Private R&D Investment in the U.S. Food and Agricultural Sectors and Research Collaboration in the Life Sciences

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

KYUSEON LEE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Adviser: Philip G. Pardey
Co-Adviser: Steve J. Miller

April 2019

Acknowledgements

The journey to the completion of my Ph.D. would not have been possible without the support and help of a number of people. First of all, I am deeply grateful to my advisor, Philip G. Pardey for his guidance, support, and patience throughout my doctoral program. No words can express my deepest gratitude toward him. I would like to especially thank my co-advisor, Steve J. Miller for his valuable insights and suggestions for my work. His guidance and encouragement helped me write this dissertation. Many thanks also go to my dissertation committee members, Terrance M. Hurley and Gregory D. Graff for their insightful and helpful feedback.

I have also greatly benefitted from the support and interactions with my colleagues at the International Science and Technology Practice and Policy (InSTePP) Center: Michelle Hallaway, Robert Andrade, Yuan Chai, Steven Dehmer, Mason Hurley, Ali Joglekar, Connie Chan-Kang, Maxwell Mkondiwa, Senait Senay, Huichun Sun, and Xudong Rao. Thank you all. I also thank late Jason Beddow for his help and guidance at the InSTePP Center. May he rest in peace.

I also thank my study group members in the Department of Applied Economics. I am also very grateful to my friend, Haejin Hwang, for our friendship from the beginning of my study at the University of Minnesota. I also thank my friends at church.

I greatly appreciate the support from the University of Minnesota Libraries, the Office of the Vice President for Research, the Office of Institutional Research, and the Office of Human Resources at the University of Minnesota during the initiation of my research collaboration projects (Ch. 3 and 4).

My deepest gratitude goes to my family. I greatly appreciate the support, and encouragement of my parents, sister, and brother in South Korea. Without their love and understanding, I could not have been here. I am grateful to my parents-in-law for their patience and support. I deeply thank my husband, Doosoo, who always stood by me during the ups and downs of my doctorate study. Thank you for your support and effort to keep me less stressed. I also thank my daughter, Elena, who was very helpful when I was writing this dissertation. I love you. Lastly, I thank God who leads the way.

Dedication

To my parents, my husband, Doosoo, and my daughter, Elena.

Abstract

This dissertation consists of three papers that analyze the two key agents in knowledge production processes, specifically private firms and university researchers.

In the first paper, I present and discuss the details of a new set of firm-level data pertaining to R&D conducted over the period 1950-2014 by food or agriculturally related businesses operating in the United States. Within the food and agricultural sector, I identify the shifting structure of investments in machinery, agriculture and chemicals, and food and beverage processing R&D, emphasizing changes in the portfolio of firms conducting this research among other details. I also econometrically examine the associations between firm-level sales, profit rates, and R&D spending among these firms.

In the second and third papers, I focus on the nature of research collaborations among the life sciences faculty at the University of Minnesota. In the second paper, I analyze the patterns of research collaboration, by using a unique, purpose-built data set spanning 3,305 scientists from three large academic units within the University of Minnesota (UMN) spanning the period 1999-2014. Coauthorship data are used to examine the patterns of internal and external collaborations and how these patterns have changed over time. In the third paper, I use the same data as the second paper, but narrow the focus to papers that are published by only UMN affiliated authors. I examine the three particular types of information signals that researchers use in selecting research partners: perceived research productivity, knowledge complementarity, and professional familiarity. The findings have important implications not only for collaboration among academic researchers but also to collaboration within and across firms, especially when that collaboration targets the acquisition or leveraging of knowledge.

Table of Contents

List of Tables

List of Figures

1 Introduction

A hallmark of the U.S. food, health and agricultural sectors is their notable investments into research and development (R&D) and related innovative activities. These investments include life sciences R&D conducted both by public and private entities.

Firms engage in innovative effort for a host of reasons, including sustaining or increasing market share and enhancing profits. University researchers are also motivated by economic self-interest, including higher salaries and academic standing, while some are also driven by the search for new knowledge or enhancing societal well-being. Whether it be the pursuit of profits for shareholders or new knowledge for its own sake, both public and private R&D involves a commitment of substantial resources, making the study of innovative activity of increasing interest to economists.

However, as important as innovation is now seen as a key determinant of economic growth and improved well-being, our economic understanding of these phenomenon is still quite limited. It was only several decades ago when Nathan Rosenberg (1982) observed that "Economists have long treated technological phenomena as events transpiring inside a black box." In this thesis, I cast an economic eye over several aspects of the life sciences black box.

Nineteenth century U.S. agriculture was heavily influenced by the efforts of private, often individual, innovators such as Eli Whitney, who patented the cotton gin, Cyrus McCormick, whose mechanical reaper "made bread cheap," Luther Burbank who developed scores of new and improved crop varieties, and John Deer whose steel-tipped moldboard plows helped tame the U.S. prairies (Alston and Pardey 2006). As the twentieth century unfolded, R&D investments affecting the U.S. food and agricultural sectors became increasingly institutionalized, with an increasing and important role for R&D carried out by public agencies such as the U.S. Department of Agriculture (USDA) and the land grant universities. As the century drew to a close the private role in U.S. food and agricultural R&D began to rebound, and as I will show below, now significantly eclipses investments in public R&D. Notwithstanding the increasing dominance of privately funded and performed food and agricultural R&D, little is known about the firms engaged

in this research. Who does the R&D, how has the portfolio of innovating firms changed over time and what are their attributes (in terms of size, longevity, and so on), and what accounts for changes in the amount invested in R&D from year to year and over the longer term? Drawing on an entirely new firm-level database constructed for this purpose, these are the unknowns that I quantitatively investigate in Chapter 2.

In comparison, the economic dimensions of life sciences (i.e., food, health and agricultural) R&D conducted by the U.S. public sector has received more attention, especially research undertaken by academic institutions. However, as the scale of these academic agencies has grown and the nature of the life sciences has rapidly evolved over recent decades, there is still lots to learn about the nature of these innovation processes. In particular, the cross-disciplinary and inter-institutional nature of collaboration in academic R&D has gained more attention of late (Porter and Rafols 2009; Jones et al. 2008). Here I focus on the nature of research collaboration in life sciences R&D.

To do so, I again use a purpose built database pertaining to life sciences R&D conducted at a large public university, specifically the University of Minnesota (UMN). The data base spans the period 1999 to 2014 (when life sciences R&D was in a period of considerable flux) and includes three main colleges at UMN, College of Biological Sciences, College of Food, Agricultural, and Natural Resource Sciences, and the Academic Health Center. In chapter 3, I examine the patterns of (internal and external) research collaborations by UMN researchers in the life sciences and whether it changes across colleges and over time. And by using the same dataset as chapter 3, chapter 4 investigates team formation and the researchers' characteristics that affect the propensity to collaborate by using the system generalized methods of moments (GMM).

To summarize, this dissertation provides evidence on the increasing importance of R&D investment among private firms in the U.S. food and agricultural sectors and research collaboration among the life scientists at UMN. The findings and arguments in this dissertation should be of interest not only to academic researchers, but also to governments, university administrators, private firms and other institutions and stakeholders, seeking deeper insights into innovation processes within the life sciences.

2 An Examination of Private (Sectoral and Firm-Level) Investments in U.S. Food and Agricultural R&D, 1950-2014

2.1 Introduction

Firms invest in research and development (R&D) for a host of reasons, including increasing output, reducing costs, encouraging productivity and profit growth, preserving or expanding market share in existing markets, or as an entry strategy into new markets (Griliches 1988; Geroski et al. 1997; Van Reenen 1997; Gilbert and Newbery 1982; Reinganum 1983). Although the decision to engage in R&D can be strategic and diverse, once firms commit to undertake in R&D, the decision on how much to invest in R&D is inevitably influenced by each firm's past or prospective financial capacity, as reflected in its sales revenue or profit rate (often measured as the cash flow to capital ratio). In a number of manufacturing sectors, the relationship between a firm's financial position and its spending on R&D has been studied in some depth (Himmelberg and Petersen 1994; Hall and Lerner 2010). However, little research has examined this topic for the food and agricultural sectors, most likely a reflection of the substantive public R&D presence in these sectors and a paucity of suitable firm-level data.

The U.S. food and agricultural sectors have changed markedly over the past half century. Initially, most R&D investment in these two sectors was invested in agricultural input and food products, mainly from the public sector (Alston et al. 2011). However, advances in the biological sciences and the broadening of intellectual property rights (IPRs) stimulated private technological interests in the development of new crops and plant varieties (Caswell and Day-Rubenstein 2006). For the food sectors, the increasing consumption of pre-prepared food and food-away-from-home also prompted additional innovative activity (Alston and Pardey 2014). These scientific, market and regulatory changes have given rise to new incentives for private firms to invest in and benefit from R&D activity. In recent decades we have also witnessed an increase in merger and acquisition (M&A) behavior among food and agricultural companies, some seemingly

motivated by decisions to acquire technology (and associated IP rights) rather than invest in in-house R&D (King and Schimmelpfennig 2005).

Despite much commentary on the changing landscape for private food and agricultural R&D, little empirically-supported research has examined the innovative activities of firms in these sectors. The research reported in this chapter is designed to redress this lack of information. Within the food and agricultural sector, we identify the shifting structure of investments in machinery, agriculture and chemicals, and food and beverage processing R&D, emphasizing changes in the portfolio of firms conducting this research, among other details. Drawing on prior work by Mulkay et al. (2001) and Bond et al. (2005) that assessed firm-level relationships between R&D investments and the financial performance of manufacturing firms, we econometrically examine these same relationships for firms investing in R&D in the food and agricultural sectors.

We find that firms invest more in R&D when they anticipate higher future sales. Not only do firms invest more in R&D when past sales increase, they also tend to increase their investments in capital when anticipating future sales, thus lowering their cash flow to capital ratio, which, eventually, increase the R&D investment. The food and agricultural sectors also differ in terms of the magnitude of these relationships. The sizes of the coefficients are generally larger for food firms, with the exception of lagged R&D growth, suggesting that food firms are more sensitive to changes in sales when investing in R&D. These findings are robust to the elimination of outliers in R&D growth that might otherwise skew results.

This study contributes to the existing literature of firm-level innovation and R&D investment in several ways. First, while a large number of empirical studies have examined the effects of sales and profit rates on R&D investment in the manufacturing sector, there is little research that has studied this relationship in the food and agricultural sectors. This research fills that gap by providing the first, to our knowledge, empirical analysis of the drivers of private, firm-level R&D investment in these sectors. Second, our findings shed light on the increasing R&D investment in crop sciences and food technologies in the United States, and the spillover effects that investment could realize

in other national economies. In addition, the U.S-based findings of this paper can also inform our understanding of the more recent and significant increases in private spending on food and agricultural R&D among the agriculturally important middle-income countries (Pardey et al. 2016a).

2.2 Empirical Evidence and Conceptual Framework

2.2.1 Prior Evidence

R&D investments have several unique characteristics that differ in nature from the other fixed investments and costs incurred by firms. First, R&D investments involve considerable uncertainty. Even if firms ultimately benefit from their R&D investment, there is usually a long lag between the initiation of an R&D endeavor and the commercialization of the results emanating from research (Hall and Lerner 2010). Second, most R&D spending includes wages and salaries of highly educated personnel and, therefore, it is often a comparatively costly undertaking. Moreover, much of the R&D capacity of a firm is embedded in human capital that over time acquires specialized, and sometimes firm-specific, attributes. This specialization makes it more difficult for firms to tap temporary talent while also changing the hire-fire calculus, both of which have potential implications for the volatility of R&D spending relative to the other costs incurred by firms. Firms may also be reluctant to fire research personnel for fear that their embodied knowledge may be transmitted to competitors. Thus, R&D investment has high adjustment costs and past evidence (mainly pertaining to manufacturing firms) suggests firms tend to smooth their R&D spending over time (Hall 1992; Himmelberg and Petersen 1994; Hall and Lerner 2010).

Numerous empirical studies have examined the relationship between a firm's financial performance and investments in R&D, yielding mixed results. Some studies suggest there is a positive relationship between a firm's R&D spending and its cash flow using firm-level data for the U.S. and elsewhere (Hall 1992; Himmelberg and Petersen 1994; Mulkay et al. 2001; Brown et al. 2009). However, using data for British and German

firms, Bond et al. (2005) found that the ratio of cash flow to capital stock did not have a significant effect on the level of investment in R&D, although the ratio does influence whether or not a firm engages in R&D. Based on German manufacturing data, Harhoff (1998) also found that firm size had a strong impact on the relationship between cash flow and R&D investment, with smaller firms more sensitive to financial constraints. Using U.S. manufacturing data, Brown and Petersen (2009) found a significant relationship between cash flow and R&D spending for young firms, but not for mature firms. These prior, but not always consistent, results suggest that the relationship between R&D investment and the financial performance (specifically cash flow and sales volume) of a firm differ with respect to firm size, age, country, and financial and labor market rules.

However, the prior findings rely heavily on firm-level data from manufacturing and high-tech industries. Although the manufacturing industry as defined by the standard industrial classification (SIC) code includes some food processing, tobacco, and chemical companies, food and agricultural R&D companies are sparsely represented in these data sets. To the best of our knowledge none of these types of studies have been conducted for firms that are solely or significantly engaged in food and agricultural R&D. To the extent these types of studies have been conducted in relation to food and agriculture, they have relied exclusively on R&D investments made by the public sector. For example, Pardey and Craig (1989) analyzed the causal relationships between public sector agricultural research expenditures and output using state-level data from the USDA and state agricultural experiment stations (SAES).

More expansive empirical analyses of private agricultural R&D investments in the United States are a relatively recent phenomenon (see, for example, Fuglie et al. 2011 and 2012; Fuglie and Tool 2014; Fuglie 2016). However, despite the new, largely USDA-sourced, data underpinning these studies, no in-depth, firm-level analysis exists for both the food and agricultural sectors over a longer time span. Specifically, Fuglie et al. (2011) constructed global time series data for the agricultural input sectors for the period 1994-2010, but no data on sales and profit rates for these sectors are available.[1] These prior

[1] The data reported by Fuglie at al. (2011), the USDA, ERS-led database of private sector R&D spending (in the crops, animals and farm machinery sectors), include 324 firms globally, of which only 124 were

studies focused on comparing and contrasting private and public R&D spending trends, and did not attempt to empirically examine the underlying relationship between the financial constraints of firms and their R&D investment choices (Fuglie and Toole 2014; Pray and Fuglie 2015).

Notwithstanding the increased role of private sector R&D in the U.S. food and agricultural sectors, there are few firm-level data sets that encompass R&D activity for this sector, reflecting in part the enormous challenges involved in constructing such a set of data. For example, numerous firms include business segments that span multiple (SIC) sectors, making it difficult to distinguish agriculture- and food-related R&D investment from research that is directed to other (sometime unrelated) sectors.[2] To address this challenge and facilitate the research reported here, we constructed a new, replicable and comprehensive firm-level data set for food- or agriculturally-related firms operating within the U.S. that stretches back over 60 years. (See the data section below for additional details.)

2.2.2 Conceptual Framework

To assess the relationship between the financial performance of firms in the U.S. food and agricultural sectors and their propensity to invest in R&D, we deployed a variant of the error correction model (ECM) used by Mulkay et al. (2001) and Bond et al. (2005) to study a similar phenomenon in the U.S. manufacturing sector.

incorporated in the USA-Canada. Moreover, the 324 global total includes 182 companies that were operating in 2014, and 142 "legacy" companies that operated some time during 1990-2013. For comparison, version 4.0 of the InSTePP private sector food and agricultural R&D database includes 465 firms operating in the U.S. during the period 1950-2014 (with 244 of those firms operating in areas other than food and beverage processing); 322 of these firms (175 in areas other than food processing) operated in the period beginning in 1990, and 131 of those firms were operating in 2014. The implication of these comparisons is that the InSTePP database includes substantially more firms that the USDA database, at least regarding private (food and) agricultural R&D in the U.S. (Chai et al. 2019).

[2] For example, Pfizer Inc., a US-based global pharmaceutical company founded in 1849, also had long-standing agricultural division that conducted in-house R&D, which was fully spun out as Zoetis in 2013. Clearly these agriculturally related R&D data belong in our series, but not the considerable other investment Pfizer makes in human health related R&D.

For conceptual motivation, consider the simple, neo-classical, long-run model of a profit maximizing firm that produces a single good with production that involves the use of labor (L) and two kinds of capital: physical capital (K), and knowledge capital (G). The constant elasticity of substitution (CES) production function is

$$f(L_t, K_t) = A_t[\alpha L_t^{\varphi} + \beta K_t^{\varphi} + \gamma G_t^{\varphi}]^{\frac{1}{\varphi}v} \quad (1)$$

where φ and v are, respectively, the elasticities of substitution and scale, and $\alpha, \beta, \gamma \in (0,1)$ are the factor share parameters. From this, we can derive the long-run optimal level of knowledge capital as a log-linear function of output and the user cost of knowledge capital. Let g_{it} denotes the log of the desired amount of knowledge capital for firm i in period t, y_{it} denotes the log of sales, and c_{it} denotes the log of the user cost of knowledge capital. This gives the long-run optimal demand for knowledge capital as

$$g_{it} = \alpha_t + \beta y_{it} - \sigma c_{it} \quad (2)$$

where α_t is the time varying technical change parameter, and β is the (long-run) elasticity of sales to capital.

