the measure of community size as the number of other inventors in each inventors' assigned network community at $t-1$. I include this measure only in the exposure model because it relates primarily to opportunities become a broker or to have a connection to one.[9]

Patent stock. Prior research suggests that high-performing inventors may be better able to identify opportunities and and attract collaborators that allow them to occupy brokerage positions (e.g., Lee, 2010), in which case some (or all) of the association between becoming a broker (or having a connection to one) may be driven by some underlying difference in ability. To help account for this possibility, I include a measure of the size of each inventor's portfolio of patents. Because older patents may be less reflective of an inventor's current ability, I use a use a depreciated stock model, defined as

$$PS_{it} = \sum_{\tau=0}^{t} \Big(1-\delta\Big)^{t-\tau} K_{i\tau}, \tag{3.5}$$

where K is the set of patents applied for by inventor i at time t (or $t-1$ for the exposure equation) and δ is a constant, set to 0.15, that regulates the annual rate of depreciation (Azoulay et al., 2007). I include patents in the calculation going back as far as 1975 (when my data begin) but only if they were assigned to the focal inventor's primary firm for the sampled career period.

Career experience. I also include a control analogous to the one for patent

[9]The substantive results are not sensitive to the addition of this or other variables that appear only in the exposure model to the the outcome equation.

stock, but that only sums patents if they were *not* assigned to the focal inventor's primary firm for the sampled career period. This variable helps further account for differences in ability among inventors. More importantly, however, it should also help to differentiate people who are highly mobile or perhaps have only temporary affiliation with their primary focal firm, and that as a result may have systematically different patterns of connectivity.

Technological diversity. Inventors who are active in multiple fields could have greater exposure to diverse information, which may enhance their innovation, while also increasing their likelihood of working with otherwise disconnected collaborators. To account for this possibility, I control for the technological diversity of each inventor's patent portfolio as the Herfindahl index of the patents' primary classes. I include only patents that inventors applied for over the past five years, and that were assigned to his or her primary firm for the focal career period. I subtract the result from 1 so that inventors with more diverse portfolios have higher values.

Clustering. Firms may differ in the degree to which their intraorganizational collaboration networks offer opportunities for inventors to work with otherwise disconnected contacts and therefore become brokers (Sytch et al., 2012). Several factors likely shape the overall level of opportunity, but one that may be especially important is the network's cohesiveness. As an intraorganizational network becomes more cohesive (and therefore inventor's collaborators tend to also collaborate with one another), opportunities should decrease. With these considerations in mind, I control for the global level of clustering in each intraorganizational network, as of time $t-1$. The clustering coefficient is defined as

$$CC_{it} = \frac{3N_{\triangle}}{N_{\vee}} = \frac{3 \times (\text{number of triangles})}{(\text{number of connected triples})}, \tag{3.6}$$

where a triangle is a closed triad and a triple is an open triad, and higher values indi-

cate more cohesive networks. Because this measure relates primarily to an inventor's propensity to become a broker or to have a connection to one, I only include it in the model for exposure.

Communities (global). I control for the total number of communities in each firm's intraorganizational network at time $t-1$ as an additional way of capturing differences in opportunities to become a broker (or to have a connection to one). In line with my other controls for opportunity, I only include this variable in the exposure equation.

R&D spending. Inventor's firms may also differ in the extent to which they have an interest and ability to support brokerage activity and innovation at a given point in time. To help account for these potential differences in strategy and resources, I control for each firm's annual costs incurred while developing new products or services, in millions of dollars. This measure, along with my other financial controls, uses data from Compustat.

EBIT. To capture more general differences in current resources, I also control for each firm's earnings before interest and taxes, in millions of dollars.

Total assets. As a final measure of financial health, I also control for each firm's current assets, once again in millions of dollars.

Firm fixed effects. Firms may differ in other ways not captured by the controls above, but that nevertheless shape both inventors' propensities to become or have a connection to a broker and their performance at innovation. Therefore, I include dummy variables for 36 of the 37 sample firms. These fixed effects control for all firm-level differences that do not change over the study period.

Year fixed effects. Finally, to help account for various temporal differences and potential right censoring in counts of FDA approved drug products, I include dummy variables for 5 of the 6 study years.

Variable summaries and descriptive statistics are shown in Tables 3.1 and 3.2,

Table 3.1: Variable Names and Definitions

Name	Definition	Equation(s)
Dependent Variables		
Patents (weighted)	Count of patents applied for by each inventor at $t+1$, weighted by team size	Outcome
Products (weighted)	Count of patents applied for by each inventor at $t+1$ that contributed to an FDA approved drug, weighted by team size	Outcome
Products	Count of patents applied for by each inventor at $t+1$ that contributed to an FDA approved drug	Outcome
Independent Variables		
Broker	Dummy variable; 1 if the inventor's immediate network neighborhood breaks into two or more components upon the inventor's removal from the network at t, 0 otherwise	Propensity score
Connection to a broker	Dummy variable; 1 if the inventor has a connection to a broker at t, 0 otherwise	Propensity score
Controls—Inventors' Networks		
Degree centrality	Count of the number of fellow inventors to which each inventor is directly connected at $t-1$ (propensity score equation) or t (outcome equation)	Outcome and propensity score
Clustering (local)	Ratio of closed triangles to connected triples running through each inventor at $t-1$ (propensity score equation) or t (outcome equation)	Outcome and propensity score
Main component	Dummy variable; 1 if the inventor is a member of the largest connected component at $t-1$ (propensity score equation) or t (outcome equation), 0 otherwise	Outcome and propensity score
Community size	Count of the number of other inventors in each inventors' network community at $t-1$	Propensity score
Controls—Inventors' Experience		
Patent stock	Cumulative patents applied for by each inventor while at the focal firm as of $t-1$ (propensity score equation) or t (outcome equation), depreciated 15% annually	Outcome and propensity score
Career experience	Cumulative patents applied for by each inventor while not at the focal firm as of $t-1$ (propensity score equation) or t (outcome equation), depreciated 15% annually	Outcome and propensity score
Technological diversity	Herfindahl index of primary classes for all patents filed by each inventor while at the focal firm over past five years, as of $t-1$ (propensity score equation) or t (outcome equation), subtracted from 1	Outcome and propensity score
Controls—Firm's (Internal) Network		
Clustering	Ratio of closed triangles to connected triples in each firm's inventor network at $t-1$	Propensity score
Communities (global)	Count of network communities in each firm's inventor network at $t-1$	Propensity score
Controls—Firm's Resources		

Table 3.1 (Continued)

R&D spending	Sum of each focal firm's annual costs incurred relating to developing new products or services at $t-1$ (propensity score equation) or t (outcome equation) (in millions of dollars)	Outcome and propensity score
EBIT	Sum of each focal firm's earnings before interest and taxes at $t-1$ (propensity score equation) or t (outcome equation) (in millions of dollars)	Outcome and propensity score
Total assets	Sum of each focal firm's current assets at $t-1$ (propensity score equation) or t (outcome equation) (in millions of dollars)	Outcome and propensity score
Firm fixed effects	Dummy variables for 36 of the 37 sample firms	Outcome and propensity score
Controls—Other		
Year fixed effects	Dummy variables for 5 of the 6 study years	Outcome and propensity score

Table 3.2: Descriptive Statistics and Correlations[†]

	Variable	Mean	SD	Min	Max	1	2	3	4	5	6	7
1.	Patents (weighted)	0.16	0.37	0.00	21.00	1.00						
2.	Products (weighted)	0.00	0.04	0.00	1.25	0.13	1.00					
3.	Products (count)	0.01	0.13	0.00	3.00	0.10	0.85	1.00				
4.	Broker	0.13	0.34	0.00	1.00	0.19	0.03	0.03	1.00			
5.	Degree centrality	6.19	6.40	0.00	85.00	0.17	−0.01	0.01	0.24	1.00		
6.	Clustering (local)	0.73	0.37	0.00	1.00	−0.19	−0.04	−0.01	−0.35	−0.01	1.00	
7.	Main component	0.52	0.50	0.00	1.00	0.05	−0.01	0.00	0.13	0.36	0.09	1.00
8.	Community size	11.69	10.72	1.00	83.00	0.02	−0.03	−0.02	0.00	0.54	0.13	0.48
9.	Patent stock	2.38	3.36	0.44	99.97	0.47	0.03	0.04	0.31	0.62	−0.21	0.28
10.	Career experience	1.48	3.87	0.00	102.54	0.03	0.00	0.00	0.02	0.01	0.02	−0.13
11.	Technological diversity	0.14	0.23	0.00	0.87	0.16	0.00	0.01	0.36	0.36	−0.17	0.21
12.	Clustering (global)	0.59	0.11	0.00	1.00	−0.02	0.01	0.02	−0.06	−0.08	0.06	−0.33
13.	Communities (global)	123.96	53.98	1.00	219.00	−0.06	−0.03	−0.03	−0.02	0.04	0.06	−0.13
14.	R&D spending	1712.40	1001.74	9.18	4435.00	0.03	−0.01	−0.02	−0.01	−0.02	0.01	−0.17
15.	EBIT	3358.01	2518.62	−198.25	9089.10	0.00	−0.01	−0.01	0.00	0.07	0.01	0.02
16.	Total assets	21000.48	14283.45	11.93	44069.30	0.01	−0.02	−0.03	−0.04	−0.01	0.04	−0.22
17.	Year	1999.04	0.81	1998.00	2000.00	0.00	−0.01	−0.01	−0.01	0.04	0.02	0.01

	Variable	8	9	10	11	12	13	14	15	16	17
8.	Community size	1.00									
9.	Patent stock	0.29	1.00								
10.	Career experience	−0.06	0.01	1.00							
11.	Technological diversity	0.17	0.38	0.01	1.00						
12.	Clustering (global)	−0.21	−0.16	0.13	−0.08	1.00					
13.	Communities (global)	0.07	−0.07	−0.01	0.00	0.23	1.00				
14.	R&D spending	0.04	−0.04	0.01	−0.06	0.11	0.60	1.00			
15.	EBIT	0.20	0.03	−0.07	−0.01	−0.14	0.45	0.74	1.00		
16.	Total assets	0.06	−0.06	0.04	−0.05	0.20	0.73	0.83	0.61	1.00	
17.	Year	0.04	0.00	0.02	−0.01	0.01	0.16	0.17	0.13	0.15	1.00

[†] $N = 41,051$ inventor-years

respectively.

3.5 Results

Table 3.3 presents estimates of the causal effect of becoming a broker (hereafter, the "broker test") and of having a connection to a broker (hereafter, the "connection test") on the three different proxies for innovation. Before I discuss the findings, it is worth noting that the size of the treatment groups across both tests are nearly identical, but there were fewer relevant control cases in the connection test. Nevertheless, the treatment and control groups are comparable on observable covariates in both tests after propensity score weighting.

The first row of estimates in Table 3.3 test Hypothesis 4, which predicts that becoming a broker will have a positive effect on an inventor's performance at innovation. Across all three dependent variables, becoming a broker has a significant, positive effect on performance, in strong support of the prediction. The effect is also large in substantive terms. Becoming a broker is expected to result in an increase of 0.206 weighted patents. That may seem like a small effect, but for perspective, note that the increase implies a 130.2% performance boost relative to the mean value of of 0.158 weighted patents. The corresponding numbers for weighted products and counts of products are also large—and surprisingly, almost identical to the relative performance boost for weighted patents—with values of 129.9% and 131.7%, respectively.[10]

Hypothesis 5 predicts that in contrast to being a broker, having a connection to one will negatively influence performance. In support of this idea, the values in the second row of estimates in Table 3.3 are all negative, and statistically significant. As one might expect, the effects are somewhat smaller in substantive terms, but still very notable. Specifically, connection to a broker leads to a decrease of 0.019

[10]Because I am using more precise values, my percentages may differ slightly from those that would be obtained using the data presented in Tables 3.2 and 3.3.

Table 3.3: Effects of Brokerage on Innovation[†]

	Patents (Weighted)	Products (Weighted)	Products (Counts)	Sample Size	
				Treatment	Control
Becoming a broker	0.206 ∗ ∗	0.005 ∗ ∗	0.018∗	953	5,939
	(0.098)	(0.002)	(0.009)		
Connection to a broker	−0.019 ∗ ∗	−0.003 ∗ ∗	−0.006∗	959	2,466
	(0.009)	(0.001)	(0.003)		

* $p < 0.1$, ** $p < 0.05$; two tailed tests.

[†] Robust standard errors are in parentheses.

in weighted patents, which implies a performance hit of roughly 12.0% relative to the mean. Interestingly, the effects are much greater for the pharmaceutical product measures; the estimates suggest performance hits of roughly 66.6% and 41.8% for weighted products and counts of products, respectively.

The large differences in the relative effects of connection to a broker for patents versus products may shed some light on the mechanism by which connection to a broker can harm an inventor's performance. For example, it is possible that brokers do pass information onto their disconnected contacts, which helps to explain the relatively small negative effect on patents. However, that information may be of lower quality, and therefore lead an inventor to make fewer contributions to drug products. Moreover, the finding also suggests that part of what connection to a broker does is limit the ability of inventors to implement their ideas, which in turn emphasizes the possibility of opportunistic behavior on the part of brokers, the costs of time and energy necessary to maintain a relationship with a broker and other contacts, and related factors.

3.6 Discussion and Conclusion

Over the past few decades, the concept of brokerage has captured the imagination of network researchers, and for good reason. Reflecting on this body of work, Reagans and Zuckerman (2008, 797) write, "it is hard to find a more precise and influential sociological theory." Even more noteworthy than its theoretical appeal are the mountains of evidence that show brokerage helps explain why people differ in many important outcomes, from performing well at work to being more creative. Given this level of interest, it is surprising that little research considers the effects of brokerage for the disconnected parties. Our lack of knowledge in this area is even more troublesome because existing theory also hints that brokers may sometimes do better at the expense of those contacts.

In an effort to develop better knowledge about the potentially negative effects of connections to brokers, I conducted an investigation of intraorganizational collaboration networks and innovation using data on more than 18,000 inventors at 37 pharmaceutical firms. I focused my investigation around two hypotheses. To establish a baseline for this study population, my first prediction was that becoming a broker would have a positive effect on an inventor's performance at innovation, a conjecture that is closely aligned with prior work. I then developed arguments to support my second hypothesis, which anticipated a negative performance effect of having a connection to a broker. I took several steps to help ensure that my tests of these hypotheses could be given a causal interpretation. First, I made use of propensity score weighting and a pretest–posttest framework in combination with a double robust estimator. This estimator is attractive because it is unbiased if either the model for exposure or the outcome (or both) are correctly specified. Second, when testing the effects of connection to a broker, I was able to focus on changes among existing contacts, where the decision to connect is exogenous to performance because it was made prior to the contact becoming a broker. And finally, I used three different proxies for innovation. My findings strongly supported both predictions. While becoming a broker was beneficial for an inventor's performance, the opposite was true for having a connection to one.

It is important to evaluate these findings in light of several limitations. My inferences about network structure rely on administrative data from the USPTO. Although my own interviews and those in published accounts (Fleming et al., 2007) suggest that these data offer useful representations of real collaboration networks, they may miss relevant connections that stem from work on projects that are not subject to patent protection. Moreover, both my patent and product data may be influenced by strategic considerations on the part of firms that have nothing to do with the underlying quality of the ideas or inventor relationships. I have attempted to

mitigate these concerns to some degree by focusing on a sector where patents are used routinely to protect intellectual property, and moreover, where there is a reasonably close correspondence between patents and actual (or potential) products. Additionally, I also sought to account for differences in firm strategy statistically, through the use of fixed effects and time-varying financial controls. A final limitation is that my claims of causality must be viewed with caution. Although the double robust estimator gives me two opportunities to correctly specify my models, the results may be biased if both of my attempts missed the mark.

Notwithstanding these limitations, my findings suggest several theoretical implications and directions for future research.

Brokers and individuals. Perhaps the most notable result of this study is its demonstration of a dark side of network brokerage, at least for innovation. All else being equal, having a connection to a broker seems to harm performance. Although this finding is consistent with observations about the control benefits of brokerage and the potential for opportunism, it is also surprising, especially in light of several qualitative studies that find evidence of a supportive role for network brokers (Hargadon and Sutton, 1997; Lingo and O'Mahony, 2010; Obstfeld, 2005). An important direction for future work, then, will be identifying cases where all else is not, in fact, equal—i.e., where personality, task, or other factors may moderate the negative effect of connection. Studies of this type would also be especially valuable for establishing the mechanisms by which ties to brokers may harm performance.

Brokers and organizations. Understanding the effects of brokers on their contacts should have special salience for organizations. To the extent that brokers gain advantages at the cost of others in their organizations, their behavior may lead to a network analogy to the tragedy of the commons (Hardin, 1968; Ostrom, 1990), in that by pursuing their own self-interest, brokers undermine the effectiveness of their larger organizations (Ibarra et al., 2005).

Table 3.4: Being and Broker and Connection to Brokers[†]

	Ties to brokers		
Broker	Yes	No	Total
Yes	4,681	712	5,393
No	22,825	12,833	35,658
Total	27,506	13,545	41,051

[†]Observations are inventor-years.

To help see why a better understanding of how brokers effect their contacts is so important for organizations, turn to Table 3.4, which cross tabulates my indicators of being a broker and having a connection to one for inventors in my sample. There are a total of 41,051 inventor-year observations reported in the table. Notice that being a broker is relatively rare, and occurs in only 5,393 of the observations, roughly 13.1% of the total. By contrast, having a connection to a broker is common, and accounts for some 27,506 cases. And, in nearly 83% of those cases, the inventor was not also a broker in the same year. What these numbers suggest is that first, knowledge of how and why having a connection to a broker influences performance is potentially of importance to a tremendous number of inventors, not a relatively marginal group. But more fundamentally, the distributions also highlight the need for future work that considers whether there may be an optimal balance within an organization between the number of brokers and the number of people who have connections to them.

Furthermore, recall that the findings from my inferential models suggest that becoming a broker leads to a roughly 130% increase in performance. This estimate is large by any standard and is also notable in comparison to the performance hits of around 12–67% that I observed for brokers' disconnected contacts. One possibility, then, is that from an organizational perspective, the superior performance of brokers outweighs their potentially negative effects. However, the relatively small number of brokers and large number of inventors who have a connection to one, as indicated by

Table 3.4, complicates this possibility. Specifically, it suggests that either the collective negatives may be greater than the collective positives, or that real organizations may struggle finding and maintaining the right balance between the two.[11]

From a broader theoretical point of view, these considerations on the individual and collective effects of brokerage also bear on an issue at the heart of two very different perspectives on social capital. Authors like Coleman (1988, 1990), Putnam (2000), and Portes and Sensenbrenner (1993) emphasize the collective benefits of collective social structures like communities. In this framework, cohesive ties among group members result in public goods like trust and monitoring. By contrast, Burt (1992) and followers emphasize the individual benefits of occupying particular types of positions in a (usually much smaller) social structure, where relationships are often competitive in nature. Little prior work has sought to determine whether these two forms of social capital are orthogonal or if they exist in fundamental opposition, such that increases on one leads to decreases in the other (c.f., Ibarra et al., 2005). Nevertheless, empirical investigations that build on my findings and explore the collective implications of brokers would help to address this issue while also allowing deeper integrating between these two important theoretical perspectives.

Evidence on the effects of context. As I note above, future research on personality, task, or other factors that differentiate among individuals will add to a deeper understanding of why having connections to brokers appears to harm performance and should also help reveal whether there are instances when these sorts of connections may be valuable. But larger contextual factors may also matter, and some evidence from this study and existing research suggests the possibility that the use of brokers—and therefore their effects—vary by organization.

To see this, consider Figures 3.4 and 3.5, which show the predicted probabilities of becoming a broker and having a connection to one, respectively, for sample

[11] Simulation may be an especially fruitful way of testing this idea in future work.