Assuming there are adjustment processes involved in attaining the long-run optimal amount of knowledge capital, we deploy a dynamic relationship between g_{it} and y_{it} with a lag length of two. In doing so, we control for the use cost of capital by including firm-specific and time-specific effects, α_t and μ_i, respectively.[3] Equation (2) can be re-written as the following dynamic equation:

$$g_{it} = \beta_1 g_{it-1} + \beta_2 g_{it-2} + \gamma_0 y_{it} + \gamma_1 y_{it-1} + \gamma_2 y_{it-2} + \alpha_t + \mu_i + \varepsilon_{it} \quad (3)$$

We then convert this equation to an error correction model (ECM) as in Engel and Granger (1987) because the previous model only considers the short-run relationship between knowledge capital and sales. In contrast, the error correction model specifically

[3] The relevant user cost of capital is often difficult to measure because both the output and R&D price are not typically available at the firm level. In addition, depreciation rates may also vary among firms.

distinguishes between a long-run and short-run relationship. We re-write equation (3) as an error correction model as follows:

$$\Delta g_{it} = (\beta_1 - 1)\Delta g_{it-1} + \gamma_0 \Delta y_{it} + (\gamma_0 + \gamma_1)\Delta y_{it-1} \tag{4}$$

$$+\eta(g_{it-2} - y_{it-2}) + \theta y_{it-2} + \alpha_t + \alpha_i + \varepsilon_{it}$$

$$\text{where } \eta = \beta_2 + \beta_1 - 1 \text{ and } \theta = \gamma_2 + \gamma_1 + \gamma_0 + \beta_2 + \beta_1 - 1$$

In this equation, the first three terms reflect the short-run dynamics, and the following two variables ($\eta(g_{it-2} - y_{it-2})$ and y_{it-2}) capture the long-run dynamics. When $\eta \neq 0$, then $(g_{it-2} - y_{it-2})$ is the error correction term, and represents the adjustment rate given by the gap between knowledge capital and sales. Specifically, when η is negative and significant, we can expect to see higher R&D investment when knowledge capital is less than the optimal level. Thus, the long-run dynamics of knowledge capital are driven by both changes in output and the gap between knowledge capital and output in the long-run equilibrium. Lastly, the coefficient θ allows us to test the hypothesis that the long-run elasticity of capital with respect to output is unity (i. e., $\gamma_2 + \gamma_1 + \gamma_0 + \beta_2 + \beta_1 - 1 = 0$)

Bond et al. (2005) and Mulkay et al. (2001) also included a measure of profit rate because sales growth alone does not capture the impact of profitability on R&D investment. More formally, the profit rate is defined as the cash-flow to capital ratio, calculated as follows: $\Pi_{it} = C_{it}/K_{i,t-1}$. We calculate the stock of knowledge capital using the perpetual inventory method ($K_{it} = (1 - \delta)K_{it-1} + i_{it}$), which is deflated by an investment goods deflator.[4] The cash-flow to capital stock ratio, however, cannot be interpreted as direct evidence of the financial constraints faced by each firm. Instead, the current amount of capital stock depends on expectations of future output as well as on current output and the cost of capital (Nickell 1978; Kaplan and Zingales 1997). Specifically, the current amount of capital stock can be calculated as

$$K_{it} = \alpha K_{it-1} + \beta y_{it} + \gamma E_t[y_{it+1}]$$

[4] See the Appendix for details on constructing cash flow to capital stock ratio.

where $E_t[y_{it+1}]$ is the expected value of sales at t+1 given information in period t (Bond et al. 2005). Thus, this cash-flow to capital stock variable is considered more like a measure of future profitability or future demand.

Finally, to measure the growth rate of knowledge capital, we follow Bond et al.'s (2005) approach, which substitutes the growth rate of knowledge capital with the growth rate of R&D investment, r_{it}.[5] This is because relatively few firms reported R&D expenditures over a long span of time, and thus a constructed estimate of the stock of knowledge capital formed by a weighted sum of historical series of R&D expenditures is likely to introduce measurement inconsistencies among the firms.

Thus, we estimate the following error correction model:

$$\Delta r_{it} = (\beta_1 - 1)\Delta g_{it-1} + \gamma_0 \Delta y_{it} + (\gamma_0 + \gamma_1)\Delta y_{it-1} + \eta(g_{it-2} - y_{it-2}) + \theta y_{it-2} + \lambda_0 \Pi_{it} + \lambda_1 \Pi_{it-1} + \alpha_t + \alpha_i + \varepsilon_{it} \quad (5)$$

that can be re-written as

$$\Delta r_{it} = \mu_1 \Delta r_{it-1} + \rho_0 \Delta y_{it} + \rho_1 \Delta y_{it-1} + \eta(r_{it-2} - y_{it-2}) + \theta y_{it-2} + \lambda_0 \Pi_{it} + \lambda_1 \Pi_{it-1} + \alpha_t + \alpha_i + \varepsilon_{it} \quad (6)$$

where $\mu_1 = \beta_1 - 1$, $\rho_0 = \gamma_0$, $\rho_1 = \gamma_0 + \gamma_1$, $\eta = \beta_2 + \beta_1 - 1$ and $\theta = \gamma_2 + \gamma_1 + \gamma_0 + \beta_2 + \beta_1 - 1$.

[5] Knowledge stock for firm i, G_{it}, can be defined as $G_{it} = (1 - \delta_i)G_{i,t-1} + R_{it}$ when δ_i is a depreciation rate and R_t is the R&D expenditure at time t. In steady state with growth rate v_i, a firm's knowledge stock can be written as $G_t = (1 + v_i)G_{it-1}$,

$$R_{it} = (\delta_i + v_i)G_{it-1}$$
$$= \left(\frac{\delta_i + v_i}{1 + v_i}\right) G_{it}$$
$$\text{and } r_{it} = \ln\left(\frac{\delta_i + v_i}{1 + v_i}\right) + g_{it},$$

where r_{it} is the log of R&D expenditure (Bond et al. 2005; Bean 1981).

2.3 Data

2.3.1 Data Description

The data used for this study are a newly updated and revised series developed by the International Science and Technology Practice and Policy (InSTePP) at the University of Minnesota.[6] This series blends data on publicly-traded firms in the United States obtained from Standard & Poor's Compustat North American database with data drawn from published annual reports or financial statements of larger firms.

The data consist of firm-level observations of annual sales, R&D spending, and the cash flow to capital stock ratio (profit rate). R&D expenditures represent all the costs incurred during a given year that relate to the development of new products or services, while the sales variable measures the value of gross sales in a given year excluding any credit given to customers. Cash flow is defined as the net funds available for investment, which is the sum of income before extraordinary items and depreciation and amortization. We do not add back research and development expenses in the cash flow as is common in the finance literature (Hall 1992; Brown et al. 2009), because doing so would result, by construction, in simultaneity and correlation between cash flow and R&D investment (Bond et al. 2005; Mulkay et al. 2001). All economic data are deflated to 2012 U.S. dollars using the implicit GDP deflator obtained from the Bureau of Economic Analysis (2018).

Agricultural firms include those involved in manufacturing farm machinery, seed production, and agricultural chemicals. Food-related firms include those engaged in processing and producing food, beverages, and tobacco products. These firms can also be grouped into three broad sub-sectors, specifically agricultural production and chemical, agricultural machinery, and food. Details on the SIC codes used to identify the relevant firms and to group them into three sub-sectors are listed in Table 2-1.

[Table 2-1: Standard Industrial Classification (SIC) Codes Used]

[6] For details on data collection and data processing, see Chapter 6 of Pardey et al. (2016b). This prior documentation is for version 3.5 of the series, and the updated documentation for the version 4.0 series used in this study is in progress.

We began with a total of 2,179 firms classified as food- and agriculturally-related firms according to their assigned SIC codes (Table 2-2). From this initial sample, we identified 1,786 firms that operated within the U.S. To be designated a firm operating within the U.S. (irrespective of whether the firm was domestically- or foreign-owned) companies were required to report U.S. sales as determined by the "geographic segment" field in the underlying source data.

[Table 2-2: Number of Firms in the Sample, by Sub-Sector]

To approximate the jurisdictional extent of R&D spending—specifically in this case, U.S. versus rest-of-the-world (ROW) performed R&D—we make use of historical sales by geographical segment data. For this process, we designated firms as being "large" if they reported median inflation-adjusted R&D spending for the 1950-2014 period of more than $100 million per year. For these large firms, we formed our sales shares estimates by using historical geographic segment sales data also reported in 10-k filings (or 20-f filings for foreign based firms). These data were obtained from three sources, specifically the Security and Exchange Commission's EDGAR database, the Orbis database published by Bureau van Dijk, and the S&P Captial IQ series, a subsidiary of McGraw Hill Financial. In many instances these sales data were only available back to the early 1990s, and so we backcast R&D attribution shares based on the earliest available reported sales data. We also assumed that prior to 1970 all R&D expenditures were attributable to the country in which the firm was incorporated. For the period 1970-1989, we assumed the first 75 percent of R&D expenditures were attributable to the country of incorporation, and the remaining 25 percent was attributed proportionally according to geographic segment sales. For 1990 and thereafter, we assumed that the first 50 percent of R&D expenditures were attributable to the country of incorporation, and the remaining 50 percent was attributed proportionally according to geographic segment sales. Lastly, for all firms with 1950-2014 median inflation-adjusted annual R&D expenditures less than $100 million, we assumed that 100 percent of each firms R&D activities were performed in the firm's country of incorporation.

For these 1,786 firms, we applied a similar procedure to estimate the share of a firm's R&D spending that pertained to food or agriculture as we did to approximate the jurisdictional extent of R&D investment. We do so because firms with SIC codes that belong to either the food or agricultural sector may not be exclusively engaged in either food- or agriculturally-related activities. For larger firms, it is more likely that they are partially engaged in either agricultural or food related businesses. Specifically, we assumed that the share of food or agricultural related R&D spending was congruent with the share of total sales derived from food and agriculture in the business segment sales data that are obtained along with the geographical segment data. For firms with 1950-2014 median inflation-adjusted annual R&D expenditures less than $100 million, we assumed that 100 percent of each firms R&D activities were performed within the food or agricultural sectors but with a few exceptions. We excluded most chemical firms whose first two digits of SIC codes are 28 (Chemical and allied products) because we could not confirm if their R&D activities are truly food- or agricultural- related. These industries are marked as "large" in Table 2-2.

We also imputed some missing R&D data prior to 1970. In some instances, we observed that firms reported doing R&D in 1970 or 1971 (and in following years), but not in years prior to 1970, even though they reported sales and other financial information for these earlier years. In 1972, amendments to regulation S-X requiring broader 10-K disclosures for R&D spending came into force (Miguel 1977; Securities and Exchange Commission 1972). We opted to assume that those firms disclosing R&D spending in 1970 and 1971 (and consistently in the following years) were also conducting (non-disclosed) R&D in these earlier years. We formed an estimate of the R&D spending in the years 1950 to 1969 by multiplying the research intensity ratio for 1970 (or 1971) by the corresponding annual sales values if they were available. A total of 1,280 year-firm observations (96 firms) were estimated, which is approximately 13 percent of the total observations in the dataset.

The resulting dataset is an unbalanced panel of 465 firms that conducted at least some domestic food- or agricultural R&D in any year during the period 1950-2014 (Table 2-3). Among the 465 firms, only 28 firms reported observations for all 65 years of the

panel, while 158 firms (more than 30 percent of the total firms) had entries in the series for a decade or less. The 28 firms with complete time-series coverage tended to be larger than the 158 firms with more limited data: notably $326 million per year of 1950-2014 median inflation-adjusted R&D expenditures compared with just $10.7 million per year. Fewer than 100 firms reported in the dataset have available data from the 1950s, with the number of firms in each year continuing to grow until 1975 (Figure 2-1). Over the following two decades the number of firms declined to a local low of 164 firms in 1984, rebounded to a local peak of 210 firms in 1997, and thereafter declined to just 131 firms in 2014. The steady decline in the number of firms conducting food and agricultural R&D over recent decades is witness to the increased merger and acquisition activity that has taken place during this time.

[Table 2-3: Balance of Panel]

[Figure 2-1: Number of Firms per Year]

2.3.2 The Pattern of Private R&D

We plot the time series of U.S. private food and agricultural R&D for the period 1950-2014 in Figure 2-2, and compare it with the U.S. public food and agricultural R&D series from InSTePP (Pardey et al. 2016a). Since the 1950s, the rate of growth in inflation-adjusted spending on privately performed food and agricultural R&D in the United States has consistently outpaced the corresponding public-sector growth, such that by the mid-1970s the private sector began outspending the public sector. The private-public R&D gap continued to grow, and by 2014, for every dollar of public spending there were 2.4 dollars of private investment.

[Figure 2-2: U.S. Private and Public Food and Agricultural R&D Expenditures]

The sub-sectoral composition of U.S. private food and agricultural R&D expenditures since 1950 is revealed in Figure 2-3. The preponderance of the private R&D is conducted by firms grouped within the agricultural production and chemical plus food sub-sectors. Marked fluctuations in the amount of R&D conducted by either of these sub-

sectors arises for multiple reasons, including mergers and acquisition activities, firm entry and exit, and notable shifts in the pace of investment by some of the particularly large firms for certain periods of time. Notably, firms such as DuPont and Pfizer rapidly ramped up their R&D spending in the 1990s such that agricultural production and chemical companies' R&D investment suddenly increased beginning in the mid-1990s. The total amount of R&D from the agricultural production and chemical companies dropped substantially in 1999, a reflection of a marked contraction in Monsanto's R&D spending in that year along with some R&D downsizing associated with DuPont's purchase of Pioneer Hi-Bred that same year. R&D investment in agricultural machinery, a sub-sector with the fewest firms and the lowest total R&D investment, has markedly increased since the mid-2000s, closing the gap with the agricultural production and chemical sector and food sector.

[Figure 2-3: U.S. Private Food and Agricultural R&D Expenditures, by Sub-sectors]

While both the agricultural production and chemical sector and the food sector spent more than $3.5 billion in 2014 for food and agricultural-related R&D, the share of R&D expenditures in sales were substantially different (Figure 2-4). The figure suggests that there is a substantial difference in R&D intensity across sectors: relatively high R&D intensity in agricultural machinery firms and very low R&D intensity in the food sector.[7] In contrast to Figure 2-3 where the sectoral sum of R&D investment in the food sector was similar to the agricultural production and chemical sector in most of years, this low R&D intensity implies that although food companies have invested heavily in R&D, the amount invested is relatively small compared with their sales revenues. On the other hand, the intensity of research conducted by agricultural machinery firms has fluctuated over time. This likely reflects the fact that there are fewer firms in that sector, and, thus, the sub-sectors R&D intensity is more likely to be influenced by the R&D investment decisions of individual firms and the somewhat erratic nature of firm entry and exit.

[7] R&D intensity (R&D expenditures divided by sales) represents the money firms invest in R&D in response to their sales revenue.

[Figure 2-4: Share of U.S. Private Food and Agricultural R&D Expenditures in Sales, by Sub-sectors]

While these sectoral spending totals provide useful overviews of the changing pattern of private food and agricultural R&D spending in the US, they mask insights that can be gleaned from the underlying variation in the firm-level data. Figure 2-5 shows the distribution of log R&D investment at a firm level for each sector for every decade. We first notice from the agricultural production and chemical firms in Panel A that while the median R&D investment has not changed substantially over time, the distribution of firms in each is increasingly dispersed. This indicates that the discrepancy between firms located at the top and bottom percentiles has widened over time, a phenomenon that also arises in the agricultural machinery and food sectors. However, unlike the agricultural production and chemical sector, the median R&D investment substantially dropped from 1950 to 1990 in the agricultural machinery sector. This was because the number of firms in the machinery sector substantially increased from 1950 to the early 1980s, from two firms in 1950 to fourteen firms in 1984.[8] The agricultural machinery sector has a relatively small number of firms compared with the other two sectors, making summary statistics sensitive to firms' entry and exit. The median R&D investment in the food sector gradually increased from 1950 to 2014. The distribution of food firms is right-skewed, indicating that many of the firms in this sub-sector invest very little in R&D (e.g., in 2014, 10 percent of the 56 firms in this sector spent less than $10 million on R&D, while the top five firms spent more than $200 million on R&D).

[Figure 2-5: Firm-level R&D Expenditures by Sub-sectors, Natural Logarithms]

Summary statistics in Table 2-4 reconfirm that the data are right skewed and widely dispersed. For example, the agricultural machinery sectors' median R&D expenditures are the lowest among the three sectors but the highest at the 90th percentile, suggesting that a few firms in the top percentile invest substantially in the agricultural machinery sector. We also find that other sectoral differences are notable, especially in sales. As a result, the

[8] Those two firms in 1950 are Deere & Co., and Case Corporation whose 1950-2014 median, inflation-adjusted R&D spending exceeded $100 million.

food sector reports the lowest median R&D intensity due to high sales and relatively low R&D investment, while R&D intensity is highest among agricultural machinery firms. Lastly, the profit rate variable, measured as cash flow divided by capital stock, also varies considerably among firms in each sector.[9]

[Table 2-4: Summary Statistics for U.S. Food and Agricultural R&D Investing Firms]

2.4 Estimation

Using ordinary least squares (OLS) to estimate equation (6) introduces the prospects of endogeneity, given that Δr_{it} is correlated with α_i. To deal with this kind of endogeneity, Anderson and Hsiao (1982) suggested first differencing the equation and using a two-stage least squares (2SLS) estimator with the lagged dependent variable being used as the instrument. Arellano and Bond (1991) further developed this idea and opted to use a generalized method-of-moments (GMM) estimator with first differences. The Arellano and Bond estimator first takes the first difference of all the variables and eliminates the firm-specific effects and the time-invariant variables. Then it uses all the past information of Y_{ijt} (at least two periods earlier) as instruments in a GMM procedure. The moment condition for this estimator is:

$$\mathrm{E}\left(r_{it-\rho}\Delta\varepsilon_{it}\right) = 0 \; for \; all \; \rho = 2, \dots, t-1 \text{ and } \mathrm{t} = 3, \dots, T \tag{7}$$

However, by doing so, we may experience large finite sample bias because of a weak instrument if the lagged values are weakly correlated with the first-differences.

To estimate this dynamic regression model using our unbalanced panel where many firms have only a small number of time periods for R&D investment, we use the system GMM estimator developed by Arellano and Bover (1995) and Blundell and Bond (1998). This estimator combines equation in levels and in first differences and allows us to control for unobserved firm-specific effects and to address endogeneity concerns. For the equation in levels, lagged first differences are used as instruments and

[9] For the differences between R&D firms and non-R&D firms, please see the Appendix.

for the equation in first differences, lagged levels are used as instruments. The moment condition for this estimator is:

(8) $$\mathrm{E}(\Delta r_{it-1}\varepsilon_{it}) = 0 \; for \; all \; t = 3, \ldots, T$$

Finally, to test the orthogonality between the instruments and the error term, we use the Hansen test of over-identifying restrictions to determine any correlation between the instruments and the errors.

2.5 The Drivers of R&D Investments

For the food and agricultural sectors, we found that firms tend to invest more in R&D when expectations of future sales are high (Table 2-5). First, firms invest more in R&D when past sales increase. Both contemporaneous and lagged output growth terms are positively associated with contemporaneous R&D growth, while lagged R&D growth rate terms are negatively correlated with the contemporaneous R&D growth. However, the lagged cash flow to capital stock ratio is positive only for agricultural firms. These results suggest that agricultural firms are likely to invest more in R&D when expectations of future sales are high. As stated earlier, the capital stock reflects expected future sales, which means it can be seen as a proxy for expectations that a firm holds regarding the future trajectory of sales. As a result, firms anticipating an increase in sales would invest more in capital, which lowers the cash flow to capital ratio, which, eventually, increases the amount of R&D investment.

[Table 2-5: Estimation Results]

The food and agricultural sectors also differ in terms of the magnitude of these relationships. The sizes of the estimated coefficients in Table 2-5 are generally larger for food firms, with the exception of lagged R&D growth. These differences suggest that food firms are more sensitive to changes in sales when investing in R&D than are firms in the other two sub-sectors. The relatively small R&D intensities of food firms may contribute to this stronger relationship as well, because food firms would require more sales revenue than agricultural firms to invest the same amount in R&D.

Error correction terms are only significant for food firms in column (4), indicating that food firms begin adjusting toward the long-run equilibrium level in the current period, with full adjustment taking a little more than five years. Agricultural firms, on the other hand, do not change their R&D investment with respect to deviations from long-run equilibrium. In light of these short- and long-run results, it appears that agricultural R&D investment is largely a longer-term and pre-committed undertaking, somewhat (but by no means entirely) immune to fluctuations in contemporaneous balance sheets.

Lastly, our results do not reject the null hypothesis of the Hansen test of overidentifying restrictions except for column (1). We also find no second-order serial correlation in the first-differenced residuals.

2.6 Robustness Checks

In this section, we examine whether our findings remain consistent even after controlling for differences in firm size and the variance of R&D growth per firm. As shown earlier, the amount of R&D investment among firms in both the food and agricultural sectors varies considerably. Therefore, the dependent variable, R&D growth, fluctuates substantially for small firms who invest little in R&D, whereas it remains relatively stable for the larger firms (Figure 2-6). To account for these differences, we re-estimate our model in two sub-samples: the first excludes outliers with high variance of R&D growth, while the second includes only firms with median log R&D growth above one.

[Figure 2-6: Median R&D Expenditures (Natural Logarithms) over Variance of R&D Growth]

Estimates using these restricted samples yield qualitatively similar conclusions to those from our full dataset (Table 2-6). We still observe a positive and significant association of contemporaneous output growth for both food and agricultural firms. This suggests that firms in both sectors adjust their current level of R&D investment when they observe an increase (or decrease) in sales. The lagged cash flow to capital stock ratio is also positive and significant for agricultural firms, suggesting that they invest more in

R&D when expectations of future sales are high. Once again, this profit rate variable is not significant for food firms, implying that agricultural firms are more sensitive to changes in sales expectations.

[Table 2-6: Estimation Results: Firms with Moderate Variance of R&D Growth and Greater R&D]

The size of the coefficient for contemporaneous sales growth is larger in food firms than in agricultural firms as in our previous estimation. Again, R&D spending by food firms is more sensitive to a contemporaneous demand shock than for firms in the agricultural sectors. This may be because of the relatively small R&D intensities of food firms.

For firms that invest moderately large amounts in R&D (columns (4) and (8)), their decision on how much to invest in R&D is mostly affected by contemporaneous sales growth. Lagged sales growth and profit-rate variables are not significant for these large R&D investors, indicating that they only adjust their R&D amount with respect to current sales growth not by the expectations in sales.

2.7 Conclusion

Despite the growing and now important presence of private R&D in the food and agricultural sectors, we know surprisingly little about firms and their R&D investment behavior in these sectors. We have offered two contributions to help fill this gap in our knowledge. First, we constructed a replicable and comprehensive set of firm-level R&D data for the U.S. food and agricultural sectors spanning the period 1950-2014. Second, using those data, we have econometrically investigated the relationships among firm-level sales, profit rates, and R&D spending.

This newly updated and revised series of R&D-related data for U.S. food and agricultural firms reveals a diverse and complex pattern of R&D spending. For firms that conducted any R&D over the 65-year period beginning in 1950, the range in spending is remarkable. In our sample, 175 firms each spent less than $1.0 million annually on R&D

conducted in the U.S., 47 firms spent more than $50 million per firm, on average, and two firms spent in excess of $300 million per year.

We have found that firms invest more in R&D when expectation of future sales are high. Not only do firms invest more in R&D when past sales increase, they also invest more in capital which lowers the cash flow to capital ratio when anticipating future sales. We have found consistent but weak evidence that R&D investment increases when this cash flow to capital ratio decreases. These estimated effects are larger for food firms than agricultural firms, largely due to the lower R&D intensities of the food sector. Moreover, these conclusions generally hold under robustness checks designed to eliminate outliers in R&D growth that might otherwise skew the results.

The findings of this paper call attention to differences in the ways private firms invest in R&D, during a period where private R&D spending in the food and agricultural sectors in the United States has increased substantially and now accounts for a significant share of the overall spending. Although it takes considerable time to fully realize the returns to both food and agricultural R&D (Alston et al. 2009), the way firms respond to short-term market shocks varies across these sectors. Deepening our understanding of these R&D investment decisions has consequences not only for the innovation trajectories of the U.S. food and agricultural sectors, but may also shed light on the prospective patterns of innovation in other high-income and agriculturally-important middle-income countries (e.g., Brazil India and, especially, China), where private investment in food and agricultural R&D is also rapidly on the rise.

Table 2-1

Standard Industrial Classification (SIC) Codes Used

SIC	Industry Names	Ag or Food	Sub-sectors	Notes
100	Agriculture Pd-Crops	Ag	Ag & Chemical	
200	Agriculture Pd-Livestock & Animal Spec	Ag	Ag & Chemical	
700	Agricultural Services	Ag	Ag & Chemical	
800	Forestry	Ag	Ag & Chemical	
900	Fishing, Hunting And Trapping	Ag	Ag & Chemical	
2400	Lumber And Wood Pds, (No Furniture)	Ag	Ag & Chemical	
2421	Sawmills, Planting Mills, General	Ag	Ag & Chemical	
2430	Millwood, Veneer, Plywood	Ag	Ag & Chemical	
2511	Wood Household Furniture (No Upholstered)	Ag	Ag & Chemical	
2600	Paper And Allied Products	Ag	Ag & Chemical	
2611	Pulp Mills	Ag	Ag & Chemical	
2621	Paper Mills	Ag	Ag & Chemical	
2631	Paperboard Mills	Ag	Ag & Chemical	
2650	Paperboard Containers, Boxes	Ag	Ag & Chemical	
5150	Farm-Product Raw Materials (Wholesale)	Ag	Ag & Chemical	
2800	Chemicals & Allied Prods	Ag	Ag & Chemical	Large
2820	Plastic Matl, Synthetic Resin	Ag	Ag & Chemical	Large
2821	Plastics, Resins, Elastomers	Ag	Ag & Chemical	Large
2833	Medicinal Chems, Botanical Pds	Ag	Ag & Chemical	
2834	Pharmaceutical Preparations	Ag	Ag & Chemical	Large
2860	Industrial Organic Chemicals	Ag	Ag & Chemical	Large
2870	Agriculture Chemicals	Ag	Ag & Chemical	
2890	Misc Chemical Products	Ag	Ag & Chemical	Large
3523	Farm Machinery And Equipment	Ag	Machinery	
2000	Food And Kindred Products	Food	Food	
2011	Meat Packing Plants	Food	Food	
2013	Sausage, Other Prepared Meat Pds	Food	Food	
2015	Poultry Slaughter & Process	Food	Food	
2020	Dairy Products	Food	Food	
2024	Ice Cream & Frozen Desserts	Food	Food	
2030	Can, Frozn & Presrved Fruit, Veg	Food	Food	
2033	Can, Fruit, Veg, Presrv, Jams, Jel	Food	Food	
2040	Grain Mill Products	Food	Food	
2050	Bakery Products	Food	Food	
2052	Cookies And Crackers	Food	Food	
2060	Sugar & Confectionery Prods	Food	Food	
2070	Fats And Oils	Food	Food	
2080	Beverages	Food	Food	
2082	Malt Beverages	Food	Food	
2084	Wine, Brandy & Brandy Spirits	Food	Food	
2085	Distilled And Blended Liquor	Food	Food	
2090	Misc Food Preps, Kindred Pds	Food	Food	
2092	Prep Fresh, Frozn Fish, Seafd	Food	Food	
2100	Tobacco Products	Food	Food	
2111	Cigarettes	Food	Food	
2840	Soap, Detergent, Toilet Preps	Food	Food	
9997	Industrial Conglomerates	Food	Food	Large

Source: U.S. Securities and Exchange Commission (n.d.)

Note: For industries marked as “Large”, smaller firms are determined to be non-agricultural or non-food.

Table 2-2

Number of Firms in the Sample, by Sub-Sectors

	Total	Ag				Food
		Production & Chemical			Machinery	
		Sub-total	Ag	Chemical		
"All" Publicly-traded Firms (by SIC)	2,179	1,480	443	1,037	35	665
US Publicly-traded Firms (by SIC)	1,786	1,201	334	867	33	553
US Firms deemed Food and Ag-related	992	429	334	95	33	531
US Firms that conduct Food and Ag-related R&D	465	215	147	68	29	222
US Firms that conduct NO Food and Ag-related R&D	528	215	188	27	4	310

Source: Developed by author.

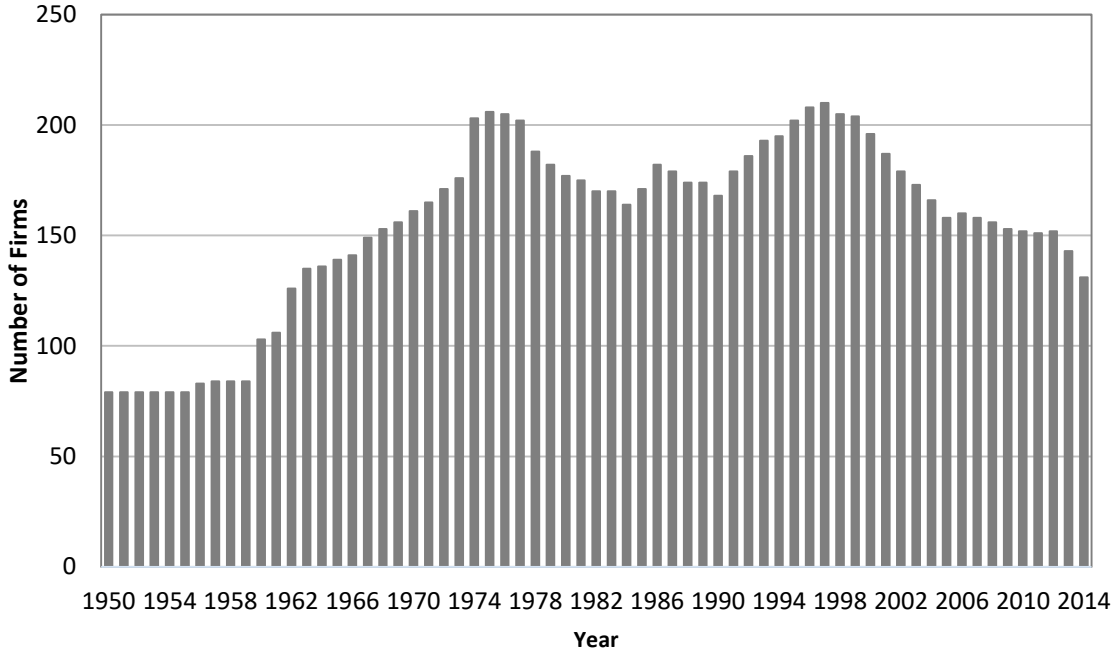
Note: In the sub-groups, Cargill is listed twice in both Ag and Food.

Table 2-3

Balance of Panel

Number of Years	Number of Firms Reporting R&D	Percentage
0 ~ 5	88	18.9%
6 ~ 10	70	15.1%
11 ~ 20	121	26.0%
21 ~ 30	75	16.1%
31~ 40	35	7.6%
41 ~ 50	21	4.5%
51 ~ 60	22	4.7%
61 ~ 64	5	1.1%
65	28	6.0%
Total	465	100.0%

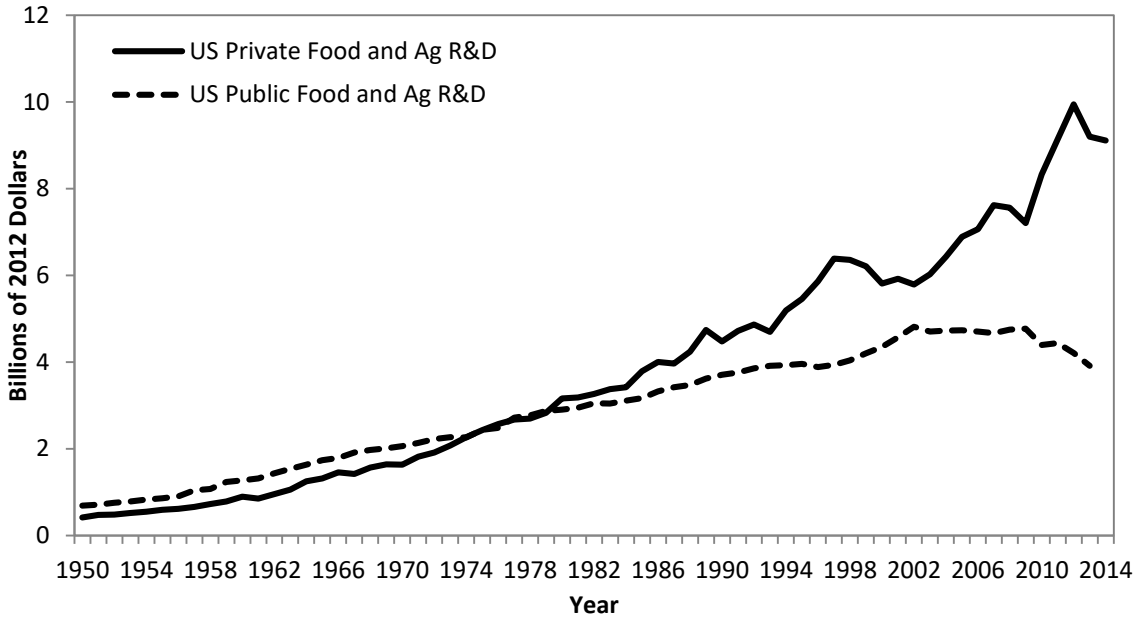
Source: Developed by author.



Source: Developed by author.

Figure 2-1

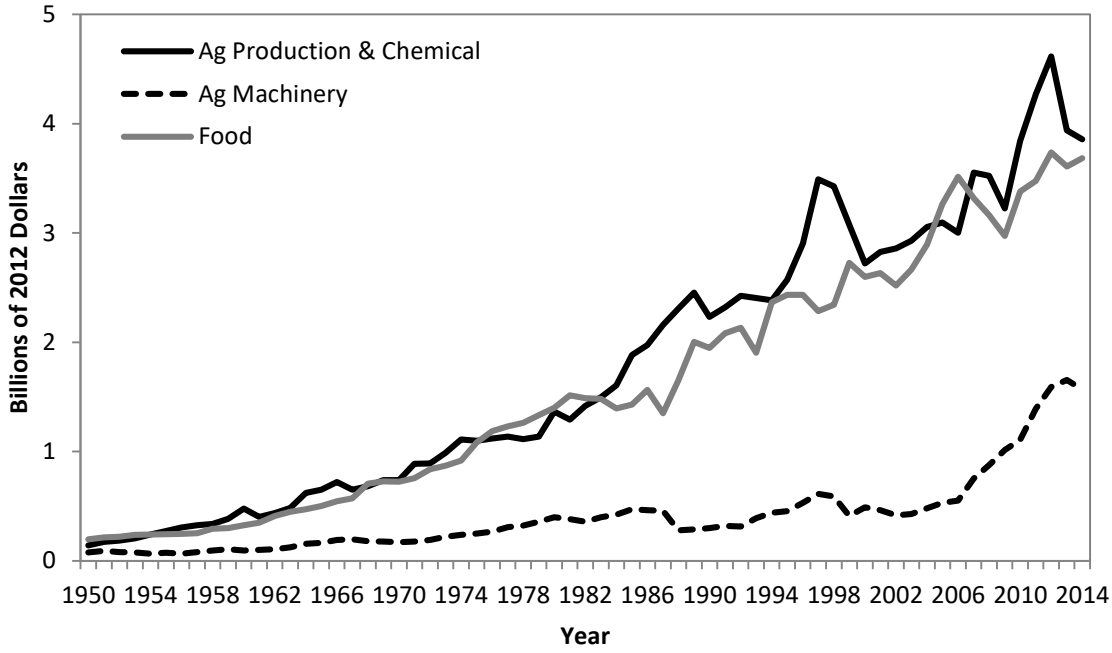
Number of Firms per Year



Source: The U.S. public food and agricultural R&D are from the InSTePP (Pardey et al. 2016), and the private U.S. series are developed by author.

Figure 2-2

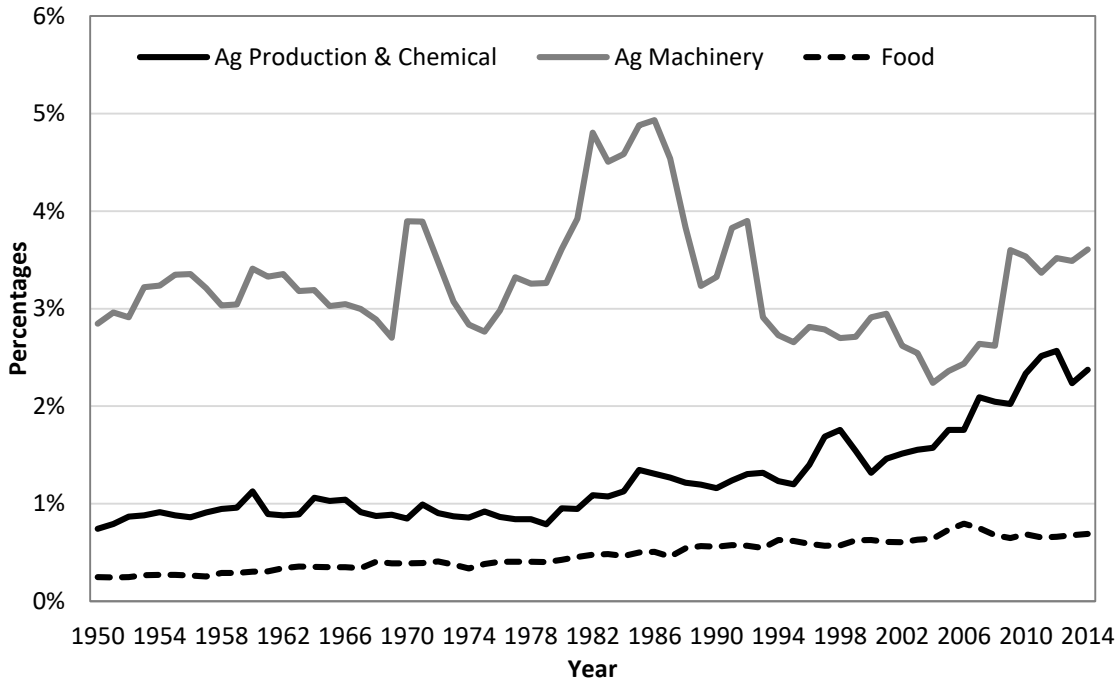
US Private and Public Food and Agricultural R&D Expenditures



Source: Developed by author.