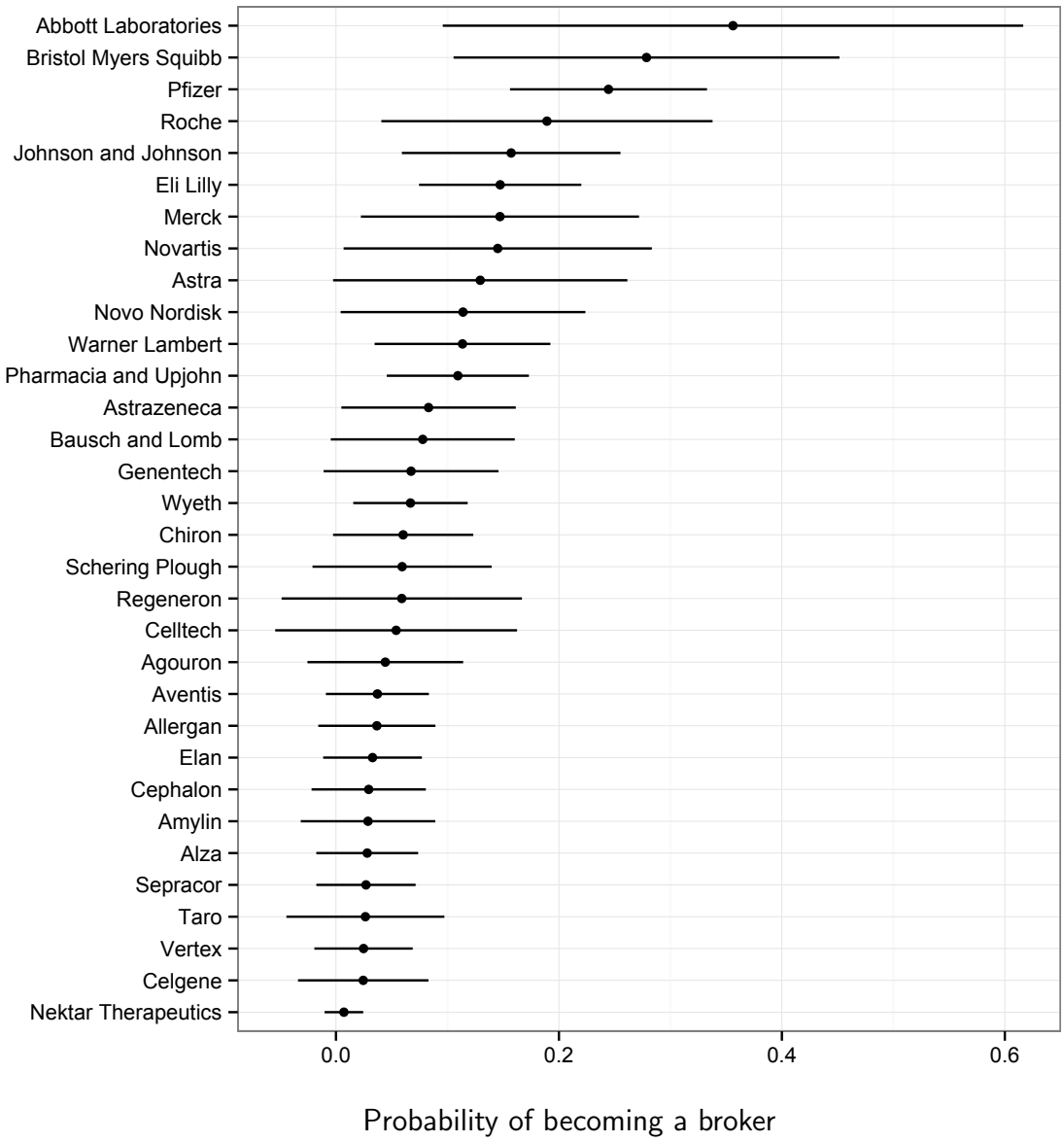

Figure 3.4: Probabilities of becoming a broker at different firms. Horizontal lines span the 95% confidence intervals for each point estimate. Values are taken from the logistic regression used to estimate the exposure (i.e., propensity score) model.

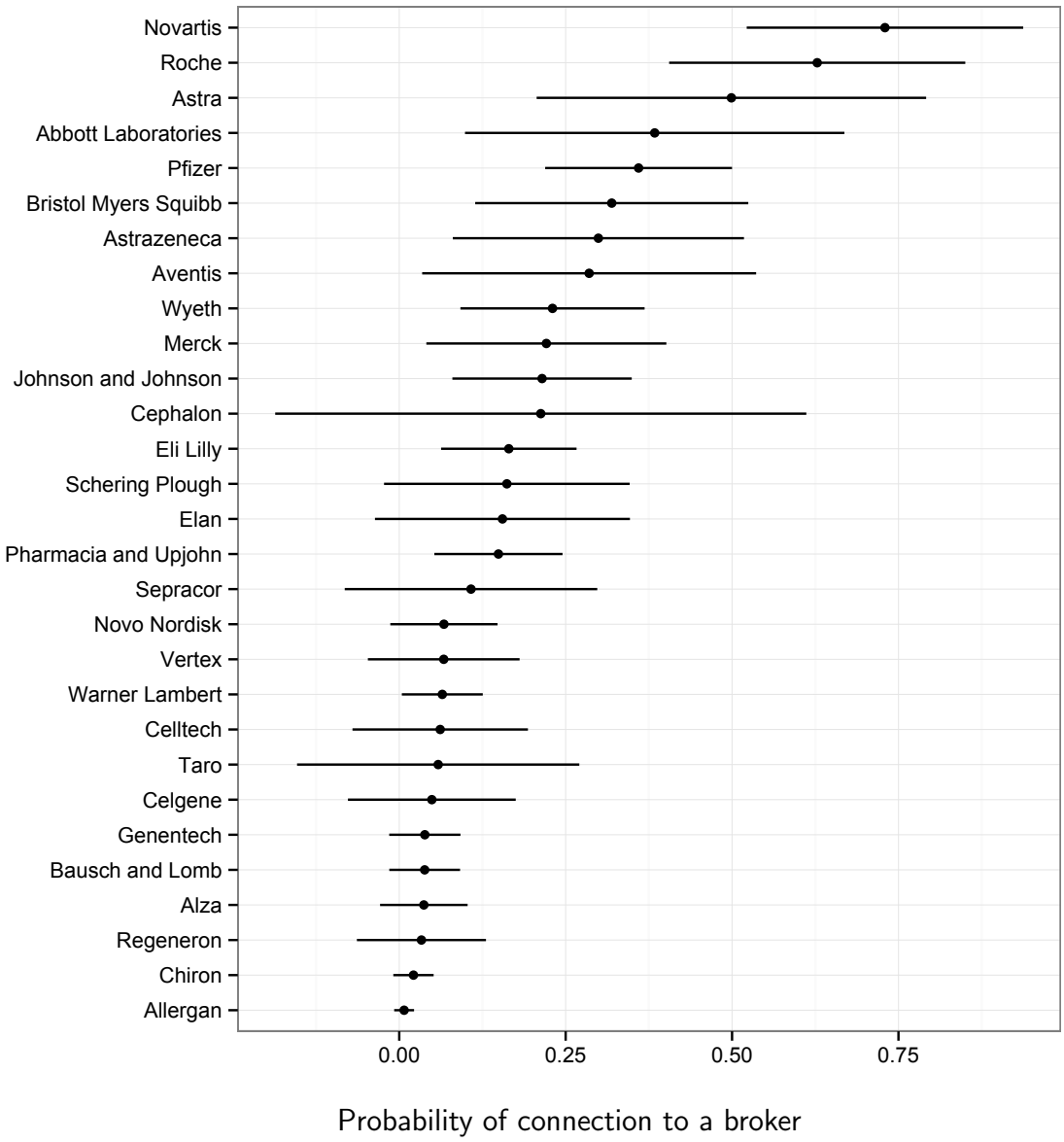

Figure 3.5: Probabilities of connection to a broker at different firms. Horizontal lines span the 95% confidence intervals for each point estimate. Values are taken from the logistic regression used to estimate the exposure (i.e., propensity score) model.

inventors, stratified by pharmaceutical firm.[12] These estimates come from their respective exposure (i.e., propensity score) models, and therefore hold constant an array of inventor- and firm-level differences. Several features of these plots are noteworthy. First, supporting the possibility of organizational differences, there is a fair amount of spread across in the predicted probabilities of both becoming a broker and having a connection to one across firms, although it is important to keep in mind that many of these differences are not statistically significant, as indicated by the overlapping 95% confidence intervals. Second, notice that in both figures, large, highly diversified pharmaceutical companies generally appear closer to the top (i.e., they have the largest probabilities on each outcome), while those that are smaller and more specialized appear near the bottom (i.e., they have the smallest probabilities). A number of factors could account for these differences in relative position. For instance, brokers may help facilitate search in large, diversified organizations, in which case there may be some positive returns to connection. A different possibility, perhaps more consistent with my findings, is that these firms simply have more internal mobility across disparate areas, which creates organizational "misfits" (Kleinbaum, 2012) who have many disconnected contacts. Finally, it is also worth noting that although there is some correspondence in the relative position of firms in Figures 3.4 and 3.5, there are also interesting deviations. Inventors at Baush and Lomb, for instance, a supplier of eye products, are well above the median with respect to their probability of becoming a broker, but they are among the lowest in terms of having a connection to one. Future research may benefit from further investigation into whether these sorts of deviations or other contextual factors moderate the effects of brokerage.

The first chapter of this dissertation also offers suggestive evidence that under some conditions, brokers may offer collective benefits. Recall that one goal of that study was to better understand what geographically isolated firms might do to help

[12]Firms only appear in either plot if they contributed an inventor that appeared in the estimation sample.

ensure their inventors are exposed to diverse ideas and perspectives. The results suggested that one important factor is network configuration. Specifically, I found that geographically isolated firms had better performance at innovation when the structure of internal collaboration networks among their inventors was more inefficient, with longer path lengths and therefore, more opportunities for inventors to serve as brokers by spanning disconnected network areas. By contrast, I also found that as firms became more geographically proximate to industry peers, network configurations that offer more opportunities for brokerage were harmful for performance.

A few other existing studies also offer indirect evidence that the collective effects of brokers may vary across organizations and even within organizations over time. For example, in a computational experiment, Lazer and Friedman (2007) examine how network structure affects the information diversity of an organization in which people (or agents, in their model) are looking for solutions to a complex problem. The authors find that less connected networks—in their case, a linear network where each person is has two neighbors—preserve diversity longer than ones that are fully connected. In the former, communication is inefficient, and therefore potential solutions derived from local knowledge are probed more thoroughly—people are less likely to converge prematurely on a local optimum. Although not directly addressed in the study, networks that exhibit greater opportunities for brokerage share some properties of a linear network. In fact, with the exception of people at either end of the line, everyone in a linear network has two disconnected contacts and is therefore by definition a broker. Extrapolating from this computational experiment, it may be that having more brokers in an intraorganizational network is good for collective performance if the goal is to facilitate exploration.

As one final example, consider Burt's (2007) study of secondhand brokerage, discussed earlier. In a closing discussion, Burt compares the association between performance and brokerage among contacts at varying degrees of separation (e.g., direct

contacts, direct contacts' contacts, and so on). The general pattern is one of exponential decline, such that increasingly indirect connections to disconnected contacts are less relevant for performance. What is more interesting, however, is that the rate of change differs across the study populations. Although it is hard to draw inferences given the small sample size (Burt has only three organizations) and other uncontrolled differences across the populations, this finding further suggests the need for deeper investigation into organizational factors.

Managing innovation. Finally, this study also has implications for managing innovation. At a minimum, my findings suggest the need for awareness that in some cases, rather than creating value, brokerage may actually be redistributing value from other members of a team or other parts of an organization. Therefore, when allocating rewards for contributions, it may be helpful to more carefully monitor or consult with brokers to get a better picture of the potentially less apparent contributions of their colleagues. Put differently, people who connect extensively with brokers may have lower than expected performance, but the reason is that they add value by helping the brokers do better. More broadly, my findings suggest that the indiscriminate use of brokers may stifle collective efficacy. Although promoting brokerage may be helpful in some cases, managers should consider doing so with caution and perhaps moderation when projects require close coordination and teamwork.

CHAPTER IV

How Knowledge Categorization Systems and Evaluation Norms Enable and Constrain Network Change in Organizations

4.1 Introduction

Two findings unite much contemporary scholarship on social networks in organizations. First, relationships among the members of an organization form networks that enable those individuals to achieve better outcomes. For example, diverse collaboration networks help people be more innovative by exposing them to heterogeneous ideas and perspectives (Burt, 2004; Tortoriello and Krackhardt, 2010). Highly connected individuals often work more efficiently because they have better access to knowledge about the capabilities of others within their organization and know where to look for assistance (Singh et al., 2010). And beyond influencing the outcomes for individuals, recent work also shows that the global structure of relationships among the members of an organization is an important feature predicting the organization's performance (Funk, 2014; Guler and Nerkar, 2012; Lazer and Friedman, 2007).

Second, findings from an array of studies reveal that the effects of networks on performance are contingent and context dependent. Much like students of organizational design emphasize that "there is no best way" to structure an organization, few, if any,

types of network configurations are helpful in all times and all places. For instance, to generate creative ideas for new products, organizations benefit from adaptable networks with weak connections that promote exploration (Fang et al., 2010; Hargadon and Sutton, 1997). Other objectives however, like actually transforming ideas into products, demand more rigid, hierarchical relations. Using this contingency lens, researchers have identified how many factors moderate the performance effects of particular network configurations, including the nature of the underlying task (Hansen, 1999), the behavioral orientation of the parties involved (Lingo and O'Mahony, 2010; Obstfeld, 2005), and their geographic location (Bell and Zaheer, 2007; Fleming et al., 2007).

As evidence continues to mount in support of the idea that intraorganizational network structures have significant but contingent effects on many important outcomes, researchers have given far less attention to an important complementary question: If some kinds of network configurations are better for reaching particular goals than others, how can organizations change established networks? What strategies are available for the leaders and members of an organization to reshape the global patterns of communication, collaboration, and other ties among their subordinates and peers? Much organizational research implies that even minor changes are likely difficult because networks have institutional and personal underpinnings and relationships are supported by routines and habits that are costly to abandon (March and Simon, 1958; Marquis, 2003; Stinchcombe, 1965). Implementing and sustaining network change therefore requires knowledge of how social and organizational factors constrain (or enable) those changes.

In recent years, scholars have made substantial progress in expanding our knowledge of network dynamics across an array of organizational settings, ranging from the evolution of alliance networks among high-technology companies (Powell et al., 2005; Sytch et al., 2012) to the persistence of collaborative relationships in academia

(Dahlander and McFarland, 2013). Despite this progress, several factors limit the ability of existing studies to account for intraorganizational network change. First, a substantial portion of research on network dynamics examines *inter*organizational relationships. Although many network mechanisms operate across levels of analysis (Brass et al., 2004; Phelps et al., 2012), relevant contextual factors like the incentives and opportunities that influence relationship formation likely differ substantially depending on whether the actors under investigation are organizations or people. Second, while recent examinations of network dynamics within organizations show how contextual factors effect people's propensities to form and maintain relationships, existing studies focus on ego networks, which are the portfolio of ties that belong to individual actors (Kleinbaum, 2012). By treating tie formation as an individual-level phenomenon, existing work overlooks how particular features of the organizational context influence the decision to establish a connection. Consequently, our knowledge of how leaders and members might reshape the global structure of an intraorganizational network remains limited.

To address the questions outlined above and extend research on network change, I conceptualize intraorganizational networks as attributes of organizations rather than of the people that constitute the nodes of those networks. Using this approach, I focus on how organizations' systems for categorizing and norms of evaluating knowledge shape bridging tie formation. Bridging ties influence many aspects of performance in knowledge-intensive organizations. A bridging tie is a connection that spans distant areas in a social network (Granovetter, 1973; Tushman, 1977). These kinds of boundary-spanning connections are often studied because they enhance integration and the flow of diverse knowledge (Davis and Eisenhardt, 2011; Grant, 1996; Henderson and Cockburn, 1994). My motivation for examining bridging ties differs. Setting aside their performance implications, bridging ties are a useful starting point for building theories of network of change because adding even a small number of con-

nections among otherwise distant actors can lead to qualitative transformations in global networks properties (Gulati et al., 2012; Watts and Strogatz, 1998). Bridging ties are therefore powerful tools for reshaping networks.

In general, knowledge categorization systems—which I define as an organization's map of its accumulated knowledge and expertise—should facilitate bridging by making it easier for the members of an organization to find relevant peers. Knowledge categorization systems take on many forms and often serve multiple functions. For example, universities rely heavily on disciplinary departments for organizing knowledge embedded in faculty, journals, books, classes, workshops, and research space (Abbott, 2001; Stinchcombe, 1990). Virtual organizations use content tagging to facilitate knowledge storage and retrieval (Golder and Huberman, 2006). And consulting firms employ electronic databases to help their employees draw on accumulated knowledge embedded in materials and expert colleagues for reuse in ongoing projects (Haas and Hansen, 2007; Hansen et al., 1999).

Despite their benefits for improving search, I argue that knowledge categorization systems can sometimes inhibit bridging and therefore network change. I hypothesize three cases that are likely to be especially pernicious: (1) when an organization's knowledge categorization system grows too large, (2) when the knowledge categorization system decouples from the actual distribution of expertise within the organization, and (3) when the standards used to evaluate knowledge contributions are variable across an organization's knowledge categorization system. I test and find support for all three hypotheses using data on 23 million knowledge-sharing exchanges among 1.3 million members of 25 online technical communities over a four-year period.

The remainder of this chapter is organized as follows. First, I draw on theories of networks and categorization to develop hypotheses that relate characteristics of organizations' knowledge categorization systems and norms of evaluating knowledge contributions to bridging tie formation. I then describe the research setting and

methods of analysis. Next, I present the results along with a battery of robustness tests that add confidence to the findings. I close with a discussion of the broader theoretical and substantive implications.

4.2 Network Change in Knowledge-Intensive Organizations

Knowledge-intensive organizations ranging from high-technology companies to hospitals face a common dilemma of finding ways to integrate groups of people with diverse backgrounds (Lawrence and Lorsch, 1967; Owen-Smith, 2001; Szulanski, 1996). For these organizations, accomplishing what may even be routine tasks—like imaging materials with electron microscopes or performing prostatectomies—requires having members who are deeply specialized in particular knowledge domains. To succeed at larger objectives, however, knowledge-intensive organizations must also foster communication and collaboration among those specialized members (Argote and Ingram, 2000; Grant, 1996; Henderson and Cockburn, 1994). For example, to develop a new product, an advanced materials firm's experts in electron microscopy must transfer knowledge to their colleagues in engineering. Likewise, for a hospital to treat a patient's prostate cancer, urologists need to interact effectively with radiation oncologists, pharmacists, nurses, and a host of other professionals.

Despite the need for integration, at least two factors limit communication and collaboration among experts in many organizations. First, specialization creates common bonds between people with similar backgrounds. Substantial research demonstrates that homophily is pervasive in social relations—people tend to associate with those who are similar to themselves (Kleinbaum et al., 2013; Kossinets and Watts, 2009; McPherson et al., 2001). Therefore, in the absence of intervening factors, communication and collaboration should be greatest among people with expertise in related domains. Second, when people do work with peers who specialize in areas other than their own, differences in technical vocabularies and problem-solving approaches

can limit the success of their interactions, at least without accompanying efforts at translation (Galison, 1997). Both factors should decrease integration and create segregated intraorganizational networks, where silos of experts interact extensively with one another but have few connections to peers who specialize in different domains.

One way for knowledge-intensive organizations to overcome these tendencies for segregation and eliminate silos of expertise is to promote tighter integration through bridging ties. By linking distant or unconnected areas of a network, bridging ties can be powerful tools for network change. For example, even a handful of distant connections dramatically reduces the number of intermediaries that members of an organization must travel through to find a person with relevant expertise—even for those people who are not an anchor of the bridge (Watts and Strogatz, 1998). By decreasing path lengths, bridging ties enhance knowledge flows within a network (Lazer and Friedman, 2007). Better knowledge flows help members of an organization become familiar with the technical vocabulary and problem-solving approaches of people from different backgrounds, and in so doing, lower barriers to integration. Moreover, to the extent that bridging ties persist over time, such connections also change the nature of homophilous behavior, as once distant actors become more alike through their collective influence over one another (Azoulay et al., 2009; Padgett and Powell, 2012).

Bridging relationships lead to substantial benefits for those who initiate them, including career advancement, compensation, and good ideas (Burt, 1992, 2004). Given these potential rewards, it seems reasonable to predict that people will naturally seek to build bridges among disconnected members of their organization. However, studies that adopt a more macro perspective suggest bridging is rare and that contextual factors have a substantial influence on the formation of such ties (Sorenson and Stuart, 2008; Sytch et al., 2012). In what follows, I build on these insights to develop hypotheses about how one important contextual factor—the systems organizations

use to map their accumulated knowledge—enables and constrains people's ability to form relationships with distant peers.

4.2.1 Knowledge Categorization Systems and Bridging Tie Formation

Most organizations use some system of categorization to keep track of what they know (Argote, 2013; Davenport and Prusak, 1998; Walsh and Ungson, 1991). Universities, for instance, rely on disciplinary departments to—among other things—cluster knowledge embodied in people (e.g., professors, students) and objects (e.g., books, electronic archives) in physical and organizational space (Stinchcombe, 1990). Professional associations establish sections to group members who have similar expertise and bundle research papers that address related topics. And online communities develop tagging systems to categorize the content added by their users. Within knowledge-intensive organizations, these systems of categorization serve as tools that people use for filtering, sorting, and screening their experience and finding relevant knowledge to accomplish their goals (March and Simon, 1958). In the search for knowledge, categories are like maps—they simplify reality and offer direction.