Figure 2-3

U.S. Private Food and Agricultural R&D Expenditures, by Sub-sectors

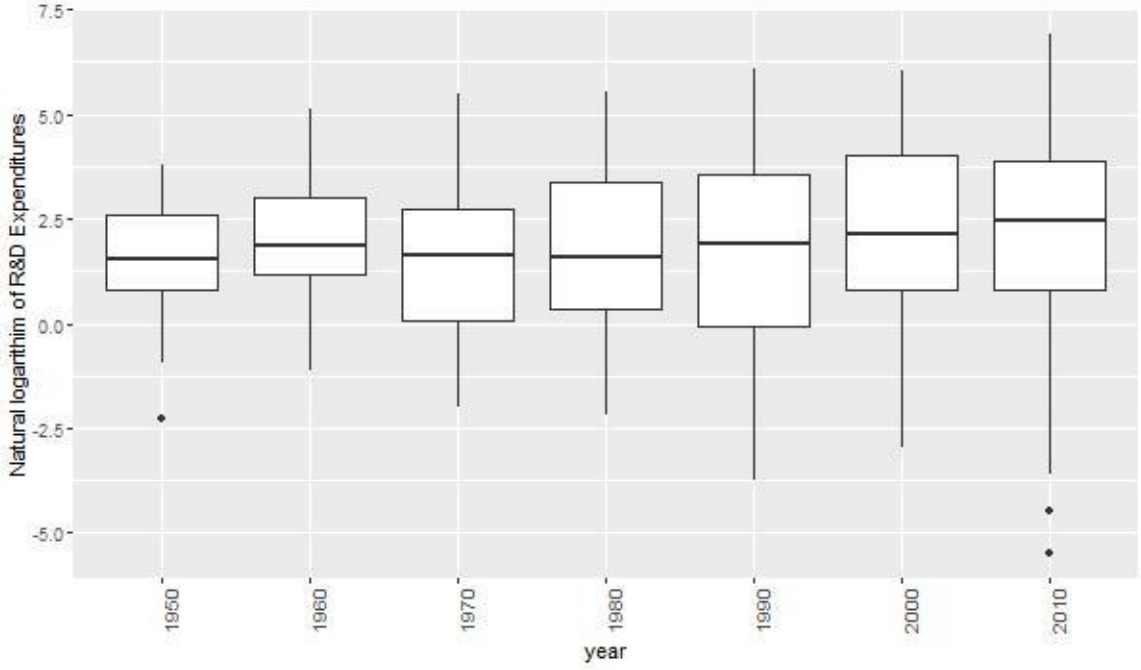


Source: Developed by author.

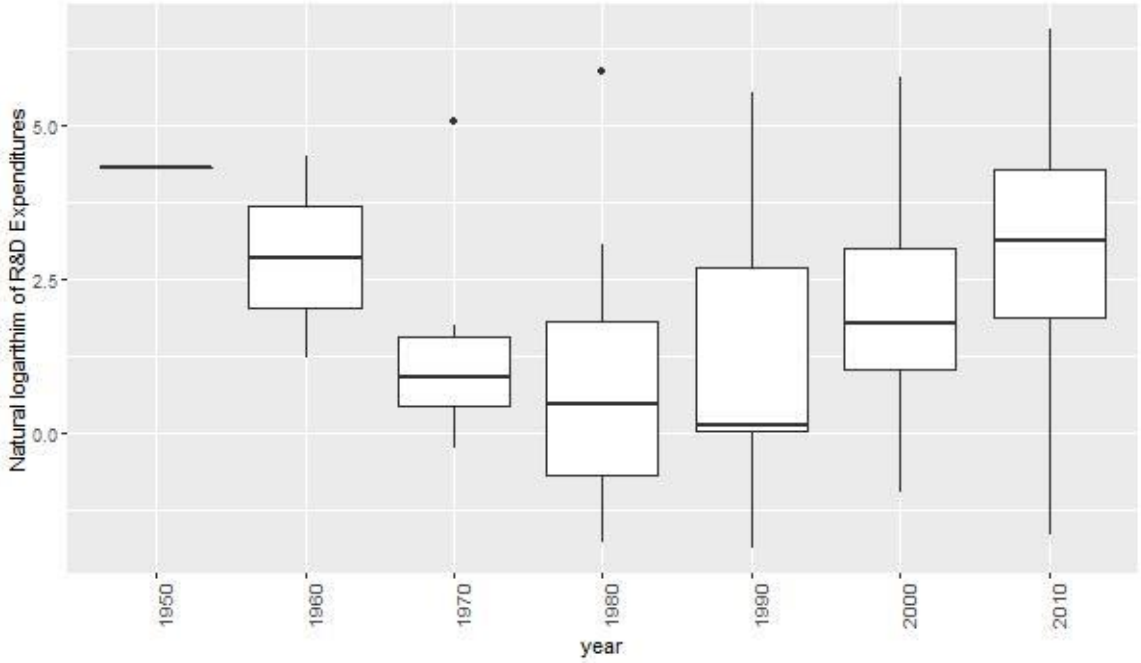
Figure 2-4

Share of U.S. Private Food and Agricultural R&D Expenditures in Sales, by Sub-sectors

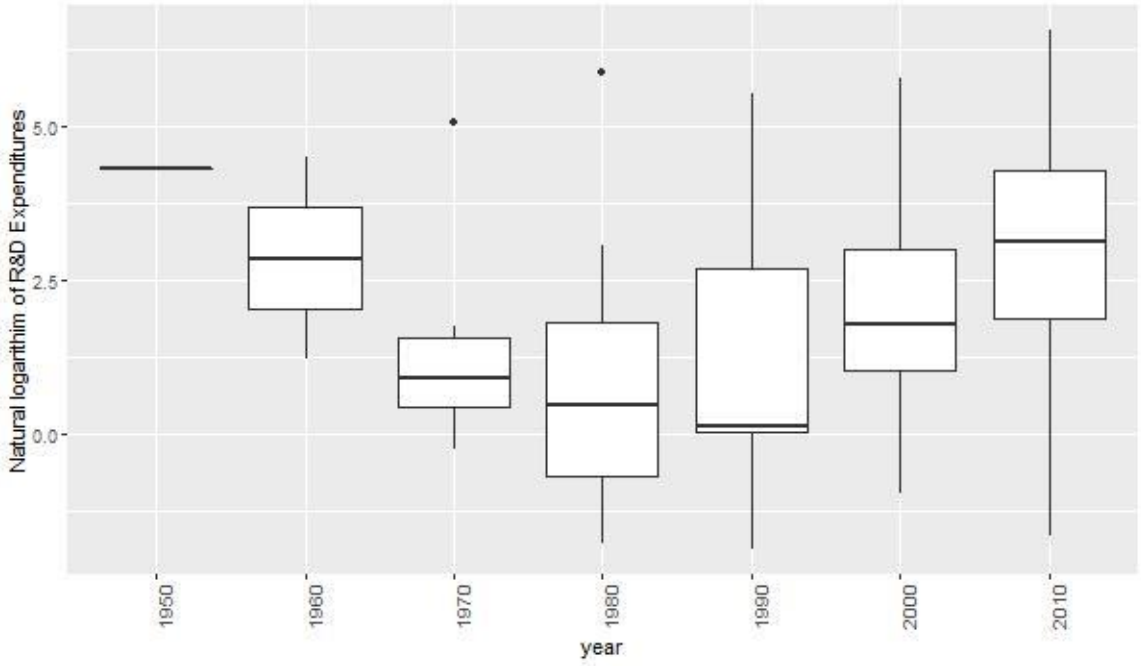
Panel A: Agricultural Production and Chemical (215 firms)



Panel B: Agricultural Machinery (29 firms)



Panel C: Food (222 firms)



Source: Developed by author.

Note: The boxplot shows the median as a horizontal line inside the box and the interquartile range (between the 25th and 75th percentiles) as the length of the box. The lines on the linearized distribution for each year indicate minimum and maximum values when they are within 1.5 times the interquartile range from either end of the box. Black dots are outliers. An outlier value is defined as a value that is smaller or larger than the 1.5 times the interquartile range The line graph connects the median R&D spending in each year.

Figure 2-5

Firm-level R&D Expenditures by Sub-sectors, Natural Logarithms

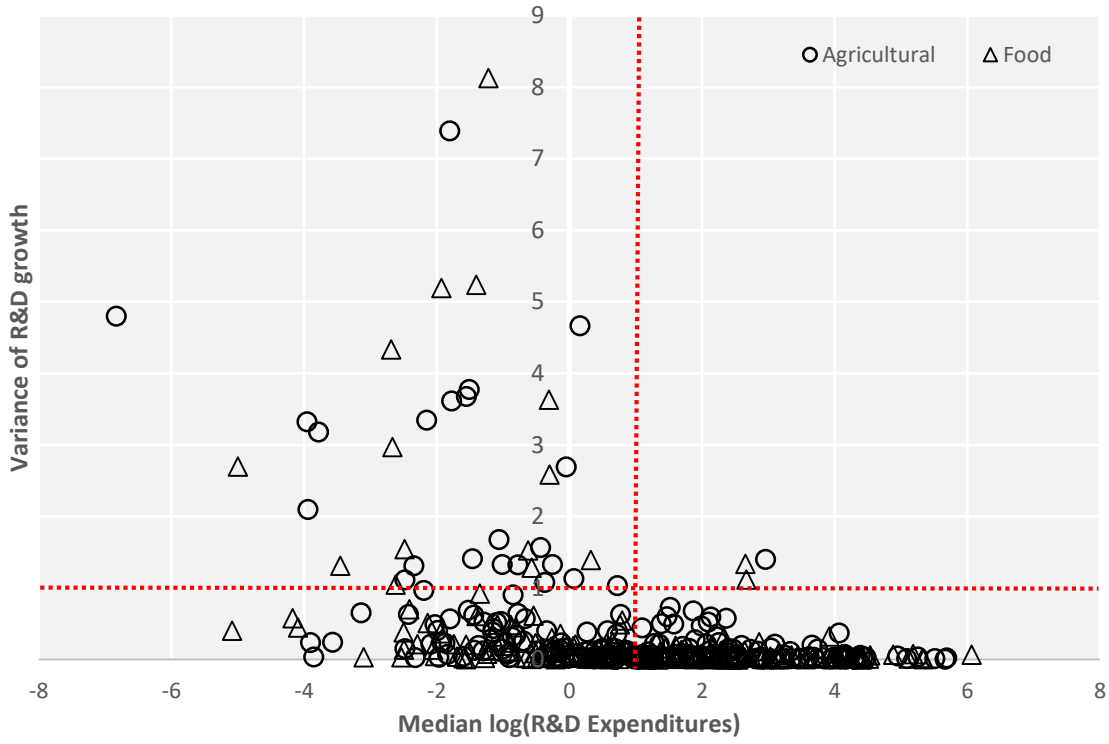
Table 2-4

Summary Statistics for U.S. Food and Agricultural R&D Investing Firms

	Number of Firms	Mean	Median	S.D.	Q10	Q90
R&D Investment (millions of 2012 dollars)						
Ag. Production & Chemical	215	33.6	5.8	86.17	0.2	87.3
Ag. Machinery	29	52.1	3.6	129.1	0.2	184.6
Food	222	27.7	4.3	63.5	0.2	75.8
Sales (millions of 2012 dollars)						
Ag. Production & Chemical	215	1,861.4	444.4	3,897.6	6.7	5,193.5
Ag. Machinery	29	1,464.3	284.6	3,203.2	13.8	4,131.3
Food	222	3,699.1	1,082.9	7,125.2	23.9	10,334.7
R&D Intensity (percentages)						
Ag. Production & Chemical	215	115.2	1.3	2,728.9	0.15	15.3
Ag. Machinery	29	3.9	2.2	23.2	0.6	4.0
Food	222	9.8	0.5	346.1	0.1	1.8
R&D Growth (in logarithm)						
Ag. Production & Chemical	215	1.7	1.8	2.2	-1.0	4.5
Ag. Machinery	29	1.7	1.3	2.4	-1.4	5.2
Food	222	1.5	1.6	2.2	-1.4	4.4
Sales Growth (in logarithm)						
Ag. Production & Chemical	215	5.7	6.1	2.6	2.2	8.6
Ag. Machinery	29	5.5	5.6	2.2	2.8	8.4
Food	222	6.6	7.0	2.3	3.8	9.3
Cash Flow/Capital Stock (percentages)						
Ag. Production & Chemical	214	-8.7	16.2	3,219.4	-95.6	57.4
Ag. Machinery	29	22.4	30.8	124.6	-25.6	98.9
Food	220	-216.1	35.0	6,535.8	-3.1	78.0

Source: Developed by author.

Note: These are only for R&D investing firms in 1950-2014. For the cash flow/capital stock variable, the data for two firms, Cargill Inc. and Mars Inc., are missing. Cargill Inc. is listed both in the agricultural production and food sectors.



Source: Developed by author.

Note: The red dotted lines indicate when the variance of R&D growth equals to one and the median R&D expenditures in natural logarithms equals to one.

Figure 2-6

Median R&D Expenditures (Natural Logarithms) over Variance of R&D Growth

Table 2-5

Estimation Results

	Agricultural		Food	
	(1)	(2)	(3)	(4)
R&D Growth, at t-1	-0.107	-0.267**	-0.172	-0.208**
	(0.084)	(0.077)	(0.160)	(0.084)
Sales Growth, at t	1.084*	0.884**	1.824**	1.515**
	(0.563)	(0.342)	(0.765)	(0.310)
Sales Growth, at t-1	0.137	0.214**	0.007	0.252*
	(0.127)	(0.108)	(0.265)	(0.135)
Sales, at t-2	-0.020	0.004	-0.122	0.101*
	(0.117)	(0.074)	(0.195)	(0.056)
Error Correction Terms	-0.031	-0.072	-0.109	-0.179*
	(0.085)	(0.113)	(0.213)	(0.096)
Cash flow to Capital Ratio, at t		-0.032		-0.147
		(0.053)		(0.113)
Cash Flow to Capital Ratio, at t-1		0.078*		0.102
		(0.042)		(0.070)
Hansen	0.046	0.074	0.147	0.184
AR(1)	0.000	0.019	0.000	0.000
AR(2)	0.217	0.790	0.536	0.679
# of Firms	209	153	179	156
Obs	3,247	2,120	3,073	1,932

Source: Developed by author.

Note: All two-step system GMM employed; Year effects included; Hansen is Hansen J test for overidentifying restrictions; AR(1) and AR(2) are tests for first and second-order serial correlation. P-values are reported next to Hansen and the serial correlation tests; Standard errors (in parentheses) are Windmeijer-corrected standard errors; For each estimation, instruments are as follows: For columns (1) and (2), r_{it-2} to r_{it-8} and y_{it-2} to y_{it-8} in the differenced equations and Δr_{it-1} and Δy_{it-1} in the levels equations. For columns (3) and (4), r_{it-2} to r_{it-24} , y_{it-2} to y_{it-24} and Π_{it-2} to Π_{it-24} in the differenced equations and Δr_{it-1}, Δy_{it-1}, $\Delta\Pi_{it-1}$ in the levels equations; ** coefficient significant at the 5 percent level, * at the 10 percent level.

Table 2-6

Estimation Results: Firms with Moderate Variance of R&D Growth and Greater R&D

	Agricultural				Food			
	Firms whose Var(R&D growth) is less than 1		Firms whose median log(R&D) exceeds 1		Firms whose Var(R&D growth) is less than 1		Firms whose median log(R&D) exceeds 1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
R&D Growth, at t-1	-0.126*	0.066	-0.066	-0.055	0.028	-0.170*	-0.103	-0.116
	(0.069)	(0.175)	(0.061)	(0.053)	(0.115)	(0.097)	(0.064)	(0.082)
Sales Growth, at t	0.810	1.228**	0.400	1.075**	1.161**	1.448**	1.508**	1.440**
	(0.595)	(0.223)	(0.511)	(0.231)	(0.413)	(0.326)	(0.245)	(0.199)
Sales Growth, at t-1	0.006	0.236	-0.048	0.040	-0.178	0.219*	-0.043	0.117
	(0.103)	(0.296)	(0.135)	(0.054)	(0.188)	(0.153)	(0.129)	(0.138)
Sales, at t-2	-0.118	0.061	-0.141	-0.015	-0.108	0.076	-0.048	0.037
	(0.108)	(0.048)	(0.156)	(0.050)	(0.106)	(0.072)	(0.065)	(0.061)
Error Correction	-0.055	-0.064	-0.053	-0.087*	-0.011	-0.201*	-0.082	-0.085
	(0.065)	(0.048)	(0.049)	(0.051)	(0.111)	(0.112)	(0.079)	(0.092)
Cash flow to Capital, at t		-0.204**		-0.044		-0.151		-0.082
		(0.074)		(0.060)		(0.122)		(0.129)
Cash flow to Capital, at t		0.171**		0.091**		0.105		0.022
		(0.063)		(0.038)		(0.066)		(0.059)
Hansen	0.258	0.100	0.080	0.131	0.304	0.188	0.719	0.234
AR(1)	0.005	0.002	0.000	0.000	0.000	0.000	0.001	0.000
AR(2)	0.562	0.451	0.105	0.546	0.421	0.536	0.864	0.702
Firms	187	144	143	116	166	151	117	109
Obs	3,097	2,059	2,641	1,810	3,019	1,907	2,393	1,558

Source: Developed by author.

Note: Two-step system GMM employed with year effects; Hansen is Hansen J test for overidentifying restrictions; AR(1) and AR(2) are tests for first and second-order serial correlation; P-values are reported next to Hansen and the serial correlation tests; Standard errors (in parentheses) are Windmeijer-corrected standard errors; For each estimation, instruments are as follows: For columns (1), (3), (5) and (7), r_{it-2} to r_{it-8} and y_{it-2} to y_{it-8} in the differenced equations and Δr_{it-1} and Δy_{it-1} in the levels equations. For column (2) and (6), r_{it-3} to r_{it-20} , y_{it-3} to y_{it-20} and Π_{it-3} to Π_{it-20} in the differenced equations and Δr_{it-2}, Δy_{it-2}, $\Delta \Pi_{it-2}$ in the levels equations. For column (4), r_{it-2} to r_{it-28} , y_{it-2} to y_{it-28} and Π_{it-2} to Π_{it-28} in the differenced equations and Δr_{it-1}, Δy_{it-1}, $\Delta \Pi_{it-1}$ in the levels equations. For column (8), r_{it-2} to r_{it-24} , y_{it-2} to y_{it-24} and Π_{it-2} to Π_{it-24} in the differenced equations and Δr_{it-1}, Δy_{it-1}, $\Delta \Pi_{it-1}$ in the levels equations; ** coefficient significant at 5 percent level, * at 10 percent level.

Appendix A

A.1 Definition of Capital Stock

To construct the capital stock variable, we follow a similar approach to Ottonello and Winberry (2018). Whereas Ottonello and Winberry (2018) used quarterly data to construct their capital stock variable, for our series we used annual data.

The details of our calculations are as follows. First, k_{jt+1}denotes the capital stock of firm *j* at the end of year *t*. For each firm we set the first value of k_{jt+1} to the level of gross plant, property, and equipment (Compustat mnemonic: *ppegt*) in the first period as obtained from Compustat. From this period onwards, we compute the evolution of k_{jt+1} using the changes of net plant, property, and equipment (*ppent*), which has significantly more observations than *ppegt* (net of depreciations). Missing values of *ppent* were linearly interpolated if the two annual adjacent values of *ppent* were available. Missing observations of *ppent* for two or more consecutive years were not linearly interpolated.