Much research focuses on the role of social relationships as tools for guiding knowledge search, often in place of the more formal categorization systems built into the structure of many complex organizations (Aral and Van Alstyne, 2011; Hansen, 1999; Reagans and McEvily, 2003). Despite their value for locating knowledge, social ties are less effective if a person requires insights on topics that are unfamiliar to his or her acquaintances. Formal categorization systems offer recourse when social search fails by helping a person identify knowledge that he or she may otherwise not encounter (Evans, 2008). Categorization systems should therefore promote bridging by allowing people to "hop" to unexplored parts of an intraorganizational network, where they can locate materials and interact with other members who have a deeper understanding of the knowledge they seek to obtain.

Organizational scholars demonstrate, however, that systems of categorization vary in their ability to convey helpful information to the people who use them (Hannan et al., 2007; Pontikes, 2012). Although many factors influence a categorization system's effectiveness, one of the most basic, yet important, is its size—i.e., the number of categories available. Systems with too few categories group together too many unrelated objects and therefore the categories convey little information about their constituent elements. Search costs increase because users need to do more manual sifting and sorting to find relevant knowledge. Communication challenges also arise when the categories people have at their disposal fail to match the complexity of their work. Consider, for instance, research on coding noise and performance in small groups. In one classic study, Christie et al. (1952) asked subjects to solve a problem that required sharing knowledge about the color of marbles using only written messages. When groups were given marbles with standard colors, they solved the problem easily, but in trials where the marbles had a cloudy and indistinctive finish, performance fell because members lacked a common language—or categorization system—for sharing knowledge (Macy Jr. et al., 1953; see also March and Simon, 1958, 161ff). Adding more categories to a system can alleviate these barriers and, in so doing, facilitate search and bridging tie formation.

Although adding categories should generally be helpful, beyond a certain point, systems may become too complex for their users to navigate and ultimately make search challenging. Prior studies across a range of settings, for instance, demonstrate that "as both the number of options and the information about options increases, people tend to consider fewer choices and to process a smaller fraction of the overall information available regarding their choices" (Iyengar and Lepper, 2000, 996) and therefore often refrain from making any choice at all (Iyengar and Kamenica, 2010; Tversky and Shafir, 1992). When an organization's categorization system offers too many options, I anticipate that because of cognitive limitations, people will use the

system less effectively to search for knowledge from unfamiliar sources. Consequently, bridging tie formation should decline. These considerations lead to a first hypothesis.

Hypothesis 6. *As the size of an organization's knowledge categorization system increases, intraorganizational bridging tie formation will also increase up to a point, beyond which increases in the size of the knowledge categorization system will decrease intraorganizational bridging tie formation.*

Put another way, a moderately differentiated knowledge categorization system should create more integrated and tightly coupled networks.

4.2.2 Knowledge Categorization Systems and Informal Knowledge Domains — The Problem of Decoupling

Knowledge categorization systems are part of an organization's formal structure. People use them as maps for—among other things—locating peers within their organization who have knowledge about subjects that lie outside their own areas of expertise. Hypothesis 6 builds on the analogy between knowledge categorization systems and maps to suggest that maintaining a balance between abstraction and detail—in terms of the overall number of categories—should be optimal for helping people connect with others by forming bridging ties.

Much like a map, however, a knowledge categorization system's effectiveness also hinges on the extent to which the categories align with meaningful clusters of knowledge as used by the members of an organization in practice. I refer to these latent groups of knowledge that emerge in practice as "informal knowledge domains." Knowledge categorization systems and informal knowledge domains often show signs of lose coupling in Weick's (1976, 3) sense—although they are responsive to one another, each "preserves its own identity and some evidence of its physical or logical separateness." Bowker and Star (1999, 232) capture this general idea in their notion

of "intimacy," which denotes the extent to which a categorization system "acknowledges common understandings that have evolved among members of the community." When it comes to categorization systems, intimacy is a good thing.

Orr's (1996) ethnographic study of technical representatives at Xerox offers a useful illustration of the distinction between knowledge categorization systems and informal knowledge domains. To help its representatives service and repair customers' photocopiers, Xerox developed an expansive technical documentation system. The system was largely organized around error codes. Given a code, the documentation explained step-by-step what a representative should do to resolve the problem. In the field, however, technicians often had to solve issues that fell outside the scope of the documentation and its highly linear, error-code-based approach to service and repair. To make their jobs easier, Xerox's technical representatives met regularly to discuss the problems they encountered in the field and exchange solutions. Through these meetings, the representatives developed their own folk knowledge of photocopier service and repair. In this example, the documentation system is a knowledge categorization system; the representatives' folk expertise is a bundle of informal knowledge domains.[1]

Many routine aspects of organizational life lead knowledge categorization systems to decouple from the underlying informal knowledge domains they seek to map. In the case of Xerox, the engineers who wrote the documentation had little firsthand experience repairing photocopiers—the people who designed the knowledge categorization system were not the ones who used it. Decoupling may also stem from the more natural evolution of knowledge within an organization, as members make new discoveries and abandon exhausted topics, but leaders fail to update or prune the categorization system. Acquisitions, spinoffs, and turnover, furthermore, all change the collective distribution of knowledge among members of an organization by either

[1]Eventually, Xerox created a database system, called Eureka, to more formally organize the technicians' repair knowledge and facilitate sharing (Brown and Duguid, 2000, 111ff).

bringing in outsiders or through the loss of active participants (Paruchuri et al., 2006; Rosenkopf and Almeida, 2003; Tzabbar, 2009).

Whatever the cause, it seems reasonable to predict that the consequences should be similar. Overlap between a knowledge categorization system and informal knowledge domains reduces some of the difficulties associated with bridging by making it easier for people to search and browse for knowledge within their organizations. Decoupling should have the opposite effect. When people have poor quality maps, they search more narrowly. If they search more narrowly, they will be less likely to venture into new settings. And if people do not venture into new settings, they are less likely to form distant ties (Feld, 1981; Sorenson and Stuart, 2008). These observations lead to a second hypothesis.

Hypothesis 7. *As decoupling between an organization's knowledge categorization system and informal knowledge domains increases, intraorganizational bridging tie formation will decrease.*

Related literature on the mismatch between formal and informal organizational structure offers some parallels to the arguments made here. Most notably, they are similar in their emphasis that, to understand many aspects of organizational behavior, it is necessary to view formal, codified rules and routines and the actual practices of members as distinct but interacting phenomenon (Brown and Duguid, 1991; Dalton, 1959). Despite these similarities, informal knowledge domains are conceptually distinct from informal social structure. The former are clusters of related knowledge, not social relationships.

4.2.3 Heterogeneous Evaluation Norms in Knowledge Categorization Systems

Hypotheses 6 and 7 propose that structural features of knowledge categorization systems influence bridging because such systems facilitate (or constrain) members'

abilities to retrieve accumulated knowledge from their organizations and locate new exchange partners. However, knowledge categorization systems are not only used for retrieval. In many organizations, they are also focal points around which new contributions to knowledge are evaluated. For example, when making hiring and promotion decisions, universities judge candidates with respect to the standards of the hiring department's discipline. Similarly, professional conferences evaluate submissions for inclusion in their programs and make awards to exceptional work according to the norms of sections and divisions.

The particular categories within a categorization system—and the members of an organization who identify with them—can vary along a number of dimensions in terms of how new contributions to knowledge are evaluated (Lamont, 2009; Rhoten and Parker, 2004). Some of these differences may stem from deeply held beliefs about what constitutes quality. Others may relate to whether convention states that low quality contributions should simply be ignored, or whether they should be policed and punished. Furthermore, categories may differ in terms of the extent to which quality can be determined by some objective standard. Regardless of origin, these and other factors should lead to the development of heterogeneous norms of evaluation along the lines of an organization's knowledge categorization system.

Heterogeneous evaluation norms are problematic for knowledge-intensive organizations that seek to foster bridging ties among their distant members. Variable evaluation creates an atmosphere of uncertainty. Research demonstrates that people tend to refrain from sharing knowledge when they feel uncertain about how their contribution will be received by their peers, especially if doing so may inadvertently reveal incompetence or otherwise cause lasting harm to their reputations (Bernstein, 2012; Edmondson, 1999; Irmer et al., 2002). When this kind of uncertainty exists, a person is likely to believe that sharing no knowledge is safer than sharing something that is potentially (or perceived to be) wrong. Therefore, to the extent that eval-

uation is highly variable across an organization and members cannot perceive how contributions will be evaluated, people should be most likely to share knowledge in familiar domains and refrain from bridging.

By contrast, in organizations where there is less uncertainty about how new, potentially low quality contributions will be received, people should feel better about venturing outside their comfort zone and are more likely to form bridging ties. For example, Hargadon and Sutton (1997) describe how IDEO, a design consulting firm, created a culture that rewarded knowledge sharing across disparate domains within the organization, and even went so far as to shun those engineers who kept potentially useful insights to themselves. These observations suggest a final hypothesis.

Hypothesis 8. *As evaluation norms become more heterogeneous across a knowledge categorization system, intraorganizational bridging tie formation will decrease.*

Note that the hypothesis does not predict that high standards or critical peers will necessarily prevent bridging but rather emphasizes the effects of uncertainty. Put differently, I predict that the less certain a person is about how he or she will be evaluated when straying from his or her comfort zone, the less likely he or she will be to do so.[2]

4.3 Research Setting

I tested the hypotheses using data on 23 million knowledge-sharing exchanges among 1.3 million members of 25 websites on the Stack Exchange Network over the

[2]Microsoft's stack ranking employee evaluation system offers an illustrative example of a policy that constrains tie formation across boundaries by increasing uncertainty about evaluation. The stack ranking system operates like a deterministic bell curve. As Eichenwald (2012) explains, "every unit [is] forced to declare a certain percentage of employees as top performers, then good performers, then average, then below average, then poor" regardless of whether or not all performed beyond expectations. Consequently, "a lot of Microsoft superstars did everything they could to avoid working alongside other top-notch developers, out of fear they would be hurt in the rankings" (Eichenwald, 2012). This, stack ranking evaluation systems may have the effect of increasing the attractiveness of working with colleagues nearby in a network because those nearby are familiar, while also decreasing the feasibility of working across divisional boundaries.

period of August, 2008 through July, 2012. As I discuss below, I conceptualize each site as an independent organization with its own intraorganizational network.

Stack Exchange is a family of online question-and-answer (Q&A) sites that span a rage of mostly technical topics, from electrical engineering to LaTeX typesetting systems. The origins of Stack Exchange date to August, 2008, when Jeff Atwood and Joel Spolsky, two prominent technology bloggers, launched Stack Overflow, the flagship site of the Network and currently one of the largest communities of programmers on the Internet (Atwood, 2008). By some estimates, new questions on Stack Overflow receive an answer in a median time of only 11 minutes (Mamykina et al., 2011). Server Fault, a sister site focused on system and network administration, was added in April of 2009, followed shortly thereafter by Super User, a site catering to general computer Q&A, in August. After an unsuccessful attempt to expand by licensing their Q&A platform for commercial use, Atwood and Spolsky founded Stack Exchange, Inc. in early 2011. Since then, dozens of sites have been added based on input from members of the community.

Several features of Stack Exchange sites set them apart from competitors and have contributed to the Network's rapid growth and continued popularity. All user generated content on Stack Exchange sites is available under a Creative Commons Attribution-ShareAlike license, meaning that posts are free to use and modify provided that attribution is made to the original author(s) and that derivative works are distributed under a similar license. Stack Exchange sites are also highly democratic. A portion of each sites's moderators are chosen through competitive elections. As a user contributes to a Q&A site, he or she gains the ability edit content, vote for questions to be closed or deleted, and a host of other privileges (Atwood, 2009).

Perhaps most importantly, Stack Exchange sites stand out for their elaborate reputation systems that incentivize high quality participation. Members are rewarded points for answering questions, but the magnitude of their reward is contingent on how

other participants evaluate the quality of the contribution. Evaluation occurs through a simple up/down voting mechanism. For each vote up on an answer, the contributor earns 10 points; a vote down, by contrast, lowers the contributor's reputation by 2. "Accepted" answers—i.e., those selected by the person who asked the question as the best response—earn their contributors 15 points. Members are also rewarded (or penalized) for asking good (or bad) questions, once again, depending on how they are evaluated by their peers.[3]

Finally, Stack Exchange sites are notable for their laser-like focus on high quality factual knowledge-transfer (as opposed to advice or opinions). Consider a recent interview with David Fullerton, Vice President of Engineering at the Network.

> Everything we do is about connecting experts with each other to ask and answer questions, so every interaction on the site is people networking. That social interaction is absolutely critical to how the site works. What separates us from a traditional social network is our focus on Q&A. We're not interested in discussion for discussion's sake or in chitchat: we want users to share information and answer real questions. We actually think communities work best when they work together to solve problems, not just come together to chat. (Begel et al., 2013, 60)

This mission distinguishes Stack Exchange sites from other online social network platforms because its primary goal is not connecting people, but rather answering questions. Unlike Facebook, Twitter, LinkedIn, and similar sites, Stack Exchange does not allow users to "friend" or "follow" one another, nor does it directly promote connectivity in other ways—ties are defined only in terms of actual knowledge sharing relationships.

Several additional considerations make Stack Exchange attractive for testing the hypotheses outlined above. Notably, all Stack Exchange sites run on a common underlying Q&A software platform. Although they differ in terms of aesthetics (e.g., color scheme, icons), they are functionally (e.g., navigation, search) identical. There-

[3]Although answering questions is by far the most effective way to build reputation, users can also earn a limited number of points in some other ways. For details, see `http://stackoverflow.com/help/whats-reputation`

fore, variation in behavior across sites should not stem from simple differences in design.

Despite these commonalities, sites on Stack Exchange can be viewed largely as independent organizations. Users who are active on more than one Stack Exchange sites have separate accounts, and reputation earned on one site does not (except in very limited circumstances) carry over to another. When browsing any particular Q&A site, there are few signs that others even exist. Only two noteworthy factors hold the different sites together. First, they have a common underlying administration—all are operated by Stack Exchange, Inc. Second, in some cases, questions can be moved from one site to another if the question is off topic for the site on which is was asked. In most cases, however, such questions will simply be closed.

Stack Exchange is also a valuable research setting because sites use a clear knowledge categorization system—all questions are coded using subject tags (Barua et al., 2012). Each site has its own independent set of tags for organizing content. New tags can be proposed by users who have sufficient reputation.[4] Over time, some unused tags are removed automatically by the Stack Exchange software. Duplicates are also occasionally removed through the expansion of a crosswalk of synonymous tags, accessible by moderators.

Finally, the available data are attractive because they allow for the observation of both successful and unsuccessful bridging ties. Although bridging often leads to positive outcomes, context also matters, and connecting diverse parties oftentimes has negative consequences (Smith, 2005; Xiao and Tsui, 2007). The top panel of Figure 4.1 below shows an example of a bridging tie (in the form of a post) that received a high rating; the middle panel of the figure, also a bridging connection, received a low evaluation.

[4]The reputation threshold is very steep. For example, on Stack Overflow, a reputation of 1,500 is required to add tags, meaning that only 1.6% of users had this privilege as of August, 2012.

Figure 4.1: Sample Posts[†]

Question: I have a script that downloads files from an FTP server using the curl command. When I run the script manually the download finishes correctly. I wrote a Java program to run the script automatically. However, when I run the program it freezes after a few minutes. By freezing, I mean that the Java program is running but the file does not continue to download anymore. What can cause this type of behavior? — *member0001*

Answer: I'm not a Java person, but rather a Unix one, and one thing seems obvious to me: The buffer on either stdout or stderr is filling up, and then curl is blocking. Does it work if you run curl in silent mode? Based on the Java documentation, it looks like you want to use getErrorStream and getInputStream. — *member0002*

Question: I wrote a Python script using android-scripting. It basically vibrates every minute (like a motivator). However, when the phone is locked with screen blanked out, I don't sense any vibration. Perhaps Android is freezing the script? Note that I am running the script as a service. Is there a way to make it work all the time regardless of the phone suspend state? — *member0007*

Answer: Here's a possible solution—use some scheduler software and start your script regularly. This way you'll not need to call time.sleep(). Maybe scripting is not a best solution for such periodic tasks. You will not face this problem if you write a simple Java app. — *member0008*

Question: It looks like the lists returned by keys() and values() methods of a dictionary are always a 1-to-1 mapping (assuming the dictionary is not altered between calling the 2 methods). If you do not alter the dictionary between calling keys() and calling values(), is it wrong to assume the above for-loop will always print True? I could not find any documentation confirming this. — *member0003*

Answer: Yes, what you observed is indeed a guaranteed property – keys(), values() and items() return lists in congruent order if the dict is not altered. iterkeys() &c also iterate in the same order as the corresponding lists. — *member0004*

Comments:
-1: no reference to the documentation (or source). — *member0005*

Dude, that's Alex Martelli, he's the author of Python in a Nutshell and The Python Cookbook. He doesn't need provide a reference. — *member0006*

@member0006: Even if it was Guido van Rossum, the Benevolent Dictator for Life, I'd still ask for a reference. — *member0005*

[†] These posts have been edited for length and adapted to help preserve the anonymity of their authors.

4.4 Data and Methods

4.4.1 Sample

I obtained all posts, edits, tags, member profiles, and other content made on Stack Exchange between August, 2008 and August, 2012 directly from the company. Stack Exchange makes these data publicly available in XML format on a quarterly basis.[5] To facilitate comparisons, several Q&A sites included in the data dump were excluded from the analyses. I used the following procedure to identify an appropriate sample. First, I began with all Q&A sites designated by Stack Exchange as having a "Science" or "Technology" focus. I chose these two categories because prior research suggests that the social structure of technical Q&A sites differs systematically from those that cater to other subject areas, where replies tend to offer opinions or advice rather than factual knowledge (Adamic et al., 2008). Of the 35 substantive sites included in the data dump, I excluded 4 that were categorized at "Life/Arts" and 6 that fell under the heading of "Culture/Recreation," leaving a total of 25. Each Q&A site on Stack Exchange also includes a companion "Meta" site, where users can report bugs, propose features, or ask about other aspects of their respective community. I exclude these sites from the core models, but find in robustness tests that the results are similar when they are included. Table 4.1 provides an overview of the sites used in the analyses.

4.4.2 Network Construction

I treat the 25 sample Stack Exchange sites as independent organizations, each with its own network of members. To measure the structure of each intraorganizational network, I use data on electronic communications made between users through posts. On Stack Exchange, posts come in three forms that can be viewed hierarchically:

[5] `http://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/`

Table 4.1: Sample Overview

Site	Category	Age (Years)†	URL	Tags†	Members†
Android Enthusiasts	Technology	1.9	http://android.stackexchange.com/	972	14,702
Ask Different	Technology	1.9	http://apple.stackexchange.com/	932	22,338
Ask Ubuntu	Technology	2.0	http://askubuntu.com/	2,444	65,447
Theoretical Computer Science	Science	2.0	http://cstheory.stackexchange.com/	408	8,371
Database Administrators	Technology	1.6	http://dba.stackexchange.com/	639	9,869
Drupal Answers	Technology	1.4	http://drupal.stackexchange.com/	700	7,769
Electrical Engineering	Technology	1.8	http://electronics.stackexchange.com/	1,092	9,068
Game Development	Technology	2.0	http://gamedev.stackexchange.com/	850	15,963
Geographic Information Systems	Technology	2.0	http://gis.stackexchange.com/	1,123	7,719
Mathematics	Science	2.0	http://math.stackexchange.com/	784	31,109
Mathematica	Technology	0.5	http://mathematica.stackexchange.com/	400	1,846
Physics	Science	1.7	http://physics.stackexchange.com/	745	9,306
Programmers	Technology	1.9	http://programmers.stackexchange.com/	1,714	53,205
IT Security	Technology	1.7	http://security.stackexchange.com/	556	10,973
Server Fault	Technology	3.3	http://serverfault.com/	4,933	89,679
SharePoint	Technology	1.7	http://sharepoint.stackexchange.com/	879	7,001
Stack Overflow	Technology	4.0	http://stackoverflow.com/	31,837	1,295,620
Cross Validated	Science	2.0	http://stats.stackexchange.com/	840	10,755
Super User	Technology	3.0	http://superuser.com/	5,066	112,577
TeX - LaTeX	Technology	2.0	http://tex.stackexchange.com/	935	14,855
Unix & Linux	Technology	2.0	http://unix.stackexchange.com/	1,527	19,075
User Experience	Technology	2.0	http://ux.stackexchange.com/	668	14,638
Web Applications	Technology	2.1	http://webapps.stackexchange.com/	832	17,747
Webmasters	Technology	2.1	http://webmasters.stackexchange.com/	778	14,562
WordPress Answers	Technology	2.0	http://wordpress.stackexchange.com/	792	15,121

† As of July 31, 2012.