A.2 Summary Statistics for R&D and Non-R&D firms

Table A-1

Summary Statistics for U.S. Food and Agricultural R&D Investing and Non-R&D Firms

	R&D firms (465 firms)			Non-R&D firms (527 firms)		
	Mean	S.D.	Median	Mean	S.D.	Median
Employment						
Agricultural Production & Chemical	15.8	25.5	3.9	4.0	7.5	1.4
Agricultural Machinery	9.9	15.4	1.8	35.2	33.4	57.5
Food	20.0	42.7	5.5	2.1	3.8	0.8
Sales						
Agricultural Production & Chemical	1,861.4	3,897.6	444.4	1,037.2	2,827.6	268.2
Agricultural Machinery	1,464.3	3,203.2	284.6	8,787.4	8,480.7	13,509.2
Food	3,699.1	7,125.2	1,082.9	624.9	1,953.1	175.1
Cash Flow/Capital Stock						
Agricultural Production & Chemical	-8.7	3,219.4	16.2	28.6	374.6	21.7
Agricultural Machinery	0.1	1.6	0.3	-2,755.0	9,926.0	14.8
Food	-1.3	28.2	0.3	-234.1	3,856.9	28.4

Source: Developed by author.

3 Patterns of Research Collaboration in the Life Sciences

3.1 Introduction

Academic research has evolved to include more team research than solitary work. Researchers are forming larger and more diverse teams in terms of number of participating researchers (Wuchty et al. 2007), disciplines, (Porter and Rafols 2009), and locations (Jones et al. 2008; Adams et al. 2005.) Policymakers, funding agencies, and research institutes have also stimulated research collaboration to meet the needs of the broader sciences (O'Brien 2012), productivity growth (Ductor 2014), and creativity (Hemlin et al. 2004). Despite the increasing professional and policy interest in multi-disciplinary and team-based research, we still know comparatively little about the nature of these collaborations and how the characteristics of researchers, their institutional circumstances, and the nature of their research may affect the propensity to collaborate.

To help fill this gap, we examine patterns and trends in research collaboration both within and across institutional boundaries, as well as what factors might drive those partnerships. In particular, we characterize internal and external research collaborations by University of Minnesota (UMN) researchers in the life sciences, and identify whether and how they change across colleges and over time. We do so using a panel dataset of publications and coauthorship information drawn from Scopus for the period 1999 to 2014. This empirical setting allows U.S. to focus on a large and economically important sector with a rich history of collaboration (Azoulay et al. 2010) while making use of detailed individual-level researcher data.

The patterns and trends we identify are broadly consistent with researchers opting to invest in cross-institution collaborations based on a comparison of the relevant benefits and costs. Researchers in the life sciences at UMN are increasingly collaborating with peers at other institutions. Those collaborations offer some value to UMN researchers, as citation rates to publications produced in those collaborations are higher than for publications lacking external collaborators. The higher value (as measured by average

citations rates per paper) of those collaborations may stem from larger sets of coauthors per publication, which could influence citation rates either through publication quality or networks for promotion. However, the rise in cross-institution collaborations does not necessarily translate to a higher number of publications per researcher (productivity). Instead, the growth in outside collaborations coincides with a rise in the number of researchers, who increasingly completed their PhD training at other institutions. Because new faculty members trained elsewhere have existing connections to researchers at other institutions, the costs of finding and establishing collaborations with other institutions are likely lower. At the level of individual colleges within UMN, however, we have found heterogeneity in benefits and costs of collaboration across colleges.

This work contributes in several ways to a relatively small literature examining how much and why researchers are collaborating across institutional boundaries. First, while the rise of multi-institution collaboration has been documented (Jones et al. 2008), we examine a broad range of possible explanations for that increase in a benefit-cost framework. In particular, we examine whether cross-institution collaborations are associated with higher per-researcher productivity or per-publication citation rates. While the latter has received some attention (Jones et al. 2008; Katz and Hicks 1997), we go farther in asking whether an established link between the number of coauthors and citations (Wuchty et al. 2007) might explain the attractiveness of large, multi-institution collaborations. As a complement to investigating the role of perceived benefits of collaboration on the incidence of multi-institution projects, we also examine the influence of costs. In particular, we highlight the potential role of existing collaborator networks that new faculty bring to their home institutions. Those existing networks reduce the costs (search and communication) of cross-institution collaborations.

The rich, individual-level dataset we employ also confers several advantages. First, we can examine heterogeneity in the patterns and drivers of collaboration at a sub-field level (e.g. Biological Sciences vs. Medical School), compared with the coarser categories used in prior work (e.g. Science and Engineering vs. Arts and Humanities in Jones et al. 2008). Second, knowing the institution at which new faculty members earned their degrees

enables U.S. to address the aforementioned question about the role of prior networks in costs associated with external collaborations. Finally, the recency of the dataset offers an updated view of the state of collaboration in the life sciences.

While we study collaboration in the context of academic research, our findings can also inform understanding of collaboration within and across firms, especially when that collaboration targets the acquisition or leveraging of knowledge. As an intangible asset, knowledge is difficult for managers to supervise and control, making hierarchical and bureaucratic management of it within a firm ineffective. Instead, knowledge creation within a firm occurs as a fluid and evolving process that arises from a community of individual agents (Powell et al. 1996). Employees thus act as independent agents to some extent, contributing actively to knowledge in the process of transforming inputs into outputs (Spender 1996).[10] This flexible and non-bureaucratic community of individuals is similar to researchers in academia, suggesting our findings may have some relevance for understanding collaborations within and across firms.

The remainder of this chapter is organized as follows. Section 3.2 lays out the conceptual framework that delineates the benefits and costs of research collaboration and possible measures of collaborations. Section 3.3 describes the data sample and the construction of variables used in the analysis. Section 3.4 describes and assesses the patterns of research collaboration at both the university and college levels, and Section 3.5 concludes.

[10] These entities are rewarded with "high-powered" incentives (Lazear 2000). The payment they receive includes the knowledge and resources that they bring to the community, as well as payment based on measured performance (Gibbons 2005).

3.2 Research Collaboration

3.2.1 Conceptual Framework

Academic researchers, like firms, are likely to collaborate if the gains from collaboration exceed the costs.[11] After all, academic researchers voluntarily work with other researchers on self-selected topics of inquiry (Wagner and Leydesdorff 2005; Wang and Hicks 2014; Zhu et al. 2013; Stephan 2012). The benefits of collaboration may entail access to tangible research resources, but are also likely to include intangible gains such as division of labor and specialization (Agrawal and Goldfarb 2008), transfer of knowledge or skills, a source of stimulation and creativity, training of apprentice researchers such as graduate students or post-doctoral researchers (Katz and Martin 1997 or Barnett et al. 1988), and access to wider network of scientists (Katz and Martin 1997).[12] Second, collaborations across different institutions tend to have more citations and get published in higher-impact journals compared to papers with authors from fewer institutions (Jones et al. 2008; Katz and Hicks 1997; Freeman and Huang 2014). One possible mechanism is that collaborations across greater distances tend to have more authors per paper (Freeman et al. 2014; Adams et al. 2005), and papers written by larger teams are more likely to have higher citations than solo-authored papers (Wuchty et al. 2007; Chung et al. 2009). Third, collaborations invite contributions of data, material, or components from multiple sources (Freeman et al. 2014).[13]

Of course, collaboration also entails costs that solitary researchers rarely face. Examples include financial costs of transporting researchers or, sometimes, equipment; both direct and indirect time costs (Katz and Martin 1997); costs of organization and

[11] This calculus has parallels in inter-firm collaboration: firms are more likely to collaborate with partners that can bring complementary assets (Nohria and Garcia-Pont 1991), defray costs or share risks (Hagedoorn 1993), improve competitive position or market power (Kogut 1988; Pfeffer and Nowak 1976), or offer skills or knowledge that a firm lacks (Hamel 1991; Powell et al. 1996; Hagedoorn 1993; Hagedoorn and Schakenraad 1994). Given the nature of academic research, it is this last potential benefit of collaboration – access to new knowledge – on which we focus.

[12] See Azoulay et al. (2017) for post-doctorate and adviser matching.

[13] Freeman et al. (2014) used survey data for Particle and Field Physics, Nanoscience and Nanotechnology, and Biotechnology and Applied Microbiology.

communication (Hudson 1996); and distribution of proper credit for joint production (Freeman et al. 2014). As teams become larger, these costs also increase, including lower consensus, higher coordination costs, and more free-riding and related control mechanisms (Lee et al. 2015). Historically, these costs also increased with spatial distance, which tended to inhibit collaboration by, for example, limiting in-person, informal exchanges of ideas that might lead to collaboration (Katz 1993; Katz and Martin 1997; Freeman et al. 2014).

The relative magnitudes of these costs and benefits vary across disciplines. For example, the incidence and extent of collaborations are greater in sciences that are laboratory based and capital intensive (in terms of both physical and human capital) than social sciences, arts and humanities, and mathematics (Laband and Tollison 2000; Newman 2004; Wuchty et al. 2007). Thus, the incentives to collaborate in the social sciences and mathematics are somewhat vague compared with the hard sciences where the research and quality often heavily rely on capital-intensive equipment or the size of the laboratory (Wuchty et al. 2007).

Through this lens, the well-documented and marked increase in collaboration spanning both institutional and national boundaries suggests that the benefits of collaboration must increasingly exceed the costs.[14] On the benefit side, funding agencies increasingly value larger-scale, capital-intensive, and interdisciplinary projects (Finholt 2000; O'Brien 2012; Millar 2013). Many fields also now have established norms of coauthorship, such that peers convey informal benefits of approval upon researchers who abide by those norms and costs upon those who don't (Valderas 2007). Lastly, the number of researchers (supply) has continued to increase since the 1950s (Wuchty et al. 2007), partially due to the growing number of graduate students and postdoctoral researchers (Stephan 2012; Freeman 2015). With more possible collaborators, the best available coll aborator is likely to have knowledge that is more valuable to a project, such that research ers are willing to bear higher costs to find those collaborators (Katz and Martin 1997).

[14] See, e.g. Laband and Tollison (2000), Newman (2004), Adams et al. (2005), Wuchty et al. (2007), and Jones et al. (2008).

On the cost side, the rise in collaboration is likely due in part to dramatic reductions in the costs of communication and transportation technologies. Those advances have weakened the relationship between distance and cost of collaboration, leading to increases in remote collaborations, including those spanning multiple institutions and countries (Coccia and Wang 2016; Hsiehchen et al. 2015; Lariviere et al. 2006; Jones et al. 2008; Adams et al. 2005; Adams 2013; Freeman et al. 2014; Agrawal and Goldfarb 2008; Catalini et al. 2016; Kim et al. 2009).

3.2.2 Quantifying Collaboration

A major challenge in studying the effects of researcher characteristics on collaboration is measuring collaboration itself. Researchers work together in many ways, not all of which are measurable. Collaborators might patent a new finding, write a research paper, or offer informal advice to other researchers (Katz and Martin 1997). Informal outputs in particular are often difficult to quantify (see Laband and Tollison 2000 for one attempt), especially in a uniform way across disciplines with different norms for acknowledgement.

Facing this measurement challenge, many researchers quantify academic collaboration using its most visible outputs: patents and publications. Patent data can be effective in analyzing the fundamental innovation process since the data include information on assignees and inventors. However, the current process of relying on patent data has critical weaknesses. First, the assignee designated on the patent–often used to determine collaboration in empirical work–need not reflect the core researchers involved in creating the innovation. Assignments are often made ex-post and may be based solely on having simply contributed funding for a portion of the work that led to the invention (Pardey and Graff 2013). Second, the United States Patent and Trademark Office (USPTO) does not provide consistent and unique identifiers for inventors, and, thus, making disambiguation costly (Li et al. 2014).

The other primary means of measuring collaboration, and the one we use here, is coauthorship of publications. Although collaboration and coauthorship are not

synonymous, using coauthorship as an indicator of collaboration has three key advantages (Katz and Martin 1997): (i) it is invariant over time and verifiable by other investigators; (ii) it is practical in a sense that the size of a sample can be very large and relatively inexpensive to acquire; and (iii) the results from a bibliometric investigation may influence collaboration practices in the long run. Thus, hereafter, we use the term collaboration to refer to coauthorship.

3.3 Data and Methods

3.3.1 Data

To examine patterns of academic research collaboration, we assemble a panel dataset tracking 3,305 life sciences faculty members[15] at UMN, their publications, and their coauthors for the period 1999-2014. Life sciences departments include the College of Food, Agricultural and Natural Resource Sciences (CFANS), College of Biological Sciences (CBS), and Academic Health Center (AHC), which encompasses six different colleges and schools: School of Dentistry, Medical School (both in the Twin Cities and Duluth), School of Nursing, College of Pharmacy, School of Public Health, and College of Veterinary Medicine.

The panel data come primarily from UMN's Data Warehouse and Elsevier's Scopus.[16] From the Data Warehouse, we obtained a list of full-time equivalent faculty who worked in UMN life sciences departments from 1999 to 2014 and their relevant job employment data (i.e., affiliated department and year of employment/leave, and personal information such as gender, ethnicity and other attributes).[17] We used a number of job-related variables, including job titles and job codes, to identify full-time equivalent faculty

[15] Among 4,161 regular faculty members who have Scopus author IDs, only 3,305 of those researchers has an article published for 1999-2014 period. See the Appendix for the details on the number of original faculty and faculty with regular appointments in each college.

[16] This study was approved by the Institutional Review Board (IRB) at UMN (Study Number**:** 1512S81383.)

[17] Data start from 1999 in order to maintain data consistency and cleanness as advised by the Office of the Vice President for Research (OVPR) and the Office of Institute of Research (OIR).

and excluded those who were listed as adjuncts, visiting scholars, and teaching and clinical positions within the university, or those who held non-regular appointments even if the job title listed them as faculty.[18] Thus, the final set of faculty included full-time faculty who were either tenure-track (including those who had tenure but were retired) or who had regular appointments with the university even if they were not tenure-track faculty.

We then matched those faculty to publication data from Scopus using Scopus' unique author IDs.[19] For each publication, we collected the number of citations, the list of coauthors, and their affiliations. For papers solely written by UMN researchers, we also collected references cited, as well as details of each coauthors' other publications, including the number of citations to and references cited by those studies.

Given our interest in institutional boundaries and their role in shaping collaboration, we filtered this dataset to articles (hereafter "papers") published by life sciences faculty reflecting research conducted at UMN. [20] Researchers, much more so than firms, are mobile and often move from one institution to another. As such, it is very difficult to identify which research collaboration was initiated and completed at UMN. Faculty members who had left UMN may merely list the new affiliation address even though all of the research was conducted at UMN. Likewise, new faculty members may also list UMN on a publication that was conducted at another institution.[21] To better reflect UMN-related publications, we used a two-year shift before and after employment. For example, for a researcher who worked at UMN from 2002 to 2005, only their publications from 2004 (two years after first employment) to 2007 (two years after leaving UMN) were

[18] Faculty includes those whose job title or job classification corresponds to: Instructor, Assistant Professor, Associate Professor, Professor, Research Professor, Research Fellow, Regents Professor, Research Associate Professor, Research Assistant Professor, Head (With Faculty Rank), Director (With Faculty Rank). See the data documentation for details.

[19] The data were downloaded from Scopus API in August 2016, May, November, and December 2017, February, November, and December 2018, and January 2019 via http://api.elsevier.com and http://www.scopus.com. See the data documentation for details on acquiring and cleaning the Scopus author IDs.

[20] We sorted out articles by using the Scopus' categorization. The types of documentation in Scopus are: abstract report, article, article in press, book, chapter, conference paper, conference review, editorial, erratum, letter, note, report, review, and short survey.

[21] See Katz and Martin (1997) for the possible scenarios of multi-institutional papers.

used for data analysis.

We categorize research collaboration into three types based on affiliations of the authors: sole-authorship, inside collaboration (all UMN authors), and outside collaboration (at least one non-UMN author). Authors with UMN affiliations were identified using affiliation IDs from Scopus, together with pre-processing to account for typographical errors and incomplete affiliation names or addresses.[22] A combination of partial names and addresses were used to identify and associate 366 distinct affiliations with UMN. We then identify a researcher as "UMN-affiliated" if at least one of the affiliation addresses resolves to UMN. For example, a paper written by two researchers, A and B, with A having only a UMN affiliation, and B being affiliated with UMN and another institution, is classified as an inside collaboration. Thus, for our analysis, outside collaboration refers to collaborations by at least one UMN faculty in a life sciences department and at least one coauthor who does not have any UMN affiliation.[23]

3.3.2 Researcher Attributes

For 3,305 life sciences faculty members at UMN, we compute the following variables to characterize researchers' personal and professional attributes.

Number of Publications. Count of articles a researcher has published based on unique publication IDs from Scopus.

Number of (distinct) coauthors. Count of (distinct) coauthors in all articles based on unique author IDs from Scopus.

Number of (distinct) UMN Coauthors. Count of (distinct) coauthors with a UMN affiliation.

[22] However, there are still some papers that do not link individual authors with their institution affiliation but provide one or two affiliation names in a paper. In these cases, if the given affiliation names include UMN, then the paper is considered an inside collaboration.

[23] See the data documentation for details on cleaning affiliation names and dealing with missing affiliation names.

Average number of authors per paper. Count of number of authors in a paper using the unique author IDs from Scopus.

Average annual citation rate. Number of forward citations (including self-citations) to a paper as of January 2019 as listed in Scopus, divided by the number of years since publication.[24]

Professional Age. Publication year of an article minus the year of the researcher's first publication plus one, i.e. a researcher is professionally one-year-old when his or her first paper is published.

3.3.3 Outliers

Before examining patterns of collaboration, we identified and eliminated a number of outlier publications to avoid providing a distorted view of research collaboration at UMN. For example, summary statistics reveal the most productive author produces almost three times the publications per year as one in the 99th percentile (Table 3-1, Panel A). These outliers are not necessarily "superstars" who are very productive and truly have huge numbers of publications, coauthors, and citations.[25] Instead, sometimes researchers have one extreme paper which is written by hundreds of coauthors and highly cited.[26]

[Table 3-1: Summary Statistics of UMN Life Sciences Faculty]

For these extreme outliers, we eliminate all of a researcher's publications in a year when a researcher in a given year exceeds the 99th percentile of the entire distribution of the number of publications, coauthors, authors per paper, and citations at least twice. As a result, 2,363 papers are dropped and the final dataset consists of 3,296 faculty members

[24] Chung et al. (2009) also adjusted the number of citations by the number of years since publication.

[25] See Azoulay et al. (2010) and Castaldi and Los (2012) for the discussion of "superstar" life sciences researchers and patenters.

[26] For example, one paper in *Radiology* in 2011 was written by 1,339 coauthors from 47 different affiliations and has already been cited 449 times (as of March 2018). These outliers skew the averages, means, and medians of our data.

and 44,022 papers. The final data (Table 3-1, Panel B) are still positively skewed, though less so, by very productive researchers.[27]

3.4 Results

3.4.1 Research Collaboration from 1999 to 2014

Collaborative research by life sciences faculty at UMN increased dramatically between 1999 and 2014 (Figure 3-1). The total number of publications nearly doubled during that period (1,887 to 3,523 papers). This doubling was associated with a significant increase in outside collaborations. Those cross-institutional collaborations surged from 57 percent of total publications in 1999 to 74 percent in 2014, while the fractions of both sole-authored papers and inside collaborations declined. The proportion of international collaborations has also increased as other researchers have noted (Adams et al. 2005; Freeman et al. 2014).