(1) questions, (2) answers, and (3) comments. Questions are the focal point around which exchanges occur, and always appear as the top post within a thread. Once a question is posted, members can respond with answers that appear below the question. Answers are arranged according to their quality, as measured by votes. Finally, comments are short communications that appear below specific questions or answers. Their purpose is to allow members to ask for or append clarifications, suggest corrections, or convey other information that does not directly answer the question that anchors the thread.

The design of Stack Exchange and the hierarchical nature of questions, answers, and comments makes it relatively straightforward to identify directed knowledge sharing relationships among members of each sample site. Answers are the most clear-cut: When a member i posts an answer to a question asked by j, I record a tie from i to j. Questions that do not receive any replies do not result in the creation of any ties.

Comments are slightly more challenging to map onto ties because the intended recipient may not be the author of the question or answer to which the comment is appended—oftentimes, comments are made by members in response to other comments. Fortunately, the Stack Exchange software allows members to specify the intended target of their remarks by including the target's name in their comment, preceded by "@". (For an example, see the bottom panel of Figure 4.1 above.) Identifying a target in this way notifies the member of the comment through the Stack Exchange system and also potentially via e-mail. Thus, for comments, I coded a tie from member i to j if i explicitly identified j in his or her post; in cases where the person making the comment does not specify a target, I code the author of the parent question or answer as the endpoint of a directed relationship originating from i.

I then aggregated the individual-level knowledge sharing relationships to approximate the global intraorganizational network structure for each sample site. Following prior research on electronic communication networks, I use a sliding window filter

approach to make these approximations (Kossinets and Watts, 2006, 2009; Moody et al., 2005). This approach uses two parameters, τ and δ, to define a global network from individual-level interactions. The first parameter, τ, specifies the size of the time window during which any pair of members in the network must have had some form of exchange in order for a tie to exist between them at time t; the second parameter, δ, defines a discrete interval according to which the the window moves, and consequently, how old ties are dropped and new ones are added. Therefore, "the instantaneous network at time t includes all dyads with nonzero strength or, equivalently, all dyads that have exchanged messages within the interval $(t-\tau, t]$" (Kossinets and Watts, 2009, 414).

In the analyses presented below, I set $\tau = 7$ days and $\delta = 1$ day. Although prior studies of electronic communication networks vary substantially in terms of window sizes, many investigations approximate global network structures using longer units of 30 days or more (Aral and Van Alstyne, 2011; Kossinets and Watts, 2009; Zhang et al., 2007). Several considerations, however, suggest that for Stack Exchange, a shorter window is appropriate. First, because many sites have substantial activity, older ties are likely less representative of contemporary communication patterns for any given Q&A community and therefore retaining them may lead to inaccurate measures. Second, although some questions remain active for long periods of time, most replies come within a relatively short window after a question is posted. Figure 4.2 shows the distribution of active questions (left panel) and replies (right panel) as a function of time for all 25 sample sites. To minimize right censoring, the figure only includes questions posted before January 1, 2012—a full eight months before the date of the data dump.

Only 25% of questions receive any answers or comments 10 days after their initial posting, and 75% of replies are made within 12 hours. Based on these considerations, I chose $\tau = 7$ days and $\delta = 1$ because they capture the vast majority relevant

activity, while also corresponding to substantively meaningful units of weeks and days, respectively. However, as I discuss in greater detail below, the findings are robust to a variety of alternative values of τ and δ.

To help clarify the data structure, consider Figures 4.3 and 4.4, which offer schematic illustrations of the intraorganizational networks of two Stack Exchange sites, Super User and Server Fault, one year after their foundings. As discussed above, each site can be viewed as an independent organization, with its own unique intraorganizational network structure, knowledge categorization system (i.e., tags), and informal knowledge domains, none of which are influenced directly by activity on other sites. Super User and Server Fault are comparable along a number of dimensions, including their age, number of members, and total tags.

For their respective sites, Figures 4.3 and 4.4 display connections among the largest communities of members. Although methodological details are reserved for Section 4.4.3, these communities correspond roughly to subgroups of members who have many connections among one another and relatively few ties outside the subgroup. Arrows correspond to to bridging ties that span communities *within* a site. The thickness of an arrow is proportional to the number of underlying exchanges between members of connected communities; the direction of an arrow corresponds to the direction of exchange. Notice that knowledge sharing among communities is not necessarily reciprocal and both figures contain asymmetric ties.

Despite some organizational similarities, Super User and Server Fault have notably different network structures. Overall, as reflected in Figures 4.3 and 4.4, with a mean of 2.6 ties per community, bridging is more common on Super User than it is on Server Fault, where the corresponding statistic is only 1.9. Moreover, Super User's top network communities are highly integrated, especially in comparison to Server Fault, where the network is centralized around the Linux community and there are few connections among smaller subgroups.

4.4.3 Dependent Variable

The dependent variable captures new bridging ties created among members of each sample site at time $t + 1$. I define bridging ties as knowledge sharing relationships between members of different network communities within a site (c.f., Sytch et al., 2012).[6] Once again, note that these "network communities" are different from the Q&A sites discussed above. The former exist only *within* a site and are defined by patterns of knowledge sharing among members, while the latter is synonymous with "a site on Stack Exchange." None of the network measures or concepts used in this chapter entail connections *across* sites.

To identify communities of members, I use a walktrap method developed by Pons and Latapy (2005). Unlike some community detection algorithms, walktrap is computationally feasible for large networks and therefore is useful for the purposes of this study given the scale of the data. Although there are many algorithms available for uncovering community structure in networks, all share the common goal of partitioning nodes into groups such that connectivity is higher among nodes within groups but lower between them (Fortunato, 2010). The basic intuition behind the walktrap method is that when traversing a network, a random walker should get stuck visiting pockets of highly interconnected nodes. Using this idea, distances between any pair of nodes i and j can be measured as the probability that a random walker moves from i to j within some predetermined number of steps, w, which I set to 4.

After obtaining these distances, I grouped nodes into network communities using Ward's hierarchical clustering algorithm (Pons and Latapy, 2005). For each network, I selected a partition that maximized the modularity function of Newman and Girvan (2004). Modularity is defined conceptually as the difference between the proportion of edges that fall within the communities identified by a given community detection

[6] I also considered several alternative measures of bridging tie formation, which I discuss in greater detail in Section 4.7.

algorithm and the proportion expected for a comparable random network. Across sample sites, the average modularity was 0.5, which is characteristic of networks with a strong community structure. To facilitate modeling and interpretation, I use the log of bridging ties in all analyses.

4.4.4 Independent Variables

Knowledge categorization system size. Hypothesis 6 proposes that there is a curvilinear (inverted U-shaped) relationship between the size of an organization's knowledge categorization system and bridging tie formation. I measure the size of each sample site's knowledge categorization system as the number of tags available for use at time t. As discussed above, tags are an essential component of all Stack Exchange sites (Barua et al., 2012). When submitting a new question, members are required to select at least one tag to characterize their post, and many users opt for more. Tags are also essential for site navigation. All Stack Exchange sites feature a prominent tab at the top of each page that allows users to filter recent posts by tag, along with a sidebar on every index page that serves a similar function. In short, tags are the backbone through which members locate and contribute to organizational knowledge.

Decoupling. Bridging is also likely to be influenced by the extent to which an organization's knowledge categorization system maps onto informal knowledge domains, as proposed by Hypothesis 7. Although tags offer a clear proxy for the structure of each site's knowledge categorization system, informal knowledge domains are more elusive. Yet measures of both are needed to estimate decoupling.

Recall that informal knowledge domains are defined as groups of related knowledge that emerge through use and practice (in contrast to the maps of those groups—i.e., categorization systems). To capture this idea empirically, I partition questions on each Stack Exchange site into latent clusters based on their degree of textual similarity.

Using this approach, it is possible to place questions that contain similar words into discrete bins that are defined independently of the tags used by members to classify those questions.

I use affinity propagation to cluster questions on each site posted between t and $t - 30$ days. Affinity propagation is an algorithm that finds exemplars among a set of data points and uses those exemplars as anchors for assembling larger groups of observations (Frey and Dueck, 2007).[7] Although I reserve a more detailed discussion of the procedure for Appendix B, several attractive features of the algorithm are worth noting. First, affinity propagation is an unsupervised learning algorithm and therefore requires no training data. Second, the algorithm does not need the number of clusters to be specified as an input parameter. This feature is especially important because I have no basis to estimate the number of latent knowledge domains for each Q&A community. Even if it were possible to make an educated guess (e.g., using knowledge about a particular community), the number of sites and observation periods covered by the sample would make doing so intractable. Finally, relative to comparable methods, affinity propagation is known to produce higher quality clustering solutions with only a fraction of the computational resources.

After clustering the questions on each site i into informal knowledge domains, I then estimate decoupling as

$$D_{it}(C, K) = \sum_{k \in K} \sum_{c \in C} p(c, k) \log \left(\frac{p(c, k)}{p(c)\, p(k)} \right), \tag{4.1}$$

where $p(c, k)$ is the joint probability distribution function of tags (C) and informal knowledge domains (K), and $p(c)$ and $p(k)$ are the marginal probability distribution functions of tags (C) and knowledge domains (K), respectively. $D_{it}(C, K)$ is

[7] Because of computational limitations, I was only able to apply the algorithm to a 7 day moving window on Stack Overflow. The substantive findings are similar if I use a 7 day window for all sites. However, because I was able to obtain higher quality partitions with more data, I chose to use the longer window when computationally feasible.

equivalent to the negative mutual information of tags (C) and knowledge domains (K) from t to $t-30$ at time t (Shannon and Weaver, 1949). Mutual information is a measure that captures how much knowledge of one random variable reveals (i.e., reduces uncertainty) about another. The measure has some conceptual parallels to the correlation coefficient, but it captures nonlinear relationships and can be used with discrete variables. Because mutual information increases monotonically with the number of discrete values (Vinh et al., 2009), I adjust $D_{it}(C, K)$ for chance agreement using the procedure discussed in Appendix C. After adjustment, I subtract the results from 1 and multiply by 100 so that larger values indicate greater decoupling and the regression coefficients correspond to unit changes.

As an illustration, consider a university where knowledge is categorized around traditional disciplinary departments, but researchers produce work that is actually highly interdisciplinary in nature. In this example, decoupling should be high (i.e., close to 100) because knowing about the distribution of departments will reveal little about the underlying informal knowledge domains. One might expect, for instance, that knowing a researcher's department would not make it easier to guess the kinds of work he or she does; likewise, knowing about a researcher's work would not be useful for discerning his or her departmental home.

Evaluation heterogeneity. Finally, Hypothesis 8 proposes that in organizations where norms of evaluation are highly variable, people will be less likely to venture outside of their comfort zones and therefore bridging should be lower. To test this hypothesis, I created a measure of evaluation heterogeneity as follows. First, for each site i at time t, I collected all votes on questions and answers associated with each tag, up to time t. I then assembled a set of values U by obtaining, for each tag, the difference between the fraction of votes up and fraction of votes down. The values of U range in theory from -1 (when all votes associated with a tag are down) to 1 (when all votes associated with a tag are up). I use total votes as of time t (rather

than votes cast within a more narrow window) because they more accurately reflect the information members will use to form impressions about the risk of asking or answering a question associated with a particular tag. To calculate the measure of evaluation heterogeneity, I use the quartile coefficient of dispersion, defined as

$$\frac{Q_3 - Q_1}{Q_3 + Q_1}, \tag{4.2}$$

where Q_1 and Q_3 are the values of the first and third quartiles of U, respectively. As with decoupling, I multiply the values by 100 so that the regression coefficients are unit changes. This measure is useful because it is designed specifically to facilitate comparisons across sets of data that vary in scale. Higher values indicate greater heterogeneity in evaluation along the dimensions of an organization's knowledge categorization system.

4.4.5 Control Variables

Interest homophily. Member's propensities for interacting with certain types of people—either through choice or opportunity—play an important role in tie formation. Prior research demonstrates that homophily—the tendency for people to associate with similar others—is one of the most basic factors driving the formation of relationships from marriages to business partnerships (Kleinbaum et al., 2013; McPherson et al., 2001; Ruef et al., 2003). In organizations where there is a greater propensity among members to interact with similar others, there should also be fewer bridging ties. For each site at time t, I measured homophily as the median cosine similarity of interests between members who communicated at any point between $t-7$ and t.

To approximate members' interests, I created a metric based on the tags within which they had prior activity. The measure ranges from 0 to 1, with higher values

indicating greater similarity. For a more detailed discussion of the procedure, see Appendix A.

Connectivity homophily. Prior research suggests that in many online Q&A communities, a relatively small number of highly active members answer a disproportionate share of questions posted by less seasoned users (Adamic et al., 2008). These individuals serve as integrators that span otherwise disconnected groups of members. By contrast, on sites where members interact with others who have similar levels of connectivity, integration—and therefore bridging—may be lower because highly active members are less likely to serve as hubs. To account for this possibility, I control for connectivity homophily using a measure of assortative mixing. Conceptually, this measure captures the extent to which the degree (i.e., number of ties) of connected nodes in a network are positively or negatively correlated (Newman, 2002).

Evaluation harshness. One alternative (and simpler) explanation for Hypothesis 8—i.e., that heterogeneous evaluation norms decrease bridging—is that members of an organization are less likely to venture outside their comfort zones as standards of evaluation increase, and that variability along the lines of a formal knowledge categorization system is irrelevant. To account for this possibility, I include a measure of evaluation harshness, which I define as the ratio of votes down to votes up across a site as of time t.

Communities (weighted). Prior research demonstrates that opportunity is a major driver of bridging tie formation (Sytch et al., 2012). To account for changes in bridging opportunities, I include a control for the number of network communities (see section 4.4.2) weighted by the number of network members (i.e., nodes) at time t.[8]

Clustering. I also control for the level of clustering in the network of each site

[8]I also tried controlling for the number of network communities and network members separately. The results are similar, however, I chose not to include those two variables in the final analyses because they introduced substantial multicollinearity in several models.

at time t. Clustering is defined as the ratio of open to closed triads in a network. The measure ranges from 0 to 1, with higher values indicating more closed triads. Conceptually, clustering captures the extent to which a person's acquaintances also know one another—a hallmark of cohesive networks (Coleman, 1988). Prior research suggests that cohesive networks tend to be more insular. Therefore, as clustering increases, members of a site should be less likely to connect with distant peers.

Repeat ties. Personal factors make intraorganizational networks resistant to change. As people build relationships with others members of their organization, they may tend to look for opportunities to interact with their close peers rather than searching more broadly. Therefore, I control for the number of repeated ties in each site's network at time t.

Centralization. As an additional check on the possibility that on some sites, bridging is driven by a small number of highly skilled members who answer questions from less able participants, I also control for the group out-degree centrality of each network at time t. The measure can be thought of as a group-level variant of out-degree centrality that allows networks of different sizes to be directly compared. For a node i, out-degree centrality is a count of the number of ties *to* (not from) other nodes. A person's out-degree centrality on a Stack Exchange site increases as he or she answers questions and comments on others' posts. The group out-degree index transforms this centrality measure into an aggregate construct by capturing variability in the scores among nodes (Wasserman and Faust, 1994, 175ff). It ranges from 0 to 1; more centralized networks have larger values.

Age. I control for the age of each site, defined as the number of days elapsed since the time of founding, divided by 30 so that the units are months. Age has a profound effect on many outcomes for organizations (Hannan, 1998; Sorensen and Toby, 2000). With respect to bridging, older sites may have more established routines and institutions that define relations among communities and, in so doing, decrease

the likelihood of connections among distant members.

Informal knowledge domains. To account for changes in the distribution of knowledge on each site, I add a control for the number of informal knowledge domains as of time t using the natural language methodology outlined in Section 4.4.3. For more details on the procedure used to obtain this measure, see Appendix B.

New members. The distribution of knowledge and expertise within an organization changes as new members join, which may make it harder for existing participants to search using established routines and therefore alter the propensity for bridging. To capture these influences and other effects of proximate growth, I control for the number of new members to join each site at time t.

Prominent member loss. One major reason why networks resist change is that they have institutional foundations. Over time, sites develop informal, largely taken-for-granted rules and conventions that guide the question-answering and asking behavior of at least some of their members. As a crude proxy for institutional transformations that might impinge on bridging, I control for the (logged) number of prominent members—defined as those who are among the top 5% in terms of Q&A activity—of each site at time t who were inactive over the past 30 days. The results are similar when I use different thresholds to define prominent members.

Time effects. Finally, to account for cyclical patterns in Internet usage (Nie et al., 2002) and changes in Stack Exchange over time, I include dummy variables for the day of the week and the year of the observation, respectively, in all models.

Table 4.2 provides a summary overview of the variables used in the analyses.

4.5 Model Estimation

The dependent variable captures the log of new bridging ties created on each site at time $t + 1$. Because the measure is continuous and approximately normally distributed, I use ordinary least squares regression to estimate one-way fixed effects

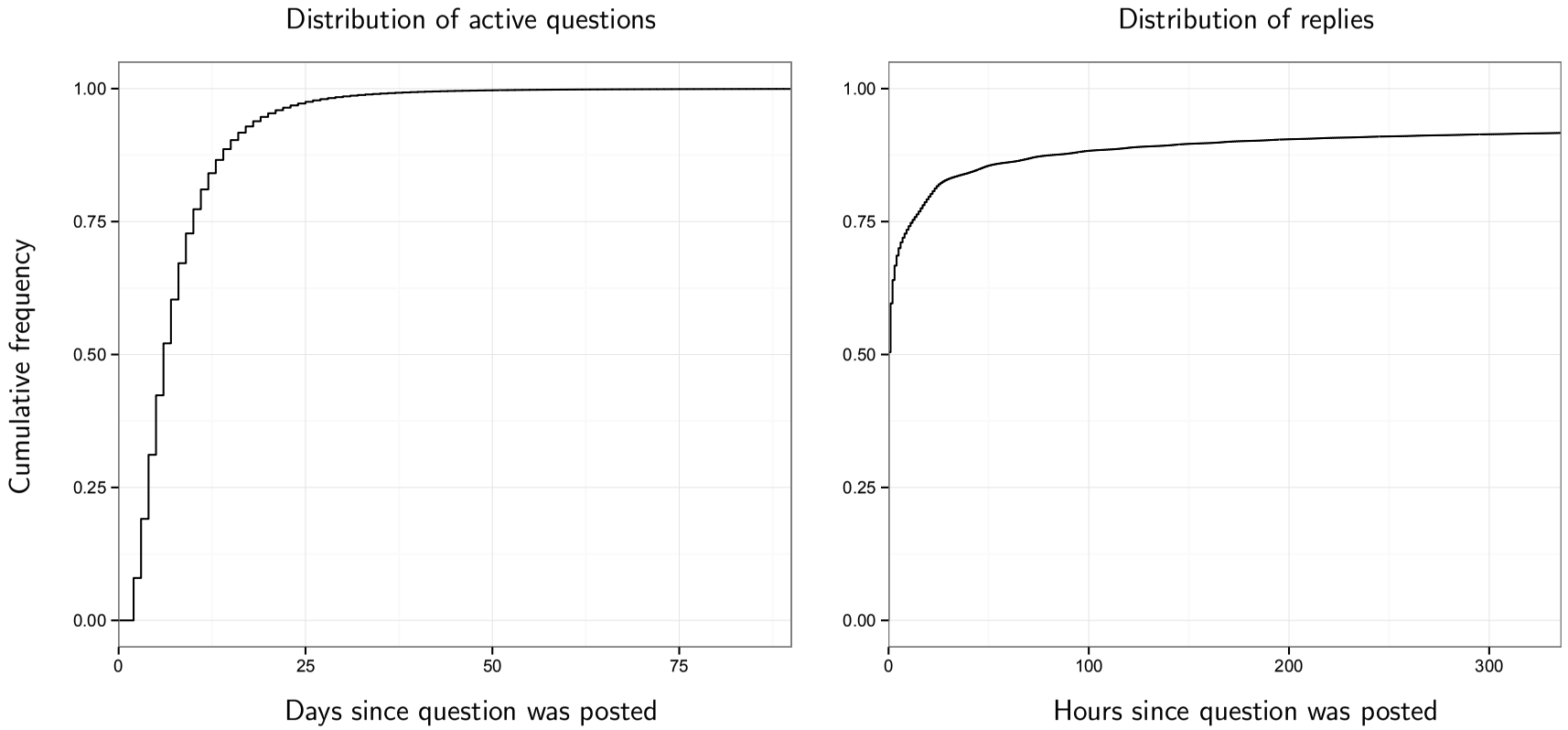

Figure 4.2: Cumulative frequency distributions of question activity.

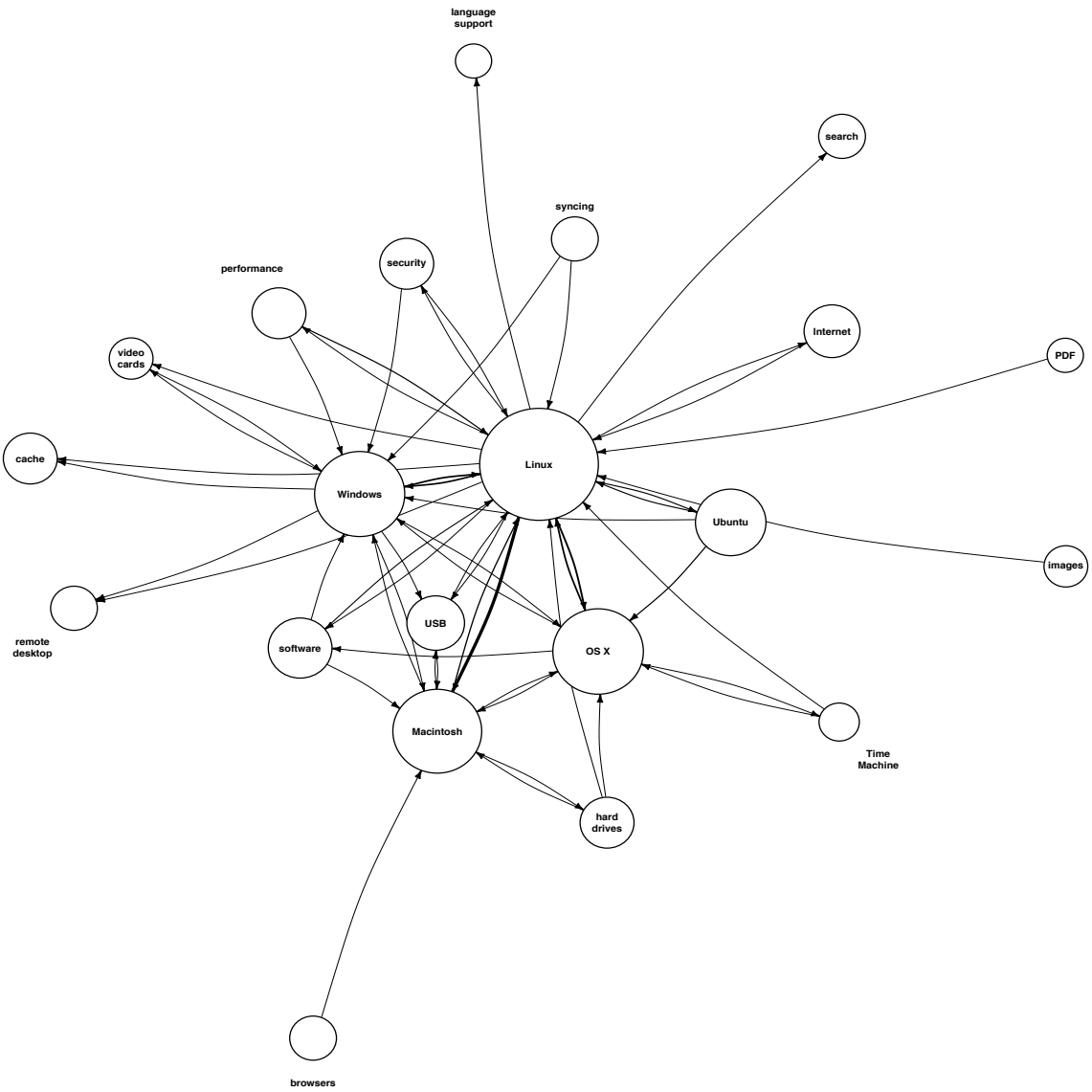

Figure 4.3: Network of top communities on Super User, July 12, 2010. Underpinning this figure are 2,197 exchanges among 1,294 members of the site. To create the diagram, I first ran a walktrap community detection algorithm on the member-level network to find groups of people who had many dense ties among one another, but relatively few ties outside the group. I then identified the most common tags used by members of each community to select a community label. Node sizes are proportional to the (logged) number of members in the community; edge width is determined by the number of ties between members of the connected communities.

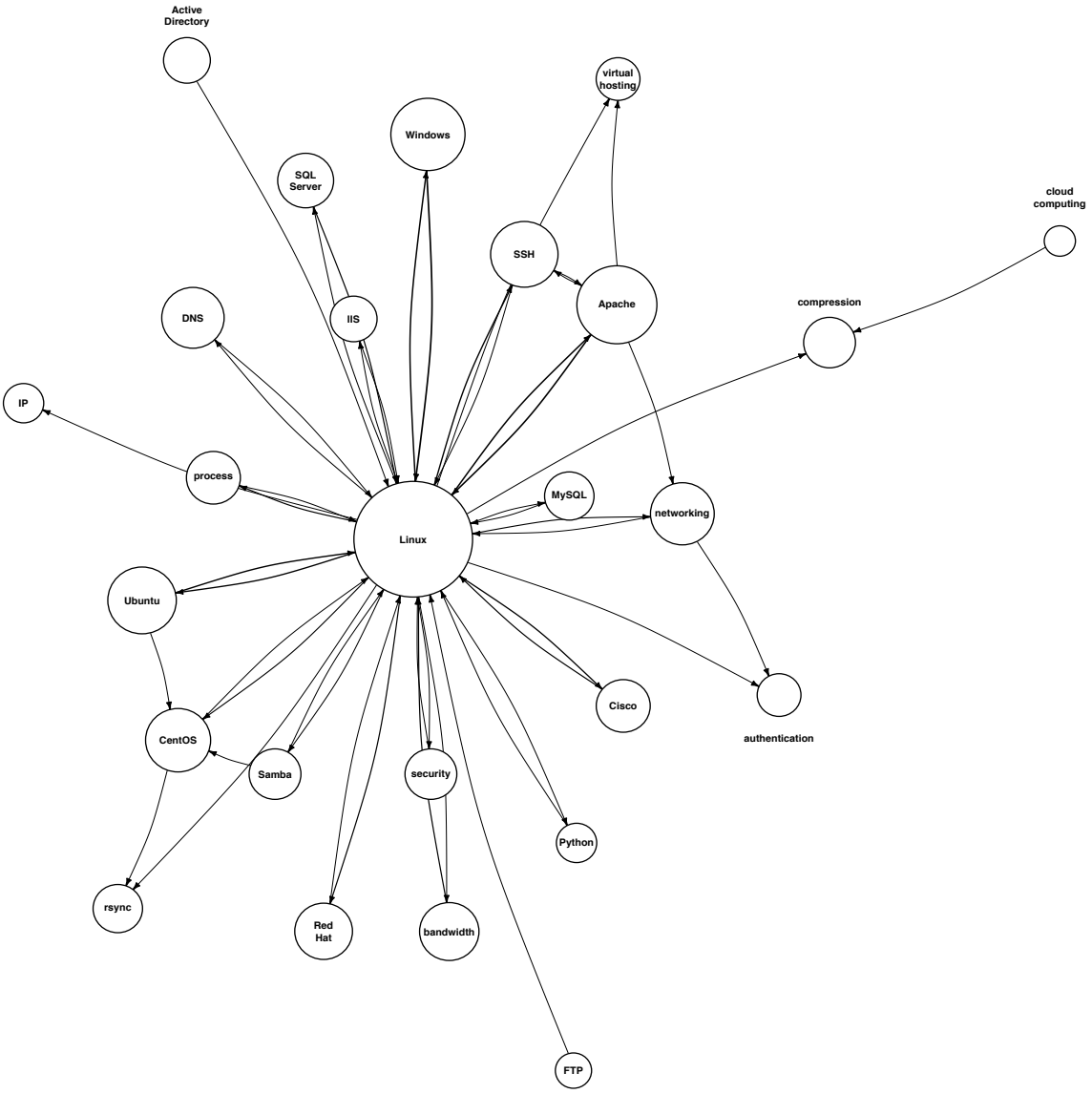

Figure 4.4: Network of top communities on Server Fault, April 30, 2010. Underpinning this figure are 1,983 exchanges among 1,188 members of the site. To create the diagram, I first ran a walktrap community detection algorithm on the member-level network to find groups of people who had many dense ties among one another, but relatively few ties outside the group. I then identified the most common tags used by members of each community to select a community label. Node sizes are proportional to the (logged) number of members in the community; edge width is determined by the number of ties between members of the connected communities.

Table 4.2: Variable Names and Definitions

Name	Definition	Panel Structure
Dependent Variables		
Bridging ties $(\log)_{(t+1)}$	New ties that span distinct network communities at time $t+1$	Updated daily as members form ties
Bridging ties $(\log)_{[t+1,t+3]}$	New ties that span distinct network communities from time $t+1$ to $t+3$	Updated daily as members form ties
Bridging ties 1 SD $(\log)_{(t+1)}$	New ties between members who differ by 1 SD or more on the similarity of their tags	Updated daily as members form ties
Bridging ties 0.5 SD $(\log)_{(t+1)}$	New ties between members who differ by 0.5 SD or more on the similarity of their tags	Updated daily as members form ties
Independent Variables		
Knowledge categorization system size $(\log)_t$	Count of the number of tags available for use	Updated daily as tags are added and removed
Decoupling$_t$	Mutual information of knowledge categorization system size and informal knowledge domains—reverse coded	Updated daily as tags are added and removed and groups evolve
Evaluation heterogeneity$_t$	Quartile dispersion of the fraction of votes up minus the fraction of votes down by tag	Updated daily as votes are cast
Controls—Members		
Interest homophily$_{[t-7,t]}$	Median similarity of tags among connected network members	Updated daily as members have activity with different tags
Connectivity homophily$_{[t-7,t]}$	Assortative mixing among network members based on degree	Updated daily as members form ties
Evaluation harshness$_{[t-7,t]}$	Fraction of votes down relative to total votes	Updated daily as votes are cast
Controls—Network Structure		
Communities (weighted)$_{[t-7,t]}$	Ratio of network communities to network members	Updated daily as members form ties
Clustering$_{[t-7,t]}$	Ratio of closed triangles to connected triples	Updated daily as members form ties
Repeat ties$_{[t-7,t]}$	Ratio of repeated ties to edges in the network	Updated daily as members form ties
Centralization$_{[t-7,t]}$	Uses out-degree centrality to capture extent to which questions are answered by a small number of members	Updated daily as members form ties
Controls—Site		
Age (months)$_t$	Number of days since the founding of a focal site, divided by 30	Updated daily as a site ages
New members $(\log)_t$	Number of new accounts created on a focal site	Updated daily as members join

Table 4.2 (Continued)

Prominent member loss $(\log)_{[t-30,t]}$	Number of members in top 5% by activity who have not posted in past 30 days	Updated daily as members ask, answer, and comment on questions
Informal knowledge domains$_t$	Number of clusters of questions based on textual similarity	Updated daily as members ask questions

panel models of the form

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \epsilon_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T_i, \tag{4.3}$$

where y_{it} is the dependent variable for site i and time t, $\mathbf{x}_{it}$ are the independent and control variables, β is a vector of coefficients to be estimated, and α_i are time-invariant, unit (site) specific effects. Fixed effects approaches are attractive because they model changes within units over time and therefore control for all time-invariant unobserved heterogeneity (Cameron and Trivedi, 2005).[9]

The panel structure of my data are such that I have many repeated observations (large T) for relatively few cases (small N). This structure differs from typical longitudinal studies of organizations—which tend to be smaller in the T dimension but larger with respect to N—and they require more careful treatment of temporal dependence. Researchers have developed a variety of estimators for small N large T panel analyses that are widely used in political economy and finance, where repeated observations of a small number of countries or securities are commonplace (Beck and Katz, 2011). I report standard errors using the covariance matrix estimator of Driscoll and Kraay (1998), which is robust to cross sectional and temporal dependence for small N, large T panels (Hoechle, 2007).

4.6 Results

Descriptive statistics and zero-order correlations are shown in Table 4.3. Although none of the sample sites ever exited the panel, only Stack Overflow was active for the entire observation period. After accounting for new entries over the course of the study, the final analysis panel contains $18,534$ site-day observations. The variance inflation factors (VIFs) were within acceptable ranges across all models.

[9] A Hausman (1978) test also clearly rejected the appropriateness of a random effects specification ($p < 0.001$).

Table 4.3: Descriptive Statistics and Correlations[†]

	Variable	Mean	SD			1	2	3	4	5	6	7
			Overall	Between	Within							
1.	Bridging ties (log)$_{(t+1)}$	3.15	1.71	1.33	0.66	1.00						
2.	Bridging ties (log)$_{[t+1,t+3]}$	4.03	1.75	1.39	0.61	0.97	1.00					
3.	Bridging ties 1 SD (log)$_{(t+1)}$	2.78	1.70	1.22	0.75	0.89	0.88	1.00				
4.	Bridging ties 0.5 SD (log)$_{(t+1)}$	4.19	1.65	1.26	0.64	0.95	0.94	0.92	1.00			
5.	Interest homophily$_{[t-7,t]}$	0.28	0.07	0.06	0.03	−0.31	−0.32	−0.16	−0.37	1.00		
6.	Connectivity homophily$_{[t-7,t]}$	−0.09	0.08	0.05	0.06	0.26	0.25	0.28	0.30	−0.09	1.00	
7.	Evaluation harshness$_{[t-7,t]}$	0.03	0.02	0.01	0.01	0.13	0.13	0.14	0.16	−0.18	0.16	1.00
8.	Communities (weighted)$_{[t-7,t]}$	0.23	0.05	0.02	0.05	−0.15	−0.15	−0.21	−0.15	−0.05	0.03	0.01
9.	Clustering$_{[t-7,t]}$	0.07	0.04	0.04	0.03	−0.33	−0.33	−0.36	−0.44	0.53	−0.24	−0.18
10.	Repeat ties$_{[t-7,t]}$	0.16	0.06	0.05	0.03	0.02	0.03	0.06	−0.06	0.44	−0.39	−0.20
11.	Centralization$_{[t-7,t]}$	0.12	0.08	0.06	0.05	−0.42	−0.42	−0.43	−0.51	0.33	−0.66	−0.18
12.	Age (months)$_t$	13.56	9.20	3.85	8.18	0.41	0.40	0.50	0.49	−0.16	0.20	0.15
13.	New members (log)$_t$	3.36	1.28	0.93	0.67	0.79	0.78	0.77	0.84	−0.32	0.33	0.21
14.	Prominent member loss (log)$_{[t-30,t]}$	4.97	2.02	1.21	1.64	0.44	0.43	0.51	0.53	−0.29	0.29	0.20
15.	Informal knowledge domains$_t$	87.75	157.86	103.77	81.43	0.78	0.79	0.80	0.80	−0.12	0.32	0.11
16.	Knowledge categorization system size (log)$_t$	6.79	1.20	0.97	0.33	0.79	0.80	0.76	0.86	−0.50	0.41	0.20
17.	Decoupling$_t$	95.04	2.11	1.80	1.19	−0.23	−0.24	−0.22	−0.27	0.24	0.09	0.07
18.	Evaluation heterogeneity$_t$	2.83	1.72	1.59	0.84	0.30	0.30	0.35	0.35	−0.24	0.21	0.64

	Variable	8	9	10	11	12	13	14	15	16	17	18
8.	Communities (weighted)$_{[t-7,t]}$	1.00										
9.	Clustering$_{[t-7,t]}$	0.14	1.00									
10.	Repeat ties$_{[t-7,t]}$	−0.02	0.50	1.00								
11.	Centralization$_{[t-7,t]}$	0.09	0.66	0.55	1.00							
12.	Age (months)$_t$	−0.11	−0.54	0.05	−0.44	1.00						
13.	New members (log)$_t$	−0.12	−0.55	−0.19	−0.59	0.64	1.00					
14.	Prominent member loss (log)$_{[t-30,t]}$	−0.14	−0.68	−0.14	−0.56	0.88	0.69	1.00				
15.	Informal knowledge domains$_t$	−0.19	−0.39	−0.05	−0.47	0.60	0.80	0.57	1.00			
16.	Knowledge categorization system size (log)$_t$	−0.13	−0.63	−0.29	−0.64	0.62	0.84	0.71	0.81	1.00		
17.	Decoupling$_t$	−0.03	0.39	−0.09	0.15	−0.24	−0.25	−0.24	−0.19	−0.20	1.00	
18.	Evaluation heterogeneity$_t$	−0.06	−0.32	−0.17	−0.31	0.35	0.39	0.42	0.26	0.39	−0.01	1.00

[†] $N = 18,534$

Table 4.4 presents OLS models of new bridging ties at time $t + 1$. Because the dependent variable is logged, the coefficients can be interpreted as elasticities or semi-elasticities. In general, the controls are consistent in terms of sign and significance across models. As expected, increases in both forms of homophily—interest and connectivity—are associated with decreases in bridging tie formation. When members of an organization tend to interact with others who have similar interests or connectivity to their own, there also tends to be fewer new connections between distant areas of the larger intraorganizational network.

Surprisingly, the weighted communities control variable never reaches statistical significance. Moreover, clustering and repeated ties both show some evidence of *increasing* the formation of ties among distant members of a site. One possible explanation for this finding is that as clustering and repeated interaction among members increases, the network structure becomes more globally differentiated; in so doing, it opens more opportunities for people to bridge distant groups (c.f., Watts and Strogatz, 1998). Therefore, clustering and repeated ties may serve as better controls for bridging opportunities than my weighted communities measure (i.e., the ratio of network communities to network members).

As anticipated, site age has a consistent negative association with bridging. The magnitude of this effect is non-negligible—a one month increase in age corresponds to a decrease in bridging of about 2.3%. In unreported analyses, I also tested for a curvilinear (inverted U-shaped) relationship between age and bridging, but found little evidence to suggest that the negative effect of age softens as a site grows older. The workweek effects dummy variables are statistically significant, with highly consistent coefficient magnitudes across models. Relative to Sundays, the baseline category, site activity that takes place on Mondays is associated with roughly a 46% increase in next-day bridging tie formation.[10] This increase over the baseline diminishes through-

[10]Unless otherwise noted, all marginal effects estimates for the control variables are based on Model 1 of Table 4.4.

out the week, until Saturdays, which are associated with a 21% decrease in next-day bridging.

Hypothesis 6 proposes the existence of a curvilinear relationship between the size of an organization's knowledge categorization system and bridging tie formation. Increases in the number of categories should help people locate peers outside their immediate network neighborhood, but beyond a certain point, additional refinements may prove to be cognitively overwhelming. Model 2 of Table 4.4 tests this hypothesis by adding measures of knowledge categorization system size and a quadratic term, both of which are significant in the predicted directions. However, the coefficient on the quadratic term is roughly an order of magnitude smaller than the main effect of knowledge categorization system size. This suggests that although there is evidence of a curvilinear relationship between knowledge categorization system size and bridging, the effect appears to flatten out as more and more categories are added, rather than fully inverting. To put the coefficients in perspective, consider that a one standard deviation increase in knowledge categorization system size is associated with more than two-and-a-half-fold increase in bridging ties. In sum, these findings offer support for Hypothesis 6.

Model 3 of Table 4.4 tests Hypothesis 7, which anticipated a negative association between decoupling and bridging tie formation. Recall that decoupling occurs when there is a lack of correspondence between an organization's knowledge categorization system and its informal knowledge domains. Model 3's negative and significant coefficient for decoupling supports the idea that bridging suffers when members do not have an accurate map of the distribution of knowledge within their organizations. The magnitude of the coefficient is also notable—a one standard deviation increase in decoupling corresponds to a 17% decrease in bridging ties at time $t + 1$.