[Figure 3-1: Total Number of Publications by Type and Year]

While overall publication output and the incidence of cross-institution collaborations have risen, those collaborations are not necessarily driving higher productivity. Indeed, productivity, as measured by papers per faculty per year, has not changed substantially from 1999 to 2014 (Figure 3-2), even declining until 2004 due to rapid faculty growth. Instead, there are simply more UMN faculty members through time, which drives up total production but not average productivity. That rise simply happens to coincide with an increase in outside collaboration.

[Figure 3-2: Total Publications per-faculty Member]

If higher rates of outside collaboration are not associated with higher publication output, researchers may engage in collaborations in hopes of boosting citations. One

[27] Silverberg and Verspagen (2007) found that the number of patent citations has both a lognormal and a Pareto distribution. Although we can assume that the number of citations for academic publications would have the similar distribution, we could not find any other analysis on this.

hypothesis is that outside collaborations have more authors, which may translate to more citations. In our data, outside collaborations do indeed have more coauthors per paper than inside collaborations across all years (Figure 3-3). Further, those outside collaborations also tend to receive more citations (Figure 3-4). Those differences in both mean and median citation rates across collaboration types are also statistically significant. Those citations may come from larger lists of coauthors, who have different networks for promotion or whose contributions may lead to higher quality work.

[Figure 3-3: Number of Coauthors per Paper and Coauthor Composition by Collaboration Type]

[Figure 3-4: Number of Citations by Type of Collaboration]

Finally, outside collaborations may also be on the rise if the costs of those collaborations are declining. We investigated one channel for declining costs: existing external networks of new UMN faculty. Specifically, we identified institutions where new faculty members received their post-graduate degree from CVs, personal or departmental websites, healthcare providers' websites, U.S. News Health, and, in a few cases, from obituaries.[28] The proportion of researchers who received their post-graduate degrees from UMN has decreased over time while those who received their degrees overseas has increased (Figure 3-5, Panel A). These non-UMN graduates have different research networks from UMN graduates: in all years, those who earned their degrees from other U.S. universities have more coauthors from other universities (Figure 3-5, Panel B). Those who have international degrees, on the other hand, tend to have a similar number of non-UMN coauthors to UMN graduates.

[Figure 3-5: Characteristics of New Faculty Hires]

[28] Among the 402 researchers hired in 2000, 2006, and 2012, we were able to obtain degree information for 339 researchers (84 percent). In most cases, we looked for the names of institution and country where new faculty members earned their Ph.D degrees. For Medical School faculty members, we looked for institutions that they got their MD degrees.

3.4.2 College-Level

In this section, we disaggregate the data to examine whether our findings of research collaboration are common among all eight life sciences colleges, or driven by only a few large colleges.[29]

The increase in the total number of publications is also notable in college-level data, except for the School of Dentistry (Figure 3-6). The increase in overall publication output is again associated with an increase in outside collaborations, except for the College of Pharmacy, whose proportion of outside collaboration has decreased through time (from 63 percent to 58 percent). The college-level data also paint a richer picture of how the benefits and costs of collaboration may vary across colleges. Unlike at the university level, per-researcher productivity does increase alongside the rise in outside collaborations for all colleges except the Medical School and the College of Pharmacy. This pattern is consistent with collaborations boosting productivity in many colleges, with the lack of support for this channel at the university level being driven by the large Medical School (Figure 3-7). The same reasoning may also explain the more mixed evidence at the college level that higher citation rates might influence the increase in collaboration. While the number of coauthors per paper remains higher for outside collaborations across colleges (Figure 3-8), the citation boost from outside collaboration is less consistent (Figure 3-9). Outside collaborations earn a substantially higher number of citations only in the Medical School, the Colleges of Biological Sciences and Veterinary Medicine.

[Figure 3-6: Types of Collaborations by College, 1999 and 2014]

[Figure 3-7: Per-faculty Total Publications by College]

[Figure 3-8: Number of Coauthors by College]

[Figure 3-9: Number of Citations by College]

This weak citation boost in outside collaborations for some colleges signals that there must be some other benefits for researchers to engage in cross-institution

[29] See the Appendix for summary statistics for college-level data.

collaborations, such as finer division of labor and specialization, or the increased supply of researchers who can contribute to data, materials, and instruments. For example, the higher incidence of outside collaborations for Public Health faculty suggests that they may have finer division of labor and specialization when collaborating, and better funding to attract more authors from outside than researchers in other colleges.[30]

3.5 Discussion and Conclusion

The perceived rise of collaborative research raises a range of questions about who collaborates, why, and whether those partnerships are ultimately beneficial. We investigate these questions using the panel data dataset of publications and coauthorship information drawn from Scopus between 1999 and 2014. We have found that the total number of publications nearly doubled during that period, and was concordant with an increase in outside collaborations. This suggests that researchers increasingly perceive that the benefits from outside collaborations outweigh the costs that arise from collaborating across institutions. Those benefits apparently do not include increased productivity at the university level, as per-publication output remains flat during the same period, with overall publication output simply rising along with the number of UMN faculty members. Instead, researchers may pursue collaborations to boost citation rates: outside collaborations have more authors per paper and more citations. On the other hand, we also find evidence consistent with declining costs of cross-institution collaboration. In recent years, a greater share of new UMN faculty receive their post-graduate degrees from non-UMN institutions, which may lower the search costs of identifying coauthors for cross-institution collaborations.

Re-examining these questions at the college level reveals patterns consistent with heterogeneity in benefits and costs of collaboration across colleges. Trends in outside collaboration, productivity, and citation rates in the Medical School align with the

[30] Researchers at the School of Public Health at UMN receive more funding from the National Institute of Health (NIH) than any other school of public health at a public university, and has the highest per-researcher funding across UMN (School of Public Health n.d.).

university-wide pattern. However, the large size of the Medical School likely masks contrasting relationships in the university-wide analysis. For example, in many other colleges, per-researcher productivity does increase through time, but citation rates to cross-institution collaborations are not consistently higher. Thus, researchers in those other colleges may seek outside collaborations in search of productivity rather than citation gains.

The findings of this paper make important contributions to the existing literature on collaborations. First, the analysis on the patterns of research collaboration in the life sciences has examined the new set of questions about why and to what extent researchers collaborate across institutional boundaries using a benefit-cost framework. Second, while higher citation rates for cross-institution collaborations have been documented (Jones et al. 2008; Katz and Hicks 1997), our analysis presents a possible channel between the number of coauthors and citations (Wuchty et al. 2007) using detailed sub-field level data. Third, we have examined the reduction in costs (search and communication) of cross-institution collaborations, by using individual-level data describing existing collaborator networks that new faculty bring to their home institutions.

With the increasing demand and attention for research collaboration across disciplines and institutions, the findings of this paper may help inform the benefits and costs that arise from cross-institution collaborations and the differences across colleges in the life sciences. That understanding can inform research administrators, funding agencies, and academia as a whole who would like to promote and benefit from research collaborations. The findings may also shed light on collaboration within and across firms, especially when that collaboration targets the acquisition or leveraging of knowledge. The findings on this paper, however, are only associations and do not provide any causal effects. One important avenue for future research is to investigate the causal link between cross-institution collaboration and the number of citations.

Table 3-1

Summary Statistics for UMN Life Sciences Faculty

Panel A: All Dataset

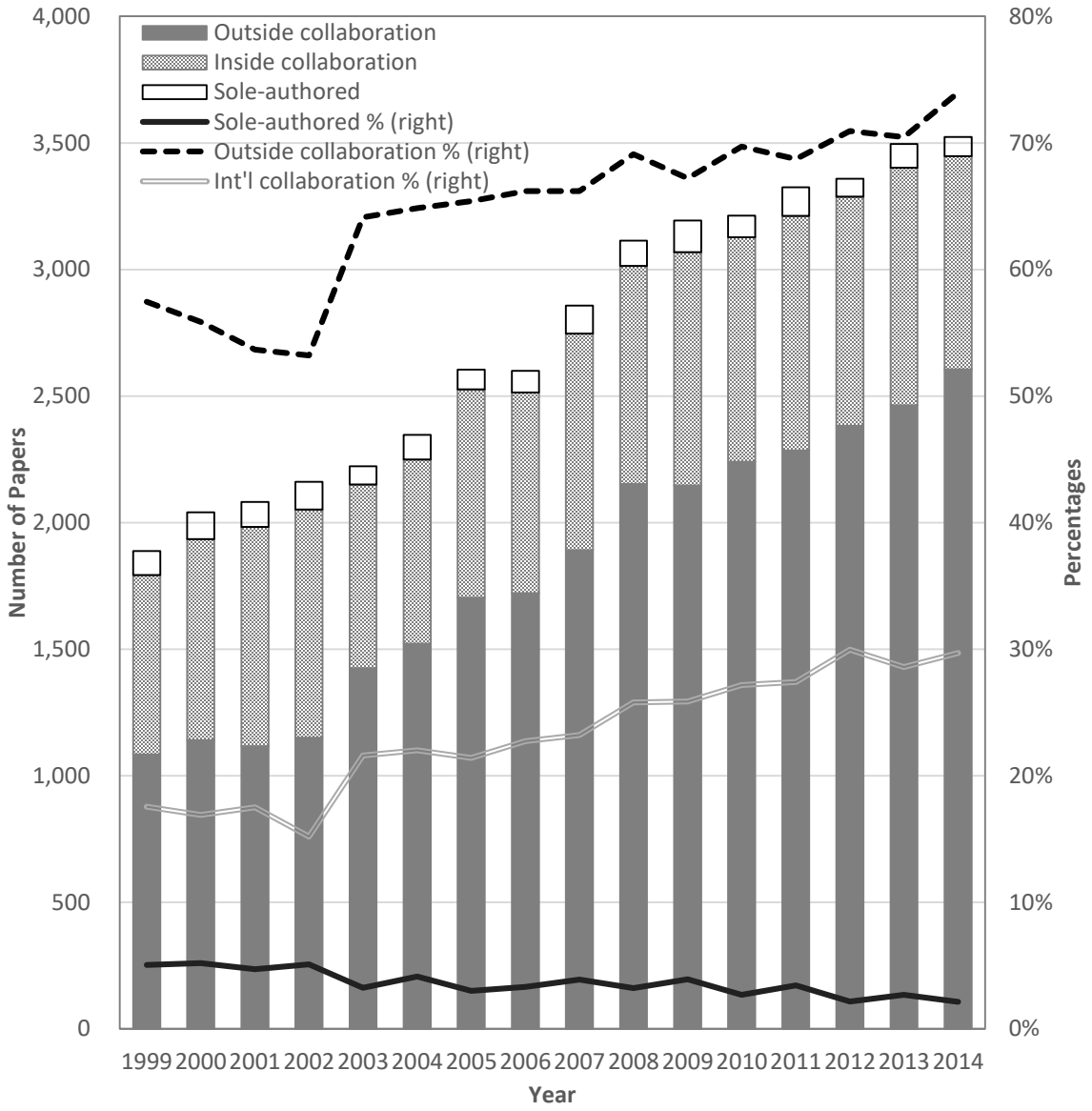
	Mean	Median	Std. D.	Min	P99	Max
Number of Publications per Researcher	3.3	2.0	3.4	1.0	17.0	51.0
Number of Coauthors per Researcher	16.2	9.0	30.3	0.0	121.0	1378.0
Average Number of Authors per Paper per Researcher	6.5	5.5	6.4	1.0	27.0	219.0
Average Number of Citations per Paper per Researcher	3.5	2.2	6.1	0.0	25.3	291.6

Panel B: Data Excluding Outliers

	Mean	Median	Std. D.	Min	P99	Max
Number of Publications per Researcher	3.2	2.0	3.1	1.0	16.0	38.0
Number of Coauthors per Researcher	14.5	9.0	18.0	0.0	91.0	325.0
Average Number of Authors per Paper per Researcher	6.3	5.5	4.6	1.0	22.0	112.0
Average Number of Citations per Paper per Researcher	3.4	2.1	5.5	0.0	23.5	153.3

Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

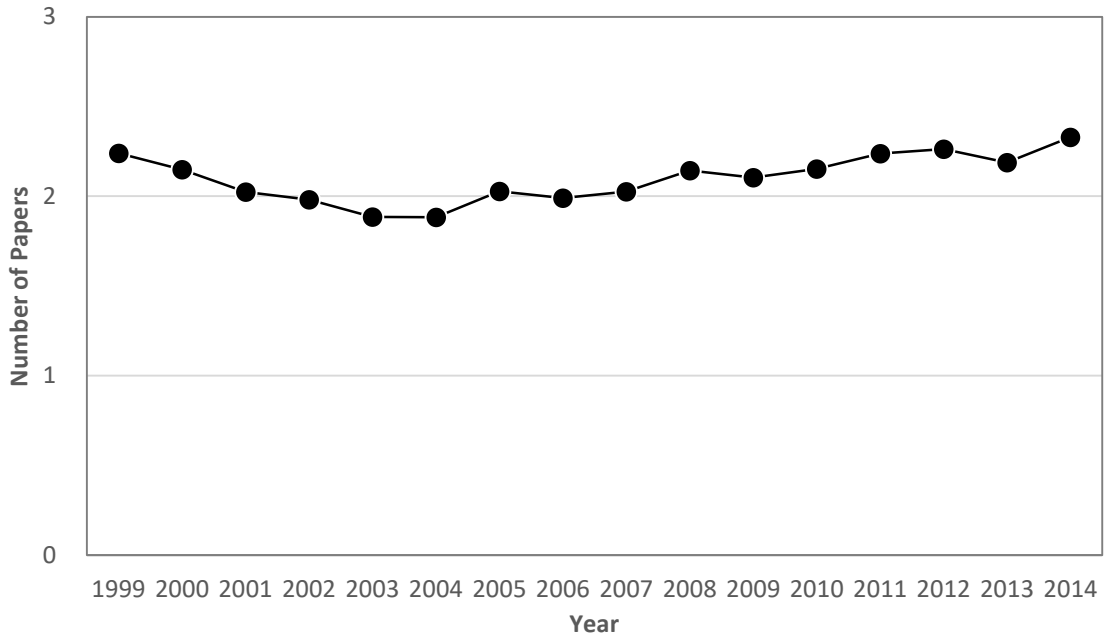
Note: The number of citations per paper is standardized, divided by the number of years since publication.



Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

Figure 3-1

Total Number of Publications by Type and Year

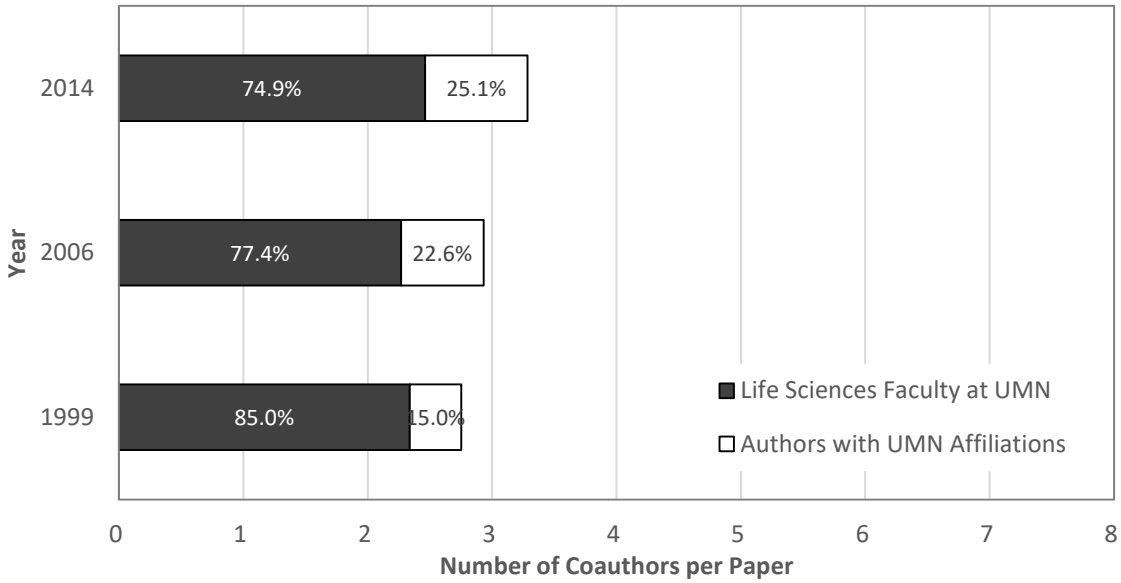


Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

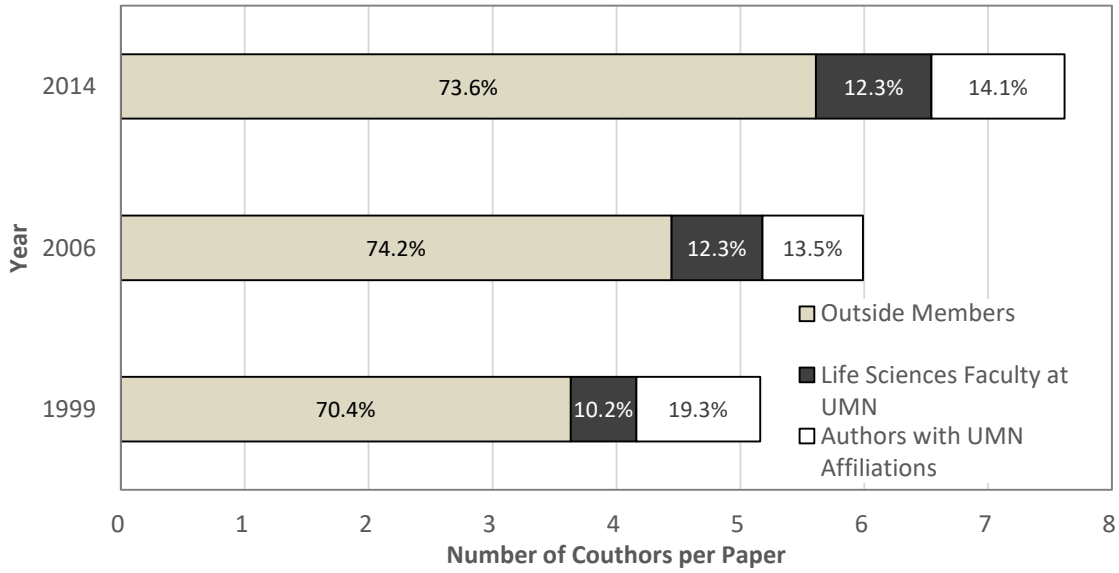
Figure 3-2

Total Publications per-faculty Member

Panel A: Number of Coauthors per Paper and Coauthor Composition for Inside Collaborations



Panel B: Number of Coauthors per Paper and Coauthor Composition for Outside Collaborations