Finally, Hypothesis 8 predicted that when norms of evaluation vary widely along the lines of an organization's knowledge categorization system, members will be less

Table 4.4: Models of Bridging Ties†

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Members								
Interest homophily$_{[t-7,t]}$	−1.531∗∗∗	−1.012∗∗	−1.568∗∗∗	−1.611∗∗∗	−1.119∗∗∗	−1.106∗∗∗	−1.662∗∗∗	−1.221∗∗∗
	(0.385)	(0.323)	(0.351)	(0.382)	(0.310)	(0.324)	(0.346)	(0.310)
Connectivity homophily$_{[t-7,t]}$	−0.897∗∗∗	−1.026∗∗∗	−0.832∗∗∗	−0.922∗∗∗	−0.961∗∗∗	−1.051∗∗∗	−0.858∗∗∗	−0.986∗∗∗
	(0.151)	(0.140)	(0.139)	(0.155)	(0.133)	(0.146)	(0.144)	(0.138)
Evaluation harshness$_{[t-7,t]}$	−0.707	−0.775	−0.713	0.004	−0.755	0.033	0.114	0.104
	(0.517)	(0.510)	(0.489)	(0.531)	(0.490)	(0.527)	(0.507)	(0.510)
Network Structure								
Communities (weighted)$_{[t-7,t]}$	0.051	0.148	0.088	0.065	0.175	0.144	0.105	0.171
	(0.166)	(0.159)	(0.158)	(0.166)	(0.155)	(0.159)	(0.158)	(0.155)
Clustering$_{[t-7,t]}$	0.413	2.174∗∗∗	1.191∗∗	0.433	2.384∗∗∗	2.299∗∗∗	1.240∗∗	2.523∗∗∗
	(0.440)	(0.433)	(0.415)	(0.429)	(0.426)	(0.436)	(0.404)	(0.429)
Repeat ties$_{[t-7,t]}$	3.397∗∗∗	2.581∗∗∗	2.921∗∗∗	3.303∗∗∗	2.388∗∗∗	2.443∗∗∗	2.796∗∗∗	2.235∗∗∗
	(0.336)	(0.280)	(0.287)	(0.332)	(0.268)	(0.278)	(0.285)	(0.267)
Centralization$_{[t-7,t]}$	−0.707∗∗	−0.412∗	−0.535∗∗	−0.773∗∗∗	−0.351+	−0.467∗	−0.606∗∗	−0.407∗
	(0.215)	(0.205)	(0.195)	(0.220)	(0.196)	(0.208)	(0.198)	(0.198)
Year Effects								
	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Workweek Effects								
Monday	0.455∗∗∗	0.456∗∗∗	0.458∗∗∗	0.455∗∗∗	0.458∗∗∗	0.456∗∗∗	0.458∗∗∗	0.458∗∗∗
	(0.014)	(0.014)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)
Tuesday	0.331∗∗∗	0.341∗∗∗	0.341∗∗∗	0.332∗∗∗	0.347∗∗∗	0.342∗∗∗	0.342∗∗∗	0.347∗∗∗
	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)	(0.017)	(0.017)	(0.016)
Wednesday	0.280∗∗∗	0.292∗∗∗	0.292∗∗∗	0.281∗∗∗	0.299∗∗∗	0.292∗∗∗	0.293∗∗∗	0.299∗∗∗
	(0.036)	(0.035)	(0.036)	(0.036)	(0.035)	(0.035)	(0.036)	(0.035)
Thursday	0.286∗∗∗	0.299∗∗∗	0.301∗∗∗	0.286∗∗∗	0.308∗∗∗	0.299∗∗∗	0.302∗∗∗	0.309∗∗∗
	(0.016)	(0.015)	(0.016)	(0.016)	(0.015)	(0.015)	(0.016)	(0.015)
Friday	0.222∗∗∗	0.234∗∗∗	0.234∗∗∗	0.222∗∗∗	0.241∗∗∗	0.234∗∗∗	0.234∗∗∗	0.241∗∗∗
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Saturday	−0.207∗∗∗	−0.198∗∗∗	−0.198∗∗∗	−0.207∗∗∗	−0.192∗∗∗	−0.197∗∗∗	−0.197∗∗∗	−0.192∗∗∗
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Site								
Age (months)$_t$	−0.023∗∗∗	−0.016∗∗∗	−0.023∗∗∗	−0.020∗∗∗	−0.017∗∗∗	−0.012∗∗	−0.020∗∗∗	−0.014∗∗
	(0.004)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
New members (log)$_t$	0.416∗∗∗	0.389∗∗∗	0.391∗∗∗	0.415∗∗∗	0.376∗∗∗	0.389∗∗∗	0.389∗∗∗	0.375∗∗∗

Table 4.4 (Continued)

	(0.017)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.015)
Prominent member loss $(\log)_{[t-30,t]}$	0.009	−0.196***	0.006	0.023	−0.166***	−0.181***	0.022	−0.149***
	(0.020)	(0.032)	(0.018)	(0.019)	(0.031)	(0.031)	(0.018)	(0.030)
Informal knowledge domains$_t$	0.002***	0.002***	0.001***	0.001***	0.002***	0.002***	0.001***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Independent Variables								
Knowledge categorization system size $(\log)_t$		1.079***			0.922***	1.067***		0.904***
		(0.102)			(0.100)	(0.108)		(0.103)
Knowledge categorization system size $(\log)_t^2$		−0.130***			−0.096***	−0.150***		−0.115***
		(0.022)			(0.023)	(0.023)		(0.024)
Decoupling$_t$			−0.087***		−0.062***		−0.090***	−0.064***
			(0.007)		(0.006)		(0.007)	(0.006)
Evaluation heterogeneity$_t$				−0.066***		−0.076***	−0.076***	−0.081***
				(0.015)		(0.014)	(0.015)	(0.014)
Constant	2.537***	2.553***	2.392***	2.811***	2.514***	2.784***	2.707***	2.757***
	(0.153)	(0.123)	(0.152)	(0.158)	(0.122)	(0.140)	(0.154)	(0.138)
N	18,534	18,534	18,534	18,534	18,534	18,534	18,534	18,534
R^2	0.34	0.37	0.36	0.35	0.38	0.37	0.36	0.38
d.f.	21	23	22	22	24	24	23	24
Sites	25	25	25	25	25	25	25	25

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; two tailed tests.

† All models are derived from conditional fixed effects OLS regressions with Driscoll-Kraay standard errors quasi-maximum (in parentheses).

likely to venture outside their comfort zones and therefore fewer bridging ties should form. Model 6 of Table 4.4 introduces the measure of evaluation heterogeneity to test this hypothesis. The coefficient is negative and statistically significant—a one standard deviation increase in evaluation heterogeneity is associated with a a 12% decrease in bridging. As hypothesized, when people are less certain about how they will be evaluated, they tend to build relationships in familiar domains. Note that this effect holds even when accounting for the overall tendency of site members to be critical of their peers, as captured by the measure of evaluation harshness. Interestingly, this latter measure never has a significant association with bridging tie formation.

4.7 Robustness Checks

I examined the robustness of the findings to a range of different model specifications and alternative explanations. First, I considered whether an interaction between the size of an organization's knowledge categorization system and the number of informal knowledge domains could provide a simpler account than the one proposed by Hypothesis 7 (on decoupling). One might predict, for example, that bridging may decline as the size of a knowledge categorization system grows large relative to the number of informal knowledge domains. Although simpler, I suggest that this alternative explanation is less appropriate because the theory proposed by this chapter does not address size of the categorization system relative to the informal distribution of knowledge, but rather only considers their relationship. I did, however, find a negative and significant association—as shown in Model 1 of Table 4.5—which suggests that as the size of the knowledge categorization system increases relative to the number of informal domains, bridging should decrease. The magnitude of the interaction, however, is small in comparison to the main effect of both knowledge categorization system size and the number of informal knowledge domains.

Table 4.5: Robustness Checks†

	Model 1 Bridging Ties	Model 2 Bridging Ties	Model 3 Bridging Ties No SO	Model 4 Bridging Ties $[t+1, t+3]$	Model 5 Bridging Ties 1 SD	Model 6 Bridging Ties 0.5 SD
Members						
Interest homophily$_{[t-7,t]}$	−1.066∗∗∗	−1.153∗∗∗	−1.446∗∗∗	−1.274∗∗∗	−2.393∗∗∗	−2.355∗∗∗
	(0.278)	(0.269)	(0.285)	(0.344)	(0.277)	(0.270)
Connectivity homophily$_{[t-7,t]}$	−0.776∗∗∗	−0.741∗∗∗	−0.794∗∗∗	−1.224∗∗∗	−0.614∗∗∗	−0.564∗∗∗
	(0.138)	(0.133)	(0.131)	(0.164)	(0.120)	(0.110)
Evaluation harshness$_{[t-7,t]}$	−0.148	−0.087	−0.030	−0.146	−0.383	−0.294
	(0.505)	(0.496)	(0.487)	(0.582)	(0.408)	(0.417)
Network Structure						
Communities (weighted)$_{[t-7,t]}$	0.007	0.034	0.074	0.127	0.137	−0.026
	(0.148)	(0.145)	(0.143)	(0.150)	(0.116)	(0.103)
Clustering$_{[t-7,t]}$	2.013∗∗∗	2.192∗∗∗	2.555∗∗∗	2.810∗∗∗	0.610	1.214∗∗∗
	(0.392)	(0.387)	(0.400)	(0.486)	(0.381)	(0.365)
Repeat ties$_{[t-7,t]}$	2.270∗∗∗	2.125∗∗∗	2.102∗∗∗	2.333∗∗∗	1.512∗∗∗	1.665∗∗∗
	(0.252)	(0.246)	(0.248)	(0.285)	(0.254)	(0.216)
Centralization$_{[t-7,t]}$	−0.153	−0.124	−0.150	−0.557∗	−0.513∗∗	−0.470∗∗
	(0.187)	(0.181)	(0.179)	(0.226)	(0.168)	(0.161)
Year Effects						
	Yes	Yes	Yes	Yes	Yes	Yes
Workweek Effects						
Monday	0.458∗∗∗	0.459∗∗∗	0.448∗∗∗	0.135∗∗∗	0.464∗∗∗	0.535∗∗∗
	(0.014)	(0.014)	(0.015)	(0.009)	(0.015)	(0.011)
Tuesday	0.382∗∗∗	0.384∗∗∗	0.370∗∗∗	−0.008	0.383∗∗∗	0.457∗∗∗
	(0.016)	(0.016)	(0.017)	(0.014)	(0.016)	(0.012)
Wednesday	0.339∗∗∗	0.342∗∗∗	0.329∗∗∗	−0.094∗	0.343∗∗∗	0.391∗∗∗
	(0.035)	(0.035)	(0.034)	(0.041)	(0.035)	(0.045)
Thursday	0.349∗∗∗	0.354∗∗∗	0.341∗∗∗	−0.199∗∗∗	0.371∗∗∗	0.413∗∗∗
	(0.015)	(0.015)	(0.016)	(0.015)	(0.015)	(0.014)
Friday	0.284∗∗∗	0.287∗∗∗	0.274∗∗∗	−0.375∗∗∗	0.269∗∗∗	0.323∗∗∗
	(0.017)	(0.017)	(0.018)	(0.014)	(0.017)	(0.015)
Saturday	−0.158∗∗∗	−0.155∗∗∗	−0.157∗∗∗	−0.360∗∗∗	−0.207∗∗∗	−0.204∗∗∗
	(0.015)	(0.015)	(0.016)	(0.012)	(0.015)	(0.013)
Site						
Age (months)$_t$	−0.010∗∗	−0.012∗∗	−0.009∗	−0.014∗∗	−0.010∗	−0.013∗∗∗

Table 4.5 (Continued)

	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)
New members $(\log)_t$	0.301∗∗∗	0.295∗∗∗	0.300∗∗∗	0.355∗∗∗	0.337∗∗∗	0.370∗∗∗
	(0.015)	(0.015)	(0.015)	(0.018)	(0.013)	(0.013)
Prominent member loss $(\log)_{[t-30,t]}$	−0.159∗∗∗	−0.136∗∗∗	−0.131∗∗∗	−0.181∗∗∗	−0.036+	−0.101∗∗∗
	(0.025)	(0.024)	(0.027)	(0.033)	(0.020)	(0.022)
Informal knowledge domains$_t$	0.028∗∗∗	0.027∗∗∗	0.008∗∗∗	0.002∗∗∗	0.000∗∗∗	0.001∗∗∗
	(0.002)	(0.002)	(0.001)	(0.000)	(0.000)	(0.000)
Independent Variables						
Knowledge categorization system size $(\log)_t$	0.893∗∗∗	0.781∗∗∗	0.493∗∗∗	1.068∗∗∗	0.946∗∗∗	0.871∗∗∗
	(0.088)	(0.084)	(0.105)	(0.117)	(0.083)	(0.091)
Knowledge categorization system size $(\log)_t^2$	−0.062∗∗	−0.041∗	−0.178∗∗∗	−0.132∗∗∗	−0.034+	−0.104∗∗∗
	(0.020)	(0.020)	(0.029)	(0.026)	(0.020)	(0.022)
Knowledge categorization system size $(\log)_t$	−0.003∗∗∗	−0.002∗∗∗				
× Informal knowledge domains	(0.000)	(0.000)				
Decoupling$_t$		−0.048∗∗∗	−0.061∗∗∗	−0.067∗∗∗	−0.055∗∗∗	−0.051∗∗∗
		(0.006)	(0.006)	(0.008)	(0.006)	(0.005)
Evaluation heterogeneity$_t$	−0.051∗∗∗	−0.056∗∗∗	−0.063∗∗∗	−0.078∗∗∗	−0.046∗∗∗	−0.048∗∗∗
	(0.012)	(0.011)	(0.012)	(0.014)	(0.011)	(0.010)
Constant	2.516∗∗∗	2.509∗∗∗	3.298∗∗∗	3.968∗∗∗	2.813∗∗∗	3.807∗∗∗
	(0.114)	(0.113)	(0.053)	(0.151)	(0.122)	(0.107)
N	18,534	18,534	17,074	18,534	18,534	18,534
R^2	0.41	0.42	0.36	0.37	0.45	0.52
d.f.	25	26	25	25	25	25
Sites	25	25	24	25	25	25

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; two tailed tests.

† All models are derived from conditional fixed effects OLS regressions with Driscoll-Kraay standard errors quasi-maximum (in parentheses).

Second, as noted earlier, Stack Overflow is both older and an order of magnitude larger than any other site on Stack Exchange. To ensure that the findings are not driven by any unusual influence of Stack Overflow, I removed the site from the sample and reran the analyses. The results, reported in Model 3 of Table 4.5, are supportive of the main findings. Interestingly, note that the coefficient estimate for the main effect of knowledge categorization system size decreases by more than 45% while the negative estimate for the quadratic term increases by nearly 55%, compared to Model 8 of Table 4.4. This suggests that relative to Stack Overflow, increases in knowledge categorization system size on other sites actually has a more inverted-U shaped association with bridging tie formation, as predicted by Hypothesis 6.

Third, although the majority of replies to questions on Stack Exchange come within just a few hours of the initial post, it is possible that a 24-hour window may be too narrow and therefore artificially truncate relevant new bridging ties, especially given that members of Stack Exchange sites come from many different time zones. To evaluate the robustness of the findings to a larger forward lag, Model 4 of Table 4.5 uses a dependent variable that includes all bridging ties formed between (and including) $t+1$ and $t+3$. The results are similar to those reported in the primary models of Table 4.4.

Finally, I considered two alternative specifications of the dependent variable that did not rely on community or network structure to measure bridging ties. Instead, these measures identify bridging among interacting members in terms of the similarity of their interests.[11] Specifically, I define bridging ties as exchanges between members who are 1 (Table 4.5, Model 5) and 0.5 (Table 4.5, Model 6) standard deviations away from the mean similarity of interacting members on their respective Stack Exchange site as of time t. Both alternative specifications are in line with the main findings.

[11]For more details on the procedure used to calculate these measures, see Appendix A.

4.8 Discussion and Conclusion

Knowledge-intensive organizations depend on effective communication and collaboration networks among their members. These networks help an organization develop new capabilities and ultimately achieve better outcomes by enhancing integration among people who are otherwise highly specialized (Grant, 1996; Lawrence and Lorsch, 1967). Building on insights from research on organizational design and studies of individual's social ties, however, recent investigations demonstrate that the degree to which communication and collaboration networks lead to better outcomes is contingent on both the global configuration of those networks and the nature of the organization's goals and environment. For example, although increasing connectivity among members of an organization is often beneficial, eliminating ties and isolating subgroups can be helpful at times for promoting innovation by enabling parallel problem solving and retaining diversity (Fang et al., 2010; Funk, 2014; Lazer and Friedman, 2007).

An important implication of these findings is that as goals evolve and environmental conditions change, organizations must also find ways to reshape their networks. Yet surprisingly little work addresses the methods organizations might use to implement such changes (Davis, 2008). In this study, I contribute to research in this area by examining how knowledge categorization systems and evaluation norms influence bridging tie formation in intraorganizational networks. Bridging ties are a useful starting point for developing theories of change because adding even a small number of such connections can dramatically alter global properties of networks. By creating and strengthening connections among otherwise distant groups of people, bridging ties transform the flow of knowledge among the members of an organization and help foster tighter integration (Kleinbaum and Tushman, 2007; Watts and Strogatz, 1998).

I hypothesize three conditions under which features of an organization's knowledge-categorization system are likely to inhibit bridging and therefore network change.

First, although size can be beneficial, when knowledge categorization systems grow too large, cognitive limitations may make them difficult to use and therefore members of an organization should be less likely to connect with distant peers. Second, knowledge categorization systems can be helpful maps, but when they become decoupled from the actual distribution of knowledge within an organization, their value is likely to fade, along with bridging. Finally, when the standards used to evaluate new and existing knowledge contributions are highly variable across an organization's knowledge categorization system, people should be less likely to share knowledge outside their comfort zones and therefore will make fewer connections with unfamiliar peers. Using data on 23 million knowledge-sharing exchanges among 1.3 million members of 25 online Q&A communities, I find strong support for all three hypotheses.

The conclusions of this study must be viewed in light of some limitations. First, despite the overall strengths, further research is needed to specify the representativeness of the Stack Exchange data and to replicate the findings in other settings. Clearly, the intraorganizational networks found on virtual Q&A communities like Stack Exchange differ from many of the kinds of connections that exist among people in brick-and-mortar organizations. Nevertheless, there are many similarities between online and offline organizations, and therefore at least some of the findings are likely to transfer with little adaptation. Recent work demonstrates, for instance, that many well-known properties of individual's 'real world' social networks are also found in virtual communities (Burt, 2013). Moreover, setting aside their correspondence to offline networks, the data are of interest in their own right. As more and more organizations adopt systems like corporate wikis, enterprise social software, and other tools that facilitate virtual collaboration, it becomes important to understand how the social networks that exist by virtue of such tools influence organizational outcomes (Mcafee, 2006). Stack Exchange shares many features with these systems.

Second, as with all observational studies, it is possible that the statistical esti-

mates are influenced by unknown or unmeasurable factors. I have sought to account for the effects of many potential confounding variables by relying only on comparisons within Q&A communities, which eliminates the influence of all time invariant heterogeneity on the statistical estimates. Moreover, I have also included controls for a broad array of factors identified by prior research as having an effect on bridging tie formation. Finally, I have found no evidence to suggest that leaders or members of Stack Exchange seek to strategically promote bridging or other tie formation, which should help to ease some concerns about one major source of endogeneity that arises in much network research (Ahuja et al., 2012). Despite these corrections, care should be taken when making inferences from the findings.

Although not without limitations, the results of this study have several noteworthy implications.

Network dynamics. This study contributes to the growing literature on the evolutionary dynamics of networks. Although researchers have recently made substantial progress in accounting for the origins of different network structures, most existing investigations focus on how individual actors—whether they are people or organizations—come to occupy advantageous positions in a social structure. Far less is known about the factors that enable and constrain the evolution of global network configurations. This oversight especially problematic—even for scholars who focus on individual actors—in light of recent studies showing that the performance effects of networks are a function of both local and global position (Paruchuri, 2010; Sytch and Tatarynowicz, 2014). Furthermore, research on the dynamics of global social structures is necessary to further develop theories that link networks to organizational outcomes. Organizations contain collections of actors; thus, the ties of individual actors may not be as relevant to collective performance as the pattern of relations among them. My analysis of the relationship between knowledge categorization systems, evaluation norms, and bridging tie formation takes one step towards filling this

gap in the literature by revealing how two pervasive features of organizations enable and constrain network change, while also suggesting that future research on network dynamics might fruitfully examine the influence of other components of organizational structure.

More broadly, my study suggests that research on the dynamics of global network structures can open new opportunities for linking theory and research with practice. Potential applications are clearest in the area of intraorganizational network change, where the findings of this chapter suggest that to promote exchange among distant groups, leaders should consider implementing policies that streamline knowledge search, and members should be more attentive to how they respond to contributions from outsiders. However, research on global network change is also likely to have valuable practical implications at larger levels of analysis. For example, regulators might try to increase or decrease the global cohesiveness of director interlock or alliance networks in order to meet certain policy objectives (Davis and Mizruchi, 1999; Mizruchi, 2013). Relatedly, research on patient transfer networks among hospitals finds that although such transfers have significant consequences for outcomes, the methods hospitals use to initiate and accept transfers from outside institutions are often unclear and the structures of the resulting networks are likely suboptimal (Iwashyna et al., 2009). Investigations of global network dynamics could therefore be attractive for leaders who seek to build more effective regional transfer networks that lower the cost and increase the quality of care.

Organizational change and growth. To ensure their survival, organizations must adapt as their environments change (O'Reilly III and Tushman, 2008; Teece et al., 1997). However, adaptation is often a major challenge, and according to some perspectives, may even add to the risk of failure (Hannan and Freeman, 1984). Although many factors inhibit the success of efforts to change, one leading cause is the unwillingness or inability of an organization's members to adapt their social relations

to match new routines. Oftentimes, resistance to change arises because some members stand to lose status and other resources (Battilana and Casciaro, 2012; Paruchuri et al., 2006). However, the challenges people face in effectively navigating a transformed organization also play a role, especially when those transformations entail substantial disruptions like the restructuring of divisions or major acquisitions. Notably, the findings of this study suggest that if efforts at change introduce decoupling between an organization's knowledge categorization system and informal knowledge domains, or if they increase uncertainty about evaluation, members will be less likely to establish ties with new (i.e., formerly distant) peers, which may curtail organizational change.