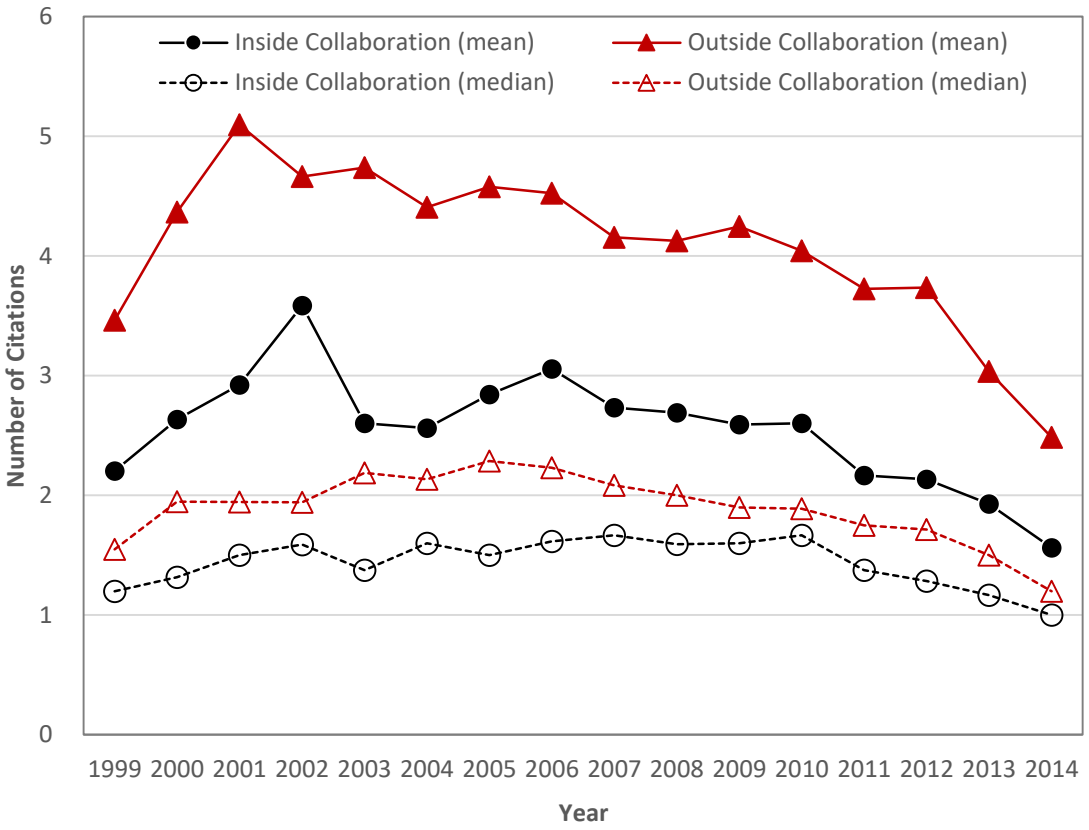


Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

Note: "Authors with UMN affiliation" refers to coauthors whose affiliation information reveals that they are affiliated to UMN but not identified as life sciences faculty. These coauthors may include UMN faculty members in non-life sciences departments, and graduate students or post-docs. "Outside members" refer to coauthors whose affiliation is not UMN. The numbers represent the number of coauthors in a paper, not the total number of authors per paper.

Figure 3-3

Number of Coauthors per Paper and Coauthor Composition by Collaboration Type



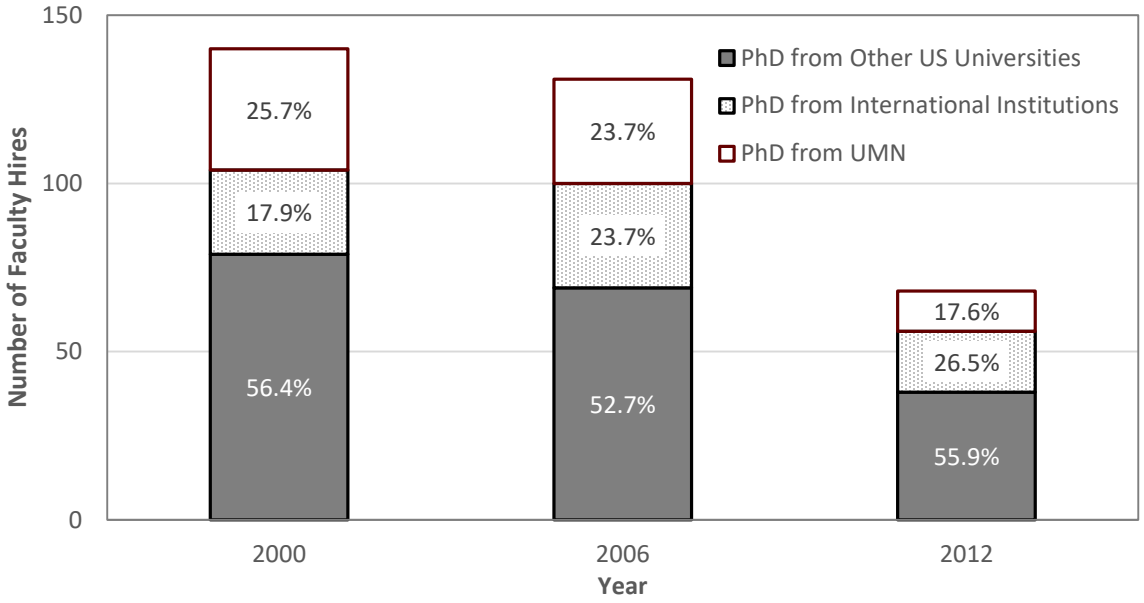
Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

Note: The figure shows forward citations of inside collaborations and outside collaborations. Year indicates the year of publication of the cited papers. The equality of the mean and median is tested each year and the null hypothesis is rejected at the five percent significance level.

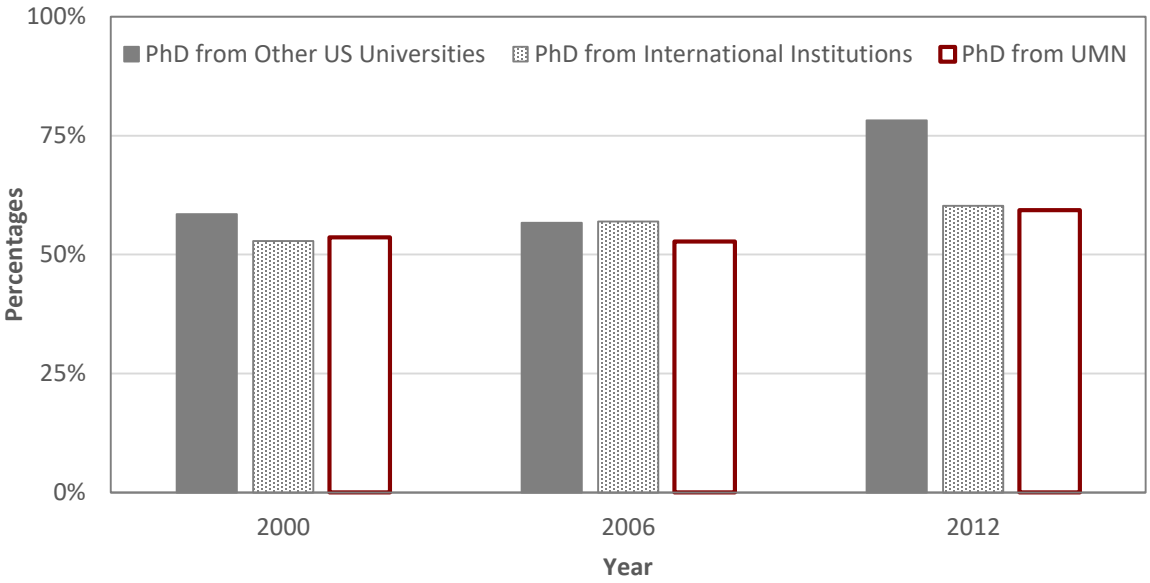
Figure 3-4

Number of Citations by Type of Collaboration

Panel A: Number and Proportion of Degrees for New Faculty Hires



Panel B: Proportion of Non-UMN Coauthors among all Coauthors



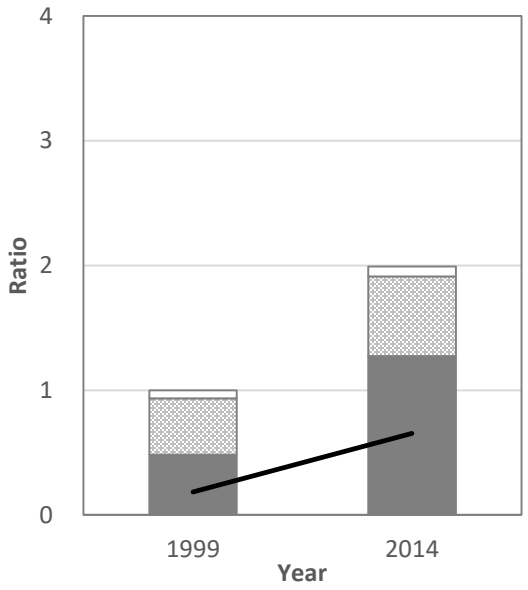
Source: Developed and calculated by author using data from Scopus, UMN Data Warehouse, and individual researchers' CVs, personal or departmental websites, healthcare providers' websites, U.S. News Health website, and, in a few cases, from obituaries.

Note: Year shows the year faculty members are hired. In Panel (b), the percentages show the proportion of non-UMN coauthors divided by the total number of coauthors from the publication that the new faculty members produced from the year they were hired until the year they leave UMN. The last year of observation is 2014.

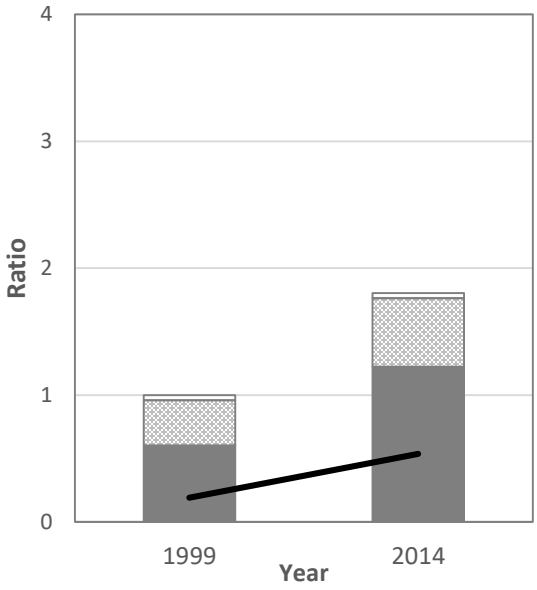
Figure 3-5

Characteristics of New Faculty Hires

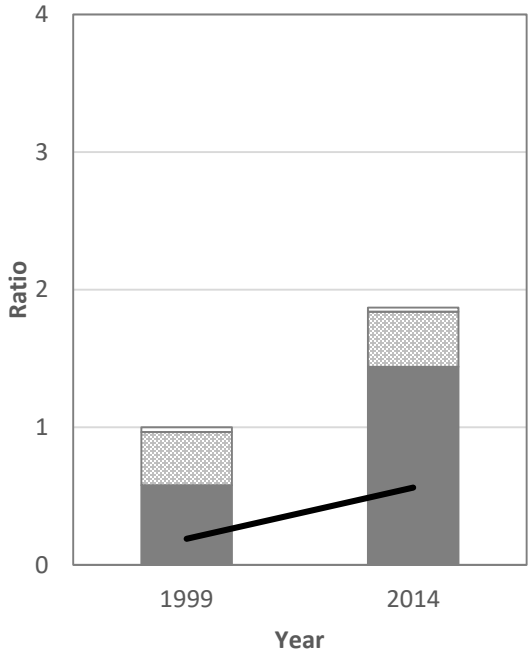
Panel A: Biological Sciences



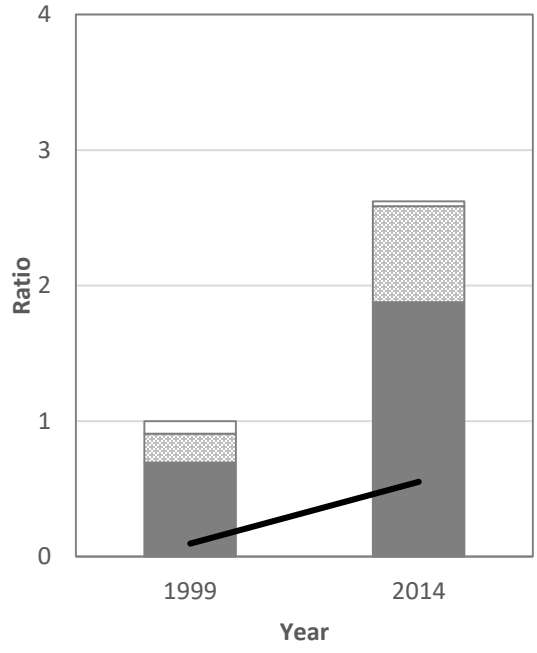
Panel B: Food, Ag, and Natural Resources Sciences

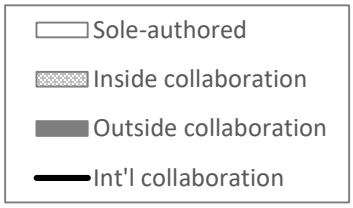


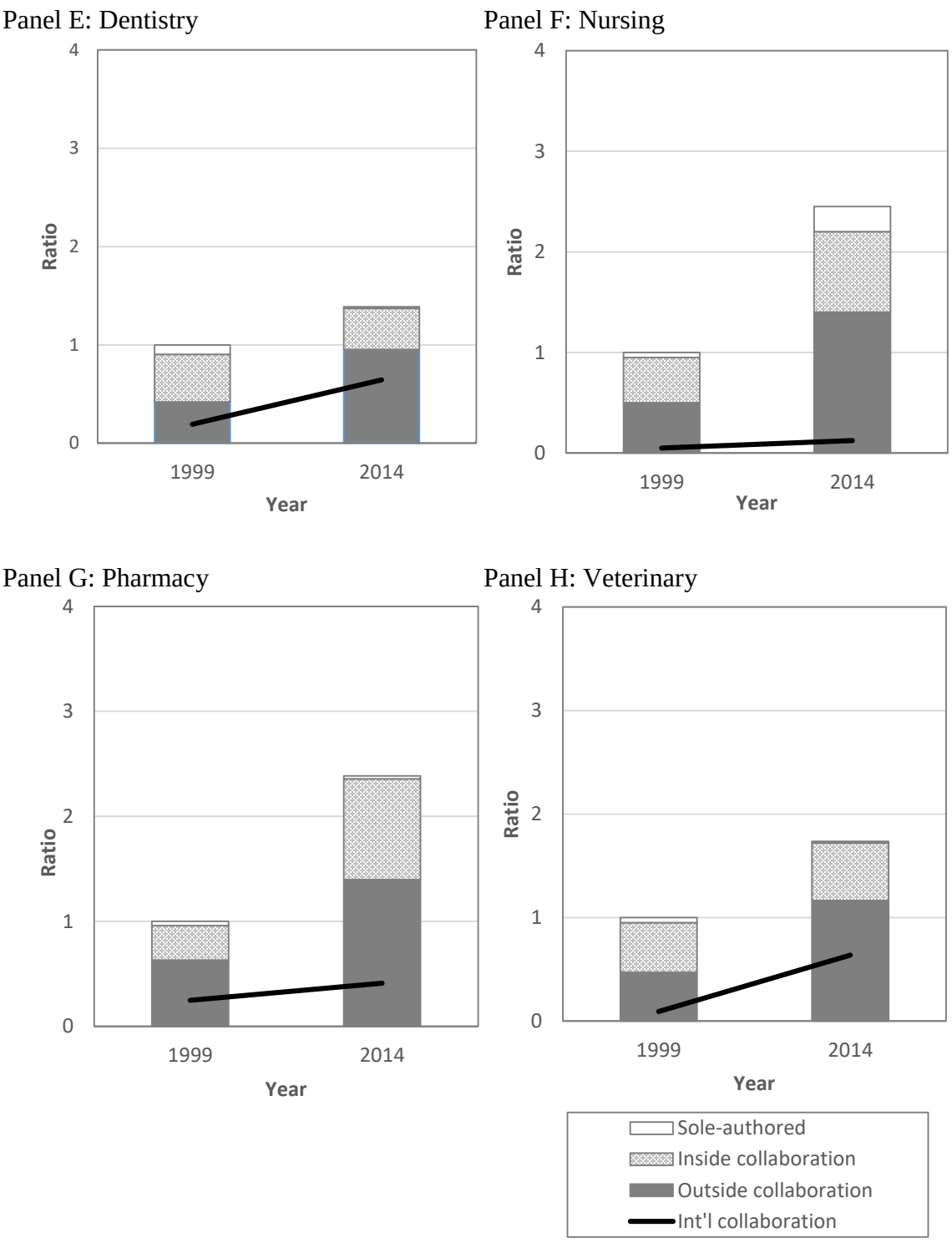
Panel C: Medical



Panel D: Public Health







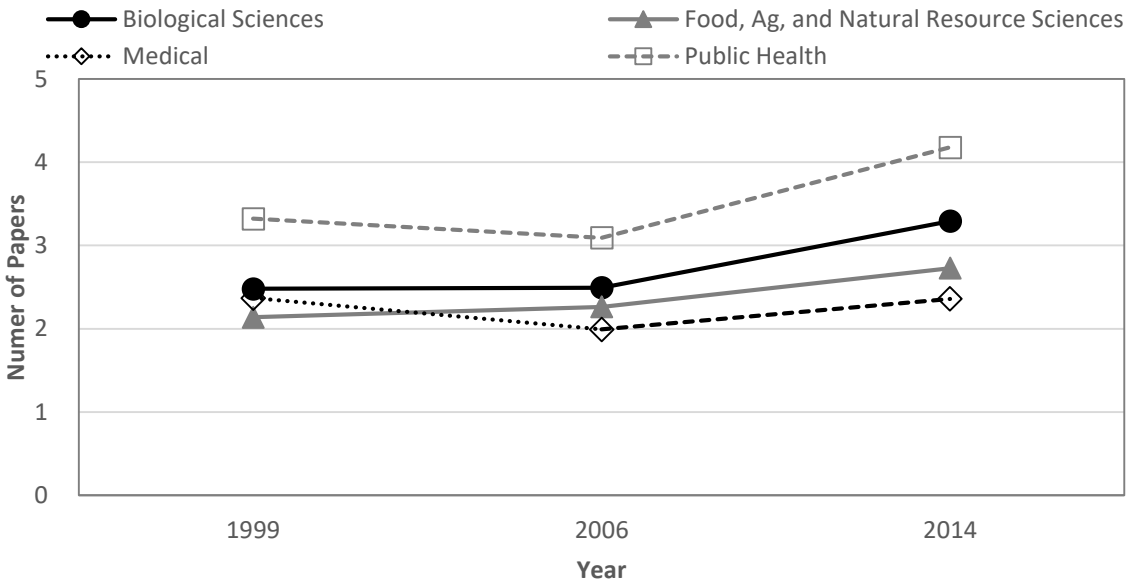
Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

Note: Height of bar graph in 1999 is set to one.