Although important in established firms, communication and collaboration network dynamics have major consequences for growth and performance in startups. For instance, Stinchcombe (1965) argues that communication patterns are central to understanding why new firms have high failure rates—what he terms the "liability of newness." "For some time until roles are defined," Stinchcombe writes, "people who need to know things are left to one side of communication channels. John thinks George is doing what George thinks John is doing" (1965, 148-9). Communication difficulties inhibit the flow of potentially valuable knowledge among relevant parties, while the ambiguous and continually evolving nature of roles in new firms poses challenges for employees in search of relevant collaborators or other assistance. Put differently, the primary benefits that social networks bring to innovation—knowledge flows and supportive relationships—are likely difficult for entrepreneurs to attain in their organizations. The findings of this study could offer guidance to entrepreneurs who seek to build more effective communication and collaboration networks within their organizations that help them reap those benefits and manage growth more effectively.

Collaboration in knowledge-intensive organizations. Finally, this study

has implications for the design of policies intended to promote collaboration. Among researchers and policymakers alike, there is growing recognition that making progress on major scientific and technological challenges requires collaborations between experts from diverse disciplines. Although integrating people with different specializations can be valuable for all types of organizations (Grant, 1996), the potential payoff—and roadblocks—to increasing collaboration are especially apparent in universities. In recent years, many academic institutions have instituted programs designed to strengthen interdisciplinary exchanges, but so far, evidence in support of their success is mixed (Porter and Rafols, 2009). The findings of this study could be useful for administrators who seek to promote tighter integration among seemingly disparate scholars at their institutions. Notably, the results suggest that challenges in locating colleagues may limit the number of collaborations that bridge disciplinary boundaries. Perhaps counterintuitively, these challenges could arise in part from efforts to enhance interdisciplinarity through the creation of centers and institutes that lie and the interstices between established disciplinary departments, at least to the extent that the new organizational units lead to decoupling between a university's predominant knowledge categorization system (e.g., disciplinary departments) and its informal knowledge domains (e.g., what scholars actually study).

Finally, the results of this study correspond with other research that suggests differences in how academic contributions are evaluated across fields of study influences the propensity for scholars to venture outside their home departments (Rhoten and Parker, 2004). In light of these findings, future research could consider the effectiveness of a variety of policies on enhancing integration across disciplines. For instance, one way for universities to increase interdisciplinary collaboration might be to reduce uncertainty about evaluation for scholars who seek to bridge across disparate fields, perhaps by adapting formal policies used to make promotions and allocate resources. Interdisciplinary research teams could also increase their effectiveness by

having explicit discussions about the yardsticks used to measure progress for their group efforts. Regardless of the approach, the prospects for successful integration in any knowledge intensive organization likely depend substantially on the ability of leaders and members to reduce unnecessary barriers to the already difficult task of bridging.

APPENDICES

APPENDIX A

Comparing Members' Interests

Several variables used in Chapter IV require a measure of the degree to which the members of a site share similar (or have divergent) interests. Although one could envision an array of possible approaches, I approximate members' interests at time t as a function of the tags in which they are active. I consider a member active within a tag c if he or she had asked, answered, or commented on a question tagged as c at any point in time between the date he or she joined a focal site and the date of measurement. More precisely, for each site at time t, I create an $m \times n$ matrix $\mathbf{S}$ in which rows index members, columns index tags, and entries correspond to the frequency of activity (i.e., questions, answers, and comments) for all member $\times$ tag dyads as of time t.

In principle, the raw frequencies recorded in $\mathbf{S}$ could be used to estimate members' interest similarities. However, tags vary dramatically in terms of their frequency of use; therefore, knowing that two members both have substantial activity in a popular tag reveals less information about their interests than if they both had even a minor amount of activity in a rare one. Similarly, members differ substantially with respect to the magnitude of their activity on a site; consequently, two people may have a relatively large absolute number of tags in common, but only because they are active

in many tags—not as a result of having similar interests.

To account for differences in tag popularity and member activity, I adjust the entries of $\mathbf{S}$ using a composite term frequency-inverse document frequency weighting (tf-idf) approach originally developed to compare documents based on the similarity of their constituent words (Manning et al., 2008). In my application, tags are analogous to terms (or words), while members correspond to documents. Inverse document frequency accounts for whether a term (tag) is rare across documents (members). For a term t, the measure is calculated as

$$\text{idf}_t = \log \frac{N}{\text{df}_\text{t}}, \tag{A.1}$$

where N is the total number of documents and df_t is the number of documents in which the term appears. A composite weight for each entry of $\mathbf{S}$ is obtained using

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_\text{t}, \tag{A.2}$$

where $\text{tf}_{t,d}$ is the raw frequency of term (tag) t for some document (member) d. To account for changes over time, I update tf-idf weights daily for each site. After applying tf-idf weights to $\mathbf{S}$, I then compare the interests of any pair of members x and y with tag vectors $x = (s_{x1}, s_{x2}, \ldots, s_{xn-1}, s_{xn})$ and $y = (s_{y1}, s_{y2}, \ldots, s_{yn-1}, s_{yn})$ using their cosine similarity, defined as

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{A.3}$$

The intuition behind this measure is that the similarity of two n-dimensional vectors can be approximated by the degree of the angle between them. Because term frequencies (or other commonly used weights) are never negative, the angle between two vectors is at most 90 degrees and therefore the measure ranges from 0 to 1. Cosine similarity is attractive over other metrics because it is not sensitive

to document length. Although doubling the size of a document would change the magnitude of its corresponding vector, the direction (and therefore angle relative to other vectors) would remain the same.

APPENDIX B

Identifying Informal Knowledge Domains

Conceptually, informal knowledge domains refer to latent groups of knowledge that emerge as members use and contribute to an organization's knowledge base. Using the Stack Exchange data, I identified informal knowledge domains by partitioning questions into clusters according to their textual similarity. I chose to focus on questions and did not supply the clustering algorithm with the text of answers or comments for several reasons. First, on Stack Exchange, only questions are tagged; therefore, by clustering questions, I am more accurately able to compare the correspondence between a site's knowledge categorization system and its informal knowledge domains. Second, as noted in the text, threads are anchored by a single question and consequently, even if they were included, comments and answers should align with clusters that are similar to their parent questions. Finally, given the size of the data, identifying clusters is computationally intensive; obtaining results for Stack Overflow alone took nearly one week using 20 processors with 500 gigabytes of memory.

To transform natural language text into a format amenable to clustering, I use a vector space model in which questions (hereafter referred to as documents) are represented as weighted vectors of terms (Salton et al., 1975). Before assembling the vectors, I perform several common preprocessing steps (Manning et al., 2008). First,

following the recommendations of Barua et al. (2012), I removed all large blocks of code, which are delimited on Stack Exchange by `<code>` and `</code>` tags. Next, I separate the strings of characters that constitute each document into distinct words (or tokens). Although seemingly straightforward, identifying accurate word boundaries can prove vexing, for example, with compound names (e.g., `New York`) or words that contain punctuation (e.g., `www.google.com`). After comparing several tokenization algorithms, I found splitting on whitespace to be ideal for identifying meaningful words in the Stack Exchange data, especially since many documents contain small snippets of code where punctuation often does not correspond to word boundaries. Following tokenization, I then convert all tokens to lowercase (e.g., `ipad`, `iPad`, `IPAD` $\Rightarrow$ `ipad`), and apply Porter's (1980) stemming algorithm to convert words to their base form (e.g., `boats`, `boating`, `boater`$\Rightarrow$ `boat`). Once I finish transforming the tokens by folding cases and stemming, I then remove common stop words like `the`, `is`, `at`, `which`, and `on` that appear frequently in almost all documents and contribute little useful information for clustering. I also eliminate extremely rare tokens that appear in very few documents and are most often misspellings.

After preprocessing, I then assemble the document vectors into a document $\times$ term matrix and apply daily-updated tf-idf weights analogous to those described in Appendix A. Even after preprocessing, the documents contain many thousands of tokens and the resulting matrices are extremely large. To facilitate computation, I reduce the dimensionality of the document $\times$ term matrices using principal components analysis. Following the dimension reduction, I then compute pairwise negative Euclidean distances among documents and perform clustering using Frey and Dueck (2007) affinity propagation algorithm, as described in Section 4.4.4. The resulting cluster assignments correspond to informal knowledge domains.

APPENDIX C

Adjusting Decoupling for Chance Agreement

The measure of decoupling I present in Equation 4.1 is equivalent to the mutual information of two discrete random variables (Shannon and Weaver, 1949). My use of mutual information is motivated by an array of research that employs information theoretic measures to evaluate agreement among different clustering algorithms. Recent simulation studies show, however, that the mutual information among two clustering solutions increases monotonically with the number of clusters (Vinh et al., 2009); therefore, it is necessary to adjust the measure to account for potential correspondence due to random chance. For the purposes of exposition, Equation 4.1 presents the uncorrected form of mutual information; however, for use in the statistical models, I compute an adjusted version of decoupling, following Vinh et al. (2010), as

$$D'_{it}(C,K) = \frac{D_{it}(C,K) - \mathbb{E}\{D_{it}(C,K)\}}{\sqrt{H(C)H(K)} - \mathbb{E}\{D_{it}(C,K)\}}, \tag{C.1}$$

where $\mathbb{E}\{D_{it}(C,K)\}$ is the expected value of decoupling and $\sqrt{H(C)H(K)}$ is the measure's upper bound. $H(C)$ and $H(K)$ refer to the entropy of tags and clusters, respectively, where entropy is defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x), \tag{C.2}$$

and $p(x)$ is the probability density function of the discrete random variable X. Following the discussion in 4.4.4, I subtract $D'_{it}(C, K)$ from 1 so that higher values correspond to greater decoupling, and multiply the resulting values by 100.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abbott, A. (2001). *Chaos of disciplines*. Chicago: University of Chicago Press.

Adamic, L. A., J. Zhang, E. Bakshy, and M. Ackerman (2008). Knowledge sharing and Yahoo Answers: Everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pp. 665–674. New York: ACM.

Agarwal, R., M. Ganco, and R. H. Ziedonis (2009). Reputations for toughness in patent enforcement: Implications for knowledge spillovers via inventor mobility. *Strategic Management Journal 30*, 1349–1374.

Agrawal, A., I. Cockburn, and J. McHale (2006). Gone but not forgotten: Knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography 6*, 571–591.

Ahuja, G. (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative Science Quarterly 45*, 425–455.

Ahuja, G., G. Soda, and A. Zaheer (2012). The genesis and dynamics of organizational networks. *Organization Science 23*, 434–448.

Allen, T. J. (1977). *Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization*. Cambridge, MA: MIT Press.

Almeida, P. and B. Kogut (1999). Localization of knowledge and the mobility of engineers in regional networks. *Management Science 45*, 905–917.

Alnuaimi, T., T. Opsahl, and G. George (2012). Innovating in the periphery: The impact of local and foreign inventor mobility on the value of Indian patents. *Research Policy 41*, 1534–1543.

Alnuaimi, T., J. Singh, and G. George (2012). Not with my own: Long-term effects of cross-country collaboration on subsidiary innovation in emerging economies versus advanced economies. *Journal of Economic Geography 12*, 943–968.

Aral, S. and M. Van Alstyne (2011). The diversity-bandwidth trade-off. *American Journal of Sociology 117*, 90–171.

Argote, L. (2013). Organizational memory. In *Organizational learning: Creating, retaining and transferring knowledge*, pp. 85–113. New York: Springer.

Argote, L. and P. Ingram (2000). Knowledge transfer: A basis for competitive advantage in firms. *Organizational Behavior and Human Decision Processes 82*, 150–169.

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. New York: Free Press.

Atwood, J. (2008). Introducing Stackoverflow.com.

Atwood, J. (2009). A theory of moderation.

Audretsch, D. B. and M. P. Feldman (1996). R&D spillovers and the geography of innovation and production. *American Economic Review 86*, 630–640.

Azoulay, P., W. Ding, and T. Stuart (2007). The determinants of faculty patenting behavior: Demographics or opportunities? *Journal of Economic Behavior & Organization 63*, 599–623.

Azoulay, P., C. Liu, and T. Stuart (2009). Social influence given (partially) deliberate matching: Career imprints in the creation of academic entrepreneurs. Working paper, Harvard Business School, Boston.

Barua, A., S. W. Thomas, and A. E. Hassan (2012). What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empirical Software Engineering 19*, 1–36.

Bathelt, H., A. Malmberg, and P. Maskell (2004). Clusters and knowledge: Local buzz, global pipelines and the process of knowledge creation. *Progress in Human Geography 28*, 31–56.

Battilana, J. and T. Casciaro (2012). Change agents, networks, and institutions: A contingency theory of organizational change. *Academy of Management Journal 55*, 381–398.

Bawa, R. (2007). Nanotechnology patent proliferation and the crisis at the U.S. Patent Office. *Albany Law Journal of Science and Technology 17*, 699–735.

Bawa, R., S. R. Bawa, S. Maebius, and C. Iver (2006). Bionanotechnology patents: Challenges and opportunities. In J. Bronzino (Ed.), *Biomedical engineering handbook*, pp. 29–1–29–16. CRC Press, Boca Raton.

Beck, N. and J. N. Katz (2011). Modeling dynamics in time-series–cross-section political economy data. *Annual Review of Political Science 14*, 331–352.

Begel, A., J. Bosch, and M.-A. Storey (2013). Social networking meets software development: Perspectives from GitHub, MSDN, Stack Exchange, and TopCoder. *Software, IEEE 30*, 52–66.

Bell, G. G. (2005). Clusters, networks, and firm innovativeness. *Strategic Management Journal 26*, 287–295.

Bell, G. G. and A. Zaheer (2007). Geography, networks, and knowledge flow. *Organization Science 18*, 955–972.

Bernstein, E. S. (2012). The transparency paradox: A role for privacy in organizational learning and operational control. *Administrative Science Quarterly 57*, 181–216.

Blundell, R., R. Griffith, and J. Van Reenen (1995). Dynamic count data models of technological innovation. *Economic Journal 105*, 333–344.

Borgatti, S. P. and R. Cross (2003). A relational view of information seeking and learning in social networks. *Management Science 49*, 432–445.

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of theory and research for the sociology of education*, pp. 241–258. New York: Greenwood Press.

Bowker, G. C. and S. L. Star (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Brass, D. J. (2009). Connecting to brokers: Strategies for acquiring social capital. In V. O. Bartkus and J. H. Davis (Eds.), *Social capital: Reaching out, reaching in*, pp. 260–274. Northampton, MA: Edward Elgar Publishing.

Brass, D. J., J. Galaskiewicz, H. R. Greve, and W. Tsai (2004). Taking stock of networks and organizations: A multilevel perspective. *Academy of Management Journal 47*, 795–817.

Breschi, S. and F. Lissoni (2009). Mobility of skilled workers and co-invention networks: An anatomy of localized knowledge flows. *Journal of Economic Geography 9*, 439–468.

Brown, J. S. and P. Duguid (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science 2*, 40–57.

Brown, J. S. and P. Duguid (2000). *The social life of information*. Boston: Harvard Business School Press.

Buhr, H. and J. Owen-Smith (2010). Networks as institutional support: Law firm and venture capital relations and regional diversity in high technology IPOs. *Research in the Sociology of Work 21*, 95–126.

Burns, T. E. and G. M. Stalker (1961). *The management of innovation*. London: Tavistock.

Burt, R. S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Burt, R. S. (1997). The contingent value of social capital. *Administrative Science Quarterly 42*, 339–365.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology 110*, 349–399.

Burt, R. S. (2005). *Brokerage and closure: An introduction to social capital*. Oxford: Oxford University Press.

Burt, R. S. (2007). Secondhand brokerage: Evidence on the importance of local structure for managers, bankers, and analysts. *Academy of Management Journal 50*, 119–148.

Burt, R. S. (2010). *Neighbor networks: Competitive advantage local and personal*. Oxford: Oxford University Press.

Burt, R. S. (2013). Structural holes in virtual worlds. Working paper, University of Chicago Booth School of Business, Chicago.

Burt, R. S. and D. Ronchi (1990). Contested control in a large manufacturing plant. In J. Weesie and H. Flap (Eds.), *Social networks through time*, pp. 127–157. Utrecht: ISOR.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and applications*. New York: Cambridge University Press.

Christensen, C. M. (1997). *The innovator's dilemma*. New York: Collins Business Essentials.

Christie, L. S., R. D. Luce, and J. Macy Jr. (1952). Communication and learning in task-oriented groups. Research Laboratory of Electronics, Technical report #231, Massachusetts Institute of Technology, Cambridge, MA.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology 94*, S95–S120.

Coleman, J. S. (1990). *Foundations of social theory*. Cambridge, MA: Harvard University Press.

Dahlander, L. and D. A. McFarland (2013). Ties that last: Tie formation and persistence in research collaborations over time. *Administrative Science Quarterly 58*, 69–110.

Dalton, M. (1959). *Men who manage*. Hoboken, NJ: John Wiley & Sons.

Darby, M. R. and L. G. Zucker (2003). Grilichesian breakthroughs: Inventions of methods of inventing and firm entry in nanotechnology. Working paper, National Bureau of Economic Research, Cambridge, MA.

Davenport, T. H. and L. Prusak (1998). *Working knowledge: How organizations manage what they know*. Boston: Harvard Business School Press.

Davis, G. F. and M. S. Mizruchi (1999). The money center cannot hold: Commercial banks in the U.S. system of corporate governance. *Administrative Science Quarterly 44*, 215–239.

Davis, J. P. (2008). Network plasticity and collaborative innovation: Processes of network reorganization. *Academy of Management Proceedings 2008*, 1–7.

Davis, J. P. and K. M. Eisenhardt (2011). Rotating leadership and collaborative innovation. *Administrative Science Quarterly 56*, 159–201.

Dodds, P. S., D. J. Watts, and C. F. Sabel (2003). Information exchange and the robustness of organizational networks. *Proceedings of the National Academy of Sciences 100*, 12516–12521.

Driscoll, J. C. and A. C. Kraay (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics 80*, 549–560.

Drucker, P. F. (1985). *Innovation and entrepreneurship*. New York: Harper & Row.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly 44*, 350–383.

Eichenwald, K. (2012). Microsoft's lost decade. *Vanity Fair 54*, 108.

Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. *Science 321*, 395–399.

Fang, C., J. Lee, and M. A. Schilling (2010). Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. *Organization Science 21*, 625–642.

Feld, S. L. (1981). The focused organization of social ties. *American Journal of Sociology 86*, 1015–1035.

Fernandez, R. M. and R. V. Gould (1994). A dilemma of state power: Brokerage and influence in the national health policy domain. *American Journal of Sociology 99*, 1455–1491.

Fernandez-Mateo, I. (2007). Who pays the price of brokerage? Transferring constraint through price setting in the staffing sector. *American Sociological Review 72*, 291–317.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science 47*, 117–132.

Fleming, L., C. King, and A. I. Juda (2007). Small worlds and regional innovation. *Organization Science 18*, 938–954.

Fleming, L., S. Mingo, and D. Chen (2007). Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly 52*, 443–475.

Florida, R. L. (2002). *The rise of the creative class*. New York: Basic Books.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*, 75–174.

Freeman, J. H. and P. G. Audia (2006). Community ecology and the sociology of organizations. *Annual Review of Sociology 32*, 145–169.

Frey, B. J. and D. Dueck (2007). Clustering by passing messages between data points. *Science 315*, 972–976.

Funk, R. J. (2014). Making the most of where you are: Geography, networks, and innovation in organizations. *Academy of Management Journal 57*, 193–222.

Galbraith, J. R. (1973). *Designing complex organizations*. Reading, MA: Addison-Wesley.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago: University of Chicago Press.

Galunic, C., G. Ertug, and M. Gargiulo (2012). The positive externalities of social capital: Benefiting from senior brokers. *Academy of Management Journal 55*, 1213–1231.

Gassmann, O. and M. von Zedtwitz (1999). New concepts and trends in international R&D organization. *Research Policy 28*, 231–250.

Gerstner, W.-C., A. König, A. Enders, and D. C. Hambrick (2013). CEO narcissism, audience engagement, and organizational adoption of technological discontinuities. *Administrative Science Quarterly 58*, 257–291.

Gittelman, M. (2007). Does geography matter for science-based firms? Epistemic communities and the geography of research and patenting in biotechnology. *Organization Science 18*, 724–741.

Glynn, A. N. and K. M. Quinn (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis 18*, 36–56.

Golder, S. A. and B. A. Huberman (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science 32*, 198–208.

Gomes-Casseres, B., J. Hagedoorn, and A. B. Jaffe (2006). Do alliances promote knowledge flows? *Journal of Financial Economics 80*, 5–33.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Applications to poisson models. *Econometrica 52*, 701–720.

Graf, H. (2011). Gatekeepers in regional networks of innovators. *Cambridge Journal of Economics 35*, 173–198.

Granovetter, M. (1992). Problems of explanation in economic sociology. In *Networks and organizations: Structure, form, and action*, pp. 25–56. Boston, MA: Harvard Business School Press.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology 78*, 1360–1380.

Grant, R. M. (1996). Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science 7*, 375–387.

Grigoriou, K. and F. T. Rothaermel (2013). Structural microfoundations of innovation: The role of relational stars. *Journal of Management 40*, 586–615.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature 28*, 1661–1707.

Gulati, R., M. Sytch, and A. Tatarynowicz (2012). The rise and fall of small worlds: Exploring the dynamics of social structure. *Organization Science 23*, 449–471.

Guler, I. and A. Nerkar (2012). The impact of global and local cohesion on innovation in the pharmaceutical industry. *Strategic Management Journal 33*, 535–549.

Haas, M. R. and M. T. Hansen (2007). Different knowledge, different benefits: Toward a productivity perspective on knowledge sharing in organizations. *Strategic Management Journal 28*, 1133–1153.

Hall, B. H., A. Jaffe, and M. Trajtenberg (2005). Market value and patent citations. *RAND Journal of Economics 36*, 16–38.

Hannan, M. T. (1998). Rethinking age dependence in organizational mortality: Logical formalizations. *American Journal of Sociology 104*, 126–164.

Hannan, M. T. and J. Freeman (1984). Structural inertia and organizational change. *American Sociological Review 49*, 149–164.

Hannan, M. T., L. Pólos, and G. R. Carroll (2007). *Logics of organization theory: Audiences, codes, and ecologies*. Princeton, NJ: Princeton University Press.

Hansen, M. T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly 44*, 82–111.

Hansen, M. T., N. Nohria, and T. Tierney (1999). What's your strategy for managing knowledge? *Harvard Business Review 77*, 106–116.

Hardin, G. (1968). The tragedy of the commons. *Science 162*, 1243–1248.

Hargadon, A. and R. I. Sutton (1997). Technology brokering and innovation in a product development firm. *Administrative Science Quarterly 42*, 716–749.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica 46*, 1251–1271.

Hemphill, C. S. and B. N. Sampat (2012). Weak patents are a weak deterrent: Patent portfolios, the orange book listing standard, and generic entry in pharmaceuticals.

Henderson, R. and I. Cockburn (1994). Measuring competence? exploring firm effects in pharmaceutical research. *Strategic Management Journal 15*, 63–84.

Hervas-Oliver, J.-L. and J. Albors-Garrigos (2009). The role of the firm's internal and relational capabilities in clusters: When distance and embeddedness are not enough to explain innovation. *Journal of Economic Geography 9*, 263–283.

Hoechle, D. (2007). Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal 7*, 281–312.

Ibarra, H., M. Kilduff, and W. Tsai (2005). Zooming in and out: Connecting individuals and collectivities at the frontiers of organizational network research. *Organization Science 16*, 359–371.

Irmer, B. E., P. Bordia, and D. Abusah (2002). Evaluation apprehension and perceived benefits in interpersonal and database knowledge sharing. *Academy of Management Proceedings 2002*, B1–B6.

Iwashyna, T. J., J. D. Christie, J. Moody, J. M. Kahn, and D. A. Asch (2009). The structure of critical care transfer networks. *Medical Care 47*, 787–793.

Iyengar, S. S. and E. Kamenica (2010). Choice proliferation, simplicity seeking, and asset allocation. *Journal of Public Economics 94*, 530–539.

Iyengar, S. S. and M. R. Lepper (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology 79*, 995–1006.

Jaffe, A. B. and M. Trajtenberg (2002). *Patents, citations, and innovations: A window on the knowledge economy*. Cambridge, MA: MIT Press.

Katila, R. and S. Shane (2005). When does lack of resources make new firms innovative? *Academy of Management Journal 48*, 814–829.

Katz, J. S. and B. R. Martin (1997). What is research collaboration? *Research Policy 26*, 1–18.

Kleinbaum, A. M. (2012). Organizational misfits and the origins of brokerage in intrafirm networks. *Administrative Science Quarterly 57*, 407–452.

Kleinbaum, A. M., T. E. Stuart, and M. L. Tushman (2013). Discretion within constraint: Homophily and structure in a formal organization. *Organization Science 24*, 1316–1336.

Kleinbaum, A. M. and M. L. Tushman (2007). Building bridges: The social structure of interdependent innovation. *Strategic Entrepreneurship Journal 1*, 103–122.

Kossinets, G. and D. J. Watts (2006). Empirical analysis of an evolving social network. *Science 311*, 88–90.

Kossinets, G. and D. J. Watts (2009). Origins of homophily in an evolving social network. *American Journal of Sociology 115*, 405–450.

Kotha, R., Y. Zheng, and G. George (2011). Entry into new niches: The effects of firm age and the expansion of technological capabilities on innovative output and impact. *Strategic Management Journal 32*, 1011–1024.

Lahiri, N. (2010). Geographic distribution of R&D activity: How does it affect innovation quality? *Academy of Management Journal 53*, 1194–1209.

Lai, R., A. D'Amour, A. Yu, Y. Sun, V. Torvik, and L. Fleming (2011). Disambiguation and co-authorship networks of the U.S. patent inventor database. Working paper, Harvard Business School, Boston.

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge, MA: Harvard University Press.

Laursen, K., F. Masciarelli, and A. Prencipe (2012). Regions matter: How localized social capital affects innovation and external knowledge acquisition. *Organization Science 23*, 177–193.

Lawrence, P. R. and J. W. Lorsch (1967). *Organization and environment: Managing differentiation and integration*. Boston: Division of Research, Graduate School of Business Administration, Harvard University.

Lazer, D. and A. Friedman (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly 52*, 667–694.

Lee, J. (2010). Heterogeneity, brokerage, and innovative performance: Endogenous formation of collaborative inventor networks. *Organization Science 21*, 804–822.

Lemley, M. A. (2005). Patenting nanotechnology. *Stanford Law Review 58*, 601–630.

Lin, N. (1999). Building a network theory of social capital. *Connections 22*, 28–51.

Lingo, E. L. and S. O'Mahony (2010). Nexus work: Brokerage on creative projects. *Administrative Science Quarterly 55*, 47–81.

Liu, Y.-Y., J.-J. Slotine, and A.-L. Barabási (2011). Controllability of complex networks. *Nature 473*, 167–173.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine 23*, 2937–2960.

Macy Jr., J., L. S. Christie, and R. D. Luce (1953). Coding noise in a task-oriented group. *Journal of Abnormal and Social Psychology 48*, 401–409.

Malmberg, A. and P. Maskell (2006). Localized learning revisited. *Growth and Change 37*, 1–18.

Mamykina, L., B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann (2011). Design lessons from the fastest Q&A site in the West. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2857–2866. New York: ACM.

Mandelbrot, B. (2010). Big Think interview with Benoît Mandelbrot.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

March, J. G. and H. A. Simon (1958). *Organizations* (2nd ed.). Cambridge, MA: Blackwell.

Marquis, C. (2003). The pressure of the past: Network imprinting in intercorporate communities. *Administrative Science Quarterly 48*, 655–689.

Marshall, A. (1890). *Principles of economics* (9th ed.). Bristol and Tokyo: Overstone Press and Kyokuto Shoten.

Mayer, H. (2011). *Entrepreneurship and innovation in second tier regions*. Northampton, MA: Edward Elgar.

Mcafee, A. P. (2006). Enterprise 2.0: The dawn of emergent collaboration. *MIT Sloan Management Review 47*, 21–28.

McEvily, B. and A. Zaheer (1999). Bridging ties: A source of firm heterogeneity in competitive capabilities. *Strategic Management Journal 20*, 1133–1156.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27*, 415–444.

Mehra, A., M. Kilduff, and D. J. Brass (2001). The social networks of high and low self-monitors: Implications for workplace performance. *Administrative Science Quarterly 46*, 121–146.

Mizruchi, M. S. (1996). What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates. *Annual Review of Sociology 22*, 271–298.

Mizruchi, M. S. (2013). *The fracturing of the American corporate elite*. Cambridge, MA: Harvard University Press.

Mizruchi, M. S., L. B. Stearns, and A. Fleischer (2011). Getting a bonus: Social networks, performance, and reward among commercial bankers. *Organization Science 22*, 42–59.

Mody, C. C. M. (2011). *Instrumental community: Probe microscopy and the path to nanotechnology*. Cambridge, MA: MIT Press.

Moody, J., D. McFarland, and S. Bender-deMoll (2005). Dynamic network visualization. *American Journal of Sociology 110*, 1206–1241.

Morgan, S. L. and C. Winship (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge: Cambridge University Press.

National Science Board (2012). *Science and engineering indicators*. Washington, DC: National Science Foundation.

Nerkar, A. and S. Paruchuri (2005). Evolution of R&D capabilities: The role of knowledge networks within a firm. *Management Science 51*, 771–785.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters 89*, 208701.

Newman, M. E. J. and M. Girvan (2004). Finding and evaluating community structure in networks. *Physical Review E 69*, 026113.

Newman, M. E. J., S. H. Strogatz, and D. J. Watts (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E 64*, 026118.

Nie, N. H., D. S. Hillygus, and L. Erbring (2002). Internet use, interpersonal relations, and sociability: A time diary study. In B. Wellman and C. Haythornthwaite (Eds.), *The Internet in everyday life*, pp. 215–243. Malden, MA: Blackwell Publishers.

Obstfeld, D. (2005). Social networks, the tertius iungens orientation, and involvement in innovation. *Administrative Science Quarterly 50*, 100–130.

Opsahl, T., F. Agneessens, and J. Skvoretz (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks 32*, 245–251.

O'Reilly, C. A. and M. L. Tushman (2004). The ambidextrous organization. *Harvard Business Review 82*, 74–83.

O'Reilly III, C. A. and M. L. Tushman (2008). Ambidexterity as a dynamic capability: Resolving the innovator's dilemma. *Research in Organizational Behavior 28*, 185–206.

Orr, J. E. (1996). *Talking about machines: An ethnography of a modern job*. Ithaca, NY: Cornell University Press.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge: Cambridge University Press.

Owen-Smith, J. (2001). Managing laboratory work through skepticism: Processes of evaluation and control. *American Sociological Review 66*, 427–452.

Owen-Smith, J. and W. W. Powell (2004). Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community. *Organization Science 15*, 5–21.

Padgett, J. F. and W. W. Powell (2012). The problem of emergence. In J. F. Padgett and W. W. Powell (Eds.), *The emergence of organizations and markets*, pp. 1–29. Princeton, NJ: Princeton University Press.

Paruchuri, S. (2010). Intraorganizational networks, interorganizational networks, and the impact of central inventors: A longitudinal study of pharmaceutical firms. *Organization Science 21*, 63–80.

Paruchuri, S., A. Nerkar, and D. C. Hambrick (2006). Acquisition integration and productivity losses in the technical core: Disruption of inventors in acquired companies. *Organization Science 17*, 545–562.

Perry-Smith, J. E. and C. E. Shalley (2003). The social side of creativity: A static and dynamic social network perspective. *Academy of Management Review 28*, 89–106.

Phelps, C., R. Heidl, and A. Wadhwa (2012). Knowledge, networks, and knowledge networks: A review and research agenda. *Journal of Management 38*, 1115–1166.

Piore, M. J. and C. F. Sabel (1984). *The second industrial divide: Possibilities for prosperity*. New York: Basic Books.

Podolny, J. M. and J. N. Baron (1997). Resources and relationships: Social networks and mobility in the workplace. *American Sociological Review 62*, 673–693.

Pons, P. and M. Latapy (2005). Computing communities in large networks using random walks. In P. Yolum, T. Güngör, F. Gürgen, and C. Özturan (Eds.), *Lecture Notes in Computer Science*, pp. 284–293. Berlin: Springer.

Pontikes, E. G. (2012). Two sides of the same coin: How ambiguous classification affects multiple audiences evaluations. *Administrative Science Quarterly 57*, 81–118.

Porter, A. L. and I. Rafols (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics 81*, 719–745.

Porter, A. L. and J. Youtie (2009). How interdisciplinary is nanotechnology? *Journal of Nanoparticle Research 11*, 1023–1041.

Porter, M. E. and S. Stern (2001). Innovation: Location matters. *Sloan Management Review 42*, 28–36.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems 14*, 130–137.

Portes, A. and J. Sensenbrenner (1993). Embeddedness and immigration: Notes on the social determinants of economic action. *American Journal of Sociology 98*, 1320–1350.

Pouder, R. and C. H. St. John (1996). Hot spots and blind spots: Geographical clusters of firms and innovation. *Academy of Management Review 21*, 1192–1225.

Powell, W. W., K. W. Koput, and L. Smith-Doerr (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly 41*, 116–145.

Powell, W. W., D. R. White, K. W. Koput, and J. Owen-Smith (2005). Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology 110*, 1132–1205.

Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

Reagans, R. and B. McEvily (2003). Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly 48*, 240–267.

Reagans, R. E. and E. W. Zuckerman (2008). All in the family: Reply to Burt, Podolny, and van de Rijt, Ban, and Sarkar. *Industrial and Corporate Change 17*, 979–999.

Reichert, J. M. (2003). Trends in development and approval times for new therapeutics in the United States. *Nature Reviews Drug Discovery 2*, 695–702.

Rhoten, D. and A. Parker (2004). Risks and rewards of an interdisciplinary research path. *Science 306*, 2046.

Romanelli, E. and C. B. Schoonhoven (2001). The local origins of new firms. In C. B. Schoonhoven and E. Romanelli (Eds.), *The entrepreneurship dynamic: Origins of entrepreneurship and the evolution of industries*, pp. 40–67. Stanford, CA: Stanford University Press.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rosenkopf, L. and P. Almeida (2003). Overcoming local search through alliances and mobility. *Management Science 49*, 751–766.

Rosvall, M. and C. T. Bergstrom (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America 105*, 1118–1123.

Rothaermel, F. T. and M. Thursby (2007). The nanotech versus the biotech revolution: Sources of productivity in incumbent firm research. *Research Policy 36*, 832–849.

Ruef, M., H. E. Aldrich, and N. M. Carter (2003). The structure of founding teams: Homophily, strong ties, and isolation among U.S. entrepreneurs. *American Sociological Review 68*, 195–222.

Salton, G., A. Wong, and C. S. Yang (1975). A vector space model for automatic indexing. *Communications of the ACM 18*, 613–620.

Saxenian, A. (1994). *Regional advantage: Culture and competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press.

Schumpeter, J. A. (1934). *The theory of economic development*. Cambridge, MA: Harvard University Press.

Shannon, C. E. and W. Weaver (1949). *The Mathematical theory of communication*. Urbana: University of Illinois Press.

Sheiness, D. and K. Canady (2006). The importance of getting inventorship right. *Nature Biotechnology 24*, 153–154.

Simmel, G. (1950). *The Sociology of Georg Simmel*. New York: The Free Press.

Singh, J. and A. Agrawal (2011). Recruiting for ideas: How firms exploit the prior inventions of new hires. *Management Science 57*, 129–150.

Singh, J. and L. Fleming (2010). Lone inventors as sources of breakthroughs: Myth or reality? *Management Science 56*, 41–56.

Singh, J., M. T. Hansen, and J. M. Podolny (2010). The world is not small for everyone: Inequity in searching for knowledge in organizations. *Management Science 56*, 1415–1438.

Smith, S. (2005). "Don't put my name on it": Social capital activation and job-finding assistance among the black urban poor. *American Journal of Sociology 111*, 1–57.

Sorensen, J. B. and E. S. Toby (2000). Aging, obsolescence, and organizational innovation. *Administrative Science Quarterly 45*, 81–112.

Sorenson, O. and P. G. Audia (2000). The social structure of entrepreneurial activity: Geographic concentration of footwear production in the United States, 1940-1989. *American Journal of Sociology 106*, 424–461.

Sorenson, O., J. W. Rivkin, and L. Fleming (2006). Complexity, networks and knowledge flow. *Research Policy 35*, 994–1017.

Sorenson, O. and T. E. Stuart (2001). Syndication networks and the spatial distribution of venture capital investments. *American Journal of Sociology 106*, 1546–1588.

Sorenson, O. and T. E. Stuart (2008). Bringing the context back in: Settings and the search for syndicate partners in venture capital investment networks. *Administrative Science Quarterly 53*, 266–294.

Spotila, J. T. (2000). Standards for defining metropolitan and micropolitan statistical areas. *Federal Register 65*, 82228–82238.

Stasser, G. and W. Titus (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology 48*, 1467–1478.

Stasser, G. and W. Titus (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology 53*, 81–93.

Stern, S. (2004). Do scientists pay to be scientists? *Management Science 50*, 835–853.

Stinchcombe, A. L. (1965). Social structure and organizations. In J. G. March (Ed.), *Handbook of organizations*, pp. 142–193. Chicago: Rand McNally.

Stinchcombe, A. L. (1990). University administration of research space and teaching loads: Managers who do not know what their workers are doing. In *Information and organizations*, pp. 312–340. Berkeley, CA: University of California Press.

Stuart, T. E., S. Z. Ozdemir, and W. W. Ding (2007). Vertical alliance networks: The case of university-biotechnology-pharmaceutical alliance chains. *Research Policy 36*, 477–498.

Sytch, M. and A. Tatarynowicz (2014). Exploring the locus of invention: The dynamics of network communities and firms' invention productivity. *Academy of Management Journal 57*, 249–279.

Sytch, M., A. Tatarynowicz, and R. Gulati (2012). Toward a theory of extended contact: The incentives and opportunities for bridging across network communities. *Organization Science 23*, 1658–1681.

Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal 17*, 27–43.

Tallman, S., M. Jenkins, N. Henry, and S. Pinch (2004). Knowledge, clusters, and competitive advantage. *Academy of Management Review 29*, 258–271.

Teece, D. J., G. Pisano, and A. Shuen (1997). Dynamic capabilities and strategic management. *Strategic Management Journal 18*, 509–533.

Terza, J. V. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics 84*, 129–154.

Tortoriello, M. and D. Krackhardt (2010). Activating cross-boundary knowledge: The role of Simmelian ties in the generation of innovations. *Academy of Management Journal 53*, 167–181.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *RAND Journal of Economics 21*, 172–187.

Tsai, W. (2001). Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance. *Academy of Management Journal 44*, 996–1004.

Tushman, M. L. (1977). Special boundary roles in the innovation process. *Administrative Science Quarterly 22*, 587–605.

Tushman, M. L. and D. A. Nadler (1978). Information processing as an integrating concept in organizational design. *Academy of Management Review 3*, 613–624.

Tversky, A. and E. Shafir (1992). Choice under conflict: The dynamics of deferred decision. *Psychological Science 3*, 358–361.

Tzabbar, D. (2009). When does scientist recruitment affect technological repositioning? *Academy of Management Journal 52*, 873–896.

Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly 42*, 35–67.

Uzzi, B. and J. Spiro (2005). Collaboration and creativity: The small world problem. *American Journal of Sociology 111*, 447–504.

Vedres, B. and D. Stark (2010). Structural folds: Generative disruption in overlapping groups. *American Journal of Sociology 115*, 1150–1190.

Vinh, N. X., J. Epps, and J. Bailey (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pp. 1073–1080. New York: ACM.

Vinh, N. X., J. Epps, and J. Bailey (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research 11*, 2837–2854.

Walsh, J. P. and G. R. Ungson (1991). Organizational memory. *Academy of Management Review 16*, 57–91.

Wasserman, S. and K. Faust (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature 393*, 440–442.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly 21*, 1–19.

Whittington, K. B., J. Owen-Smith, and W. W. Powell (2009). Networks, propinquity, and innovation in knowledge-intensive industries. *Administrative Science Quarterly 54*, 90–122.

Wolfe, R. M. (2010). U.S. businesses report 2008 worldwide R&D expense of $330 billion: Findings from new NSF survey. Technical report, National Science Foundation, Arlington, VA.

Wuchty, S., B. F. Jones, and B. Uzzi (2007). The increasing dominance of teams in production of knowledge. *Science 316*, 1036–1039.

Xiao, Z. and A. S. Tsui (2007). When brokers may not work: The cultural contingency of social capital in Chinese high-tech firms. *Administrative Science Quarterly 52*, 1–31.

Zhang, J., M. S. Ackerman, and L. Adamic (2007). Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, New York, pp. 221–230. ACM.

Zucker, L. G., M. R. Darby, and M. B. Brewer (1998). Intellectual human capital and the birth of U.S. biotechnology enterprises. *American Economic Review 88*, 290–306.

Zucker, L. G., M. R. Darby, J. Furner, R. C. Liu, and H. Ma (2007). Minerva unbound: Knowledge stocks, knowledge flows and new knowledge production. *Research Policy 36*, 850–863.