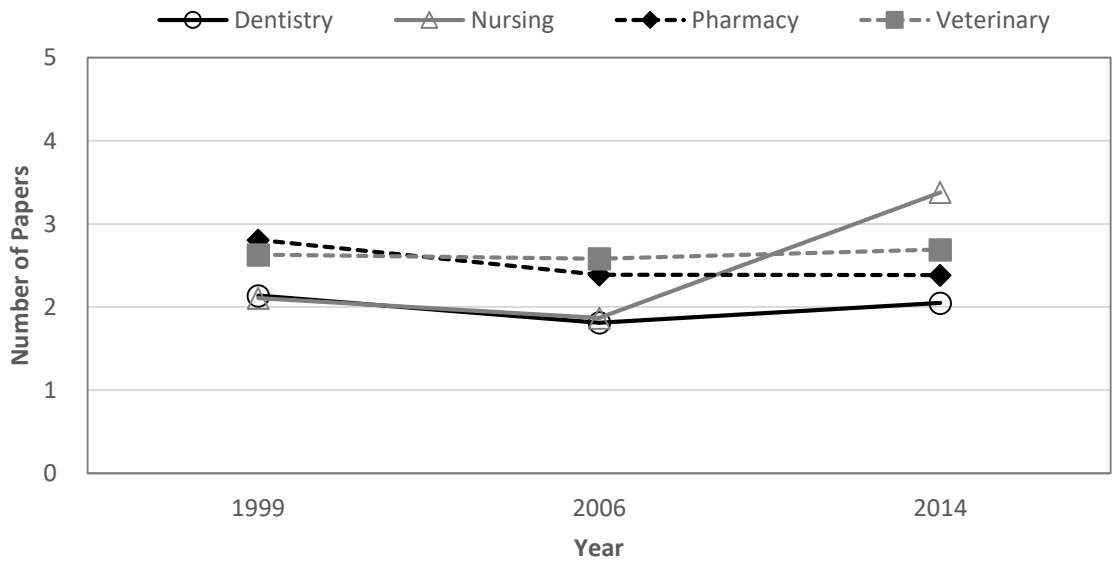
Figure 3-6

Types of Collaborations by College, 1999 and 2014

Panel A: Per-faculty total publications for College of Biological Sciences, College of Food, Agriculture, and Natural Resources Sciences, Medical School, and School of Public Health



Panel B: Per-faculty total publications for School of Dentistry, School of Nursing, College of Pharmacy, and College of Veterinary Medicine

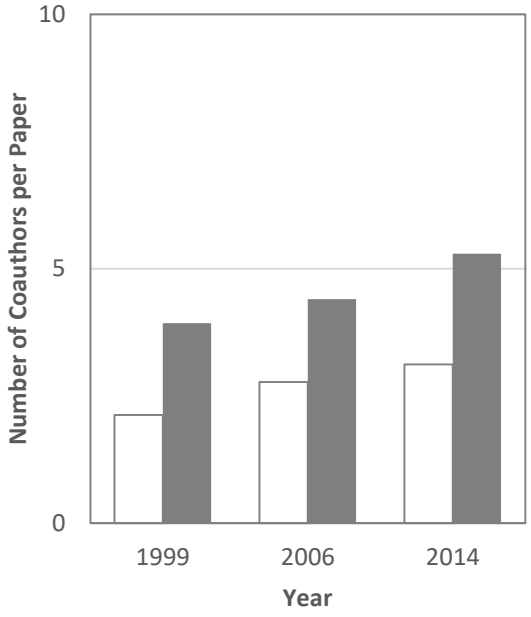


Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

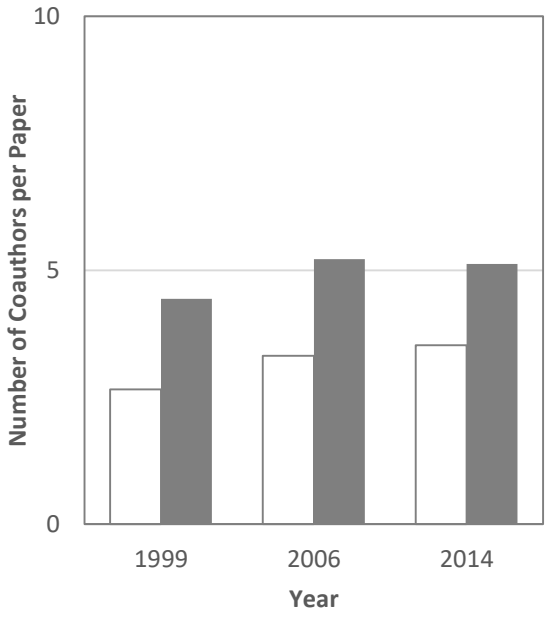
Figure 3-7

Per-faculty Total Publications by College

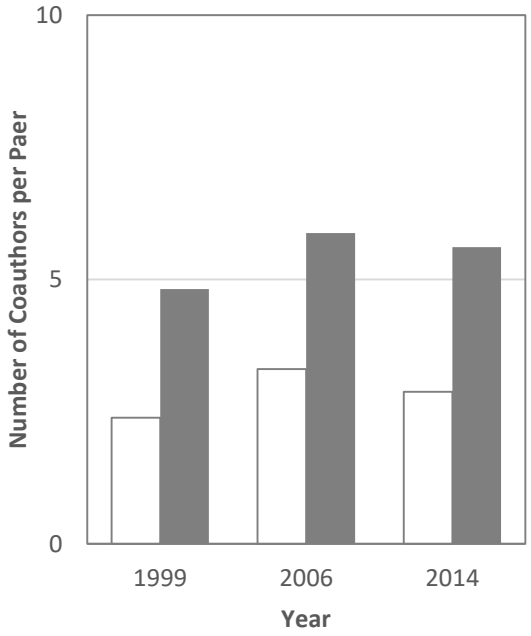
Panel A: Biological Sciences



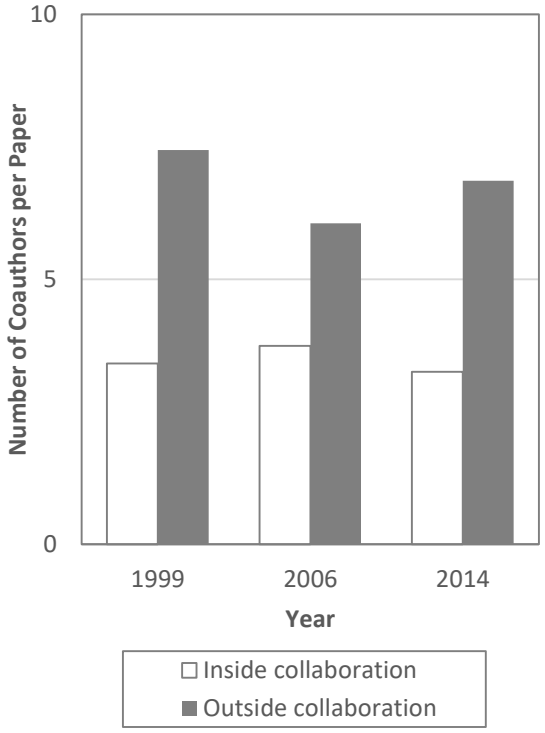
Panel B: Food, Ag, and Natural Resource Sciences



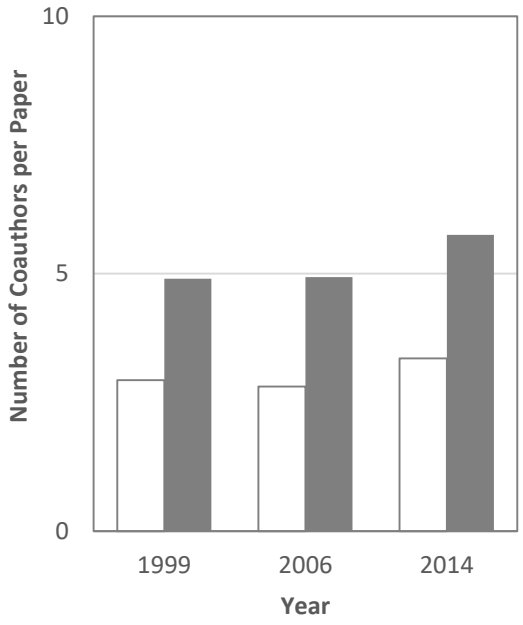
Panel C: Medical



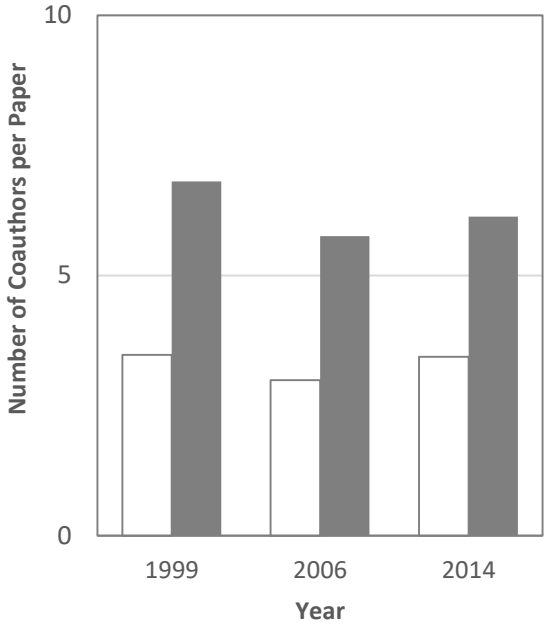
Panel D: Public Health



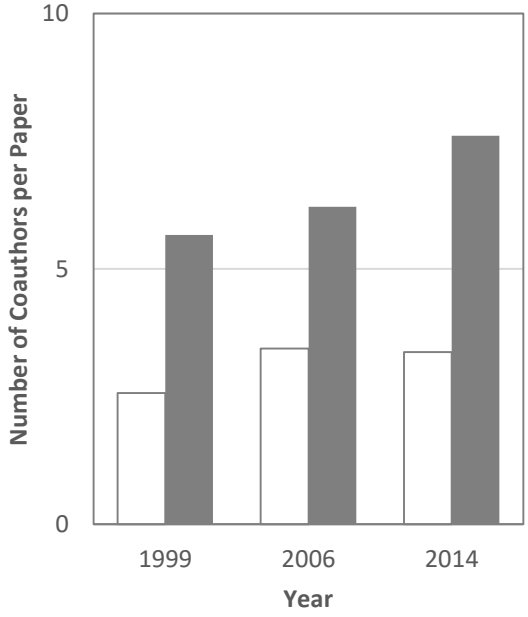
Panel E: Dentistry



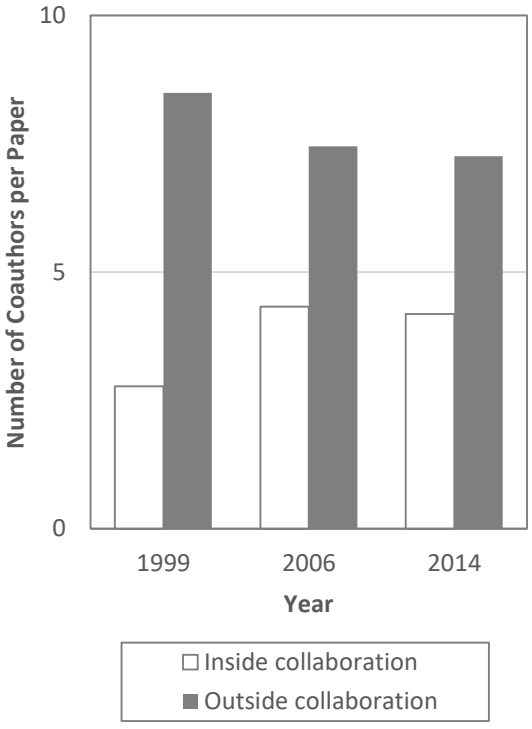
Panel F: Nursing



Panel G: Pharmacy



Panel H: Veterinary

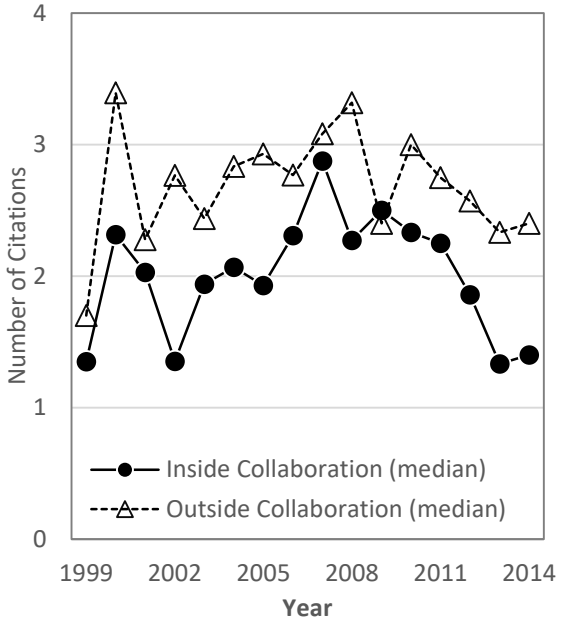


Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

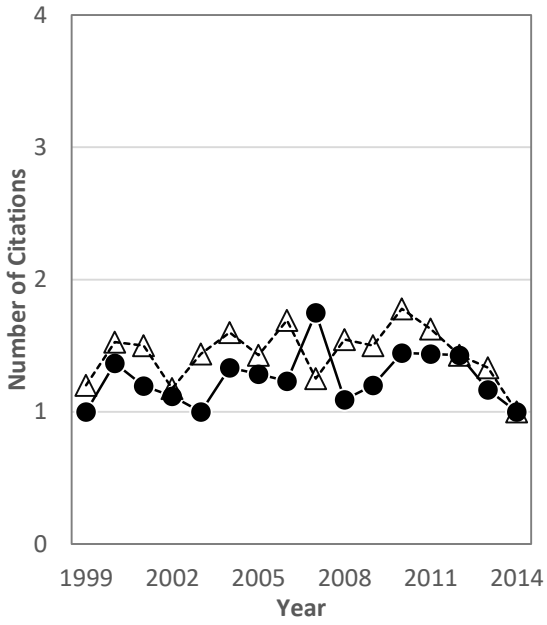
Figure 3-8

Number of Coauthors by College

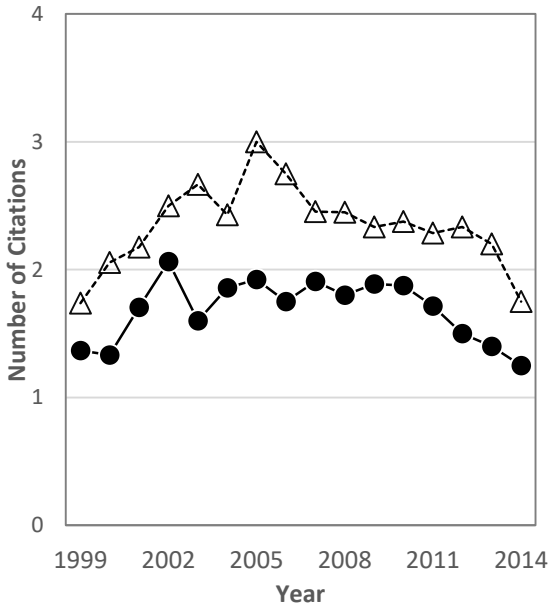
Panel A: Biological Sciences



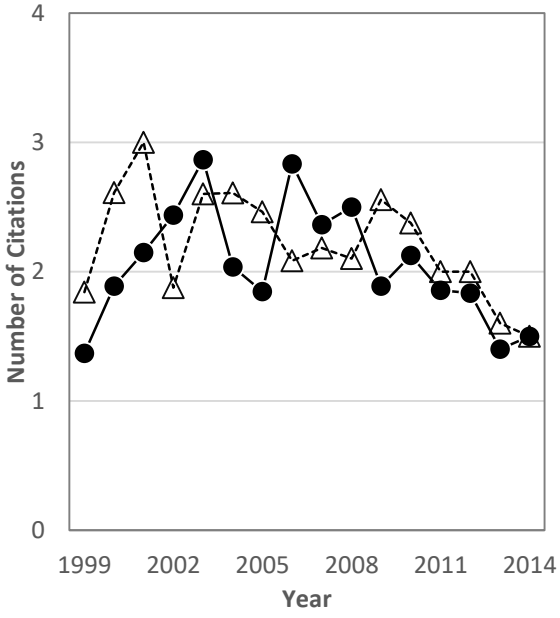
Panel B: Food, Ag, and Natural Resource Sciences



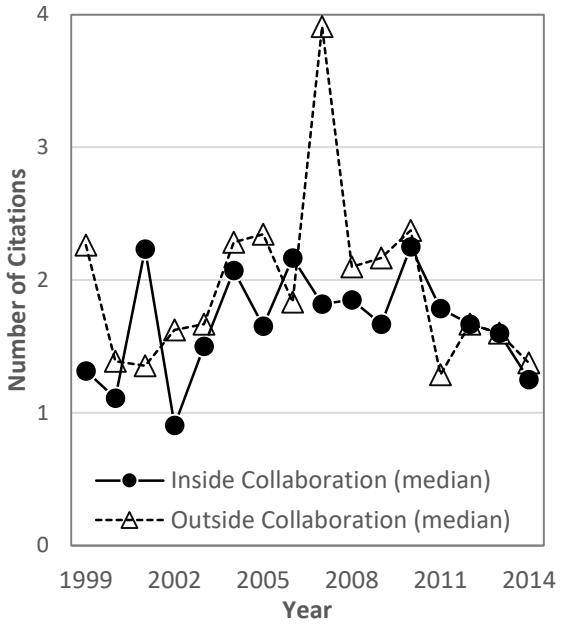
Panel C: Medical



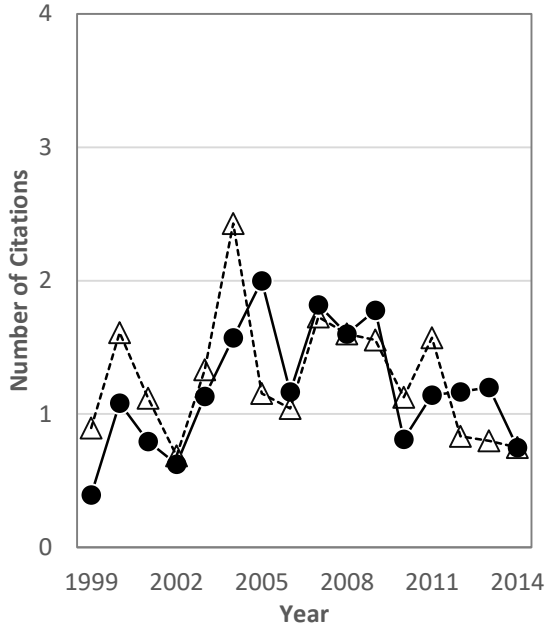
Panel D: Public Health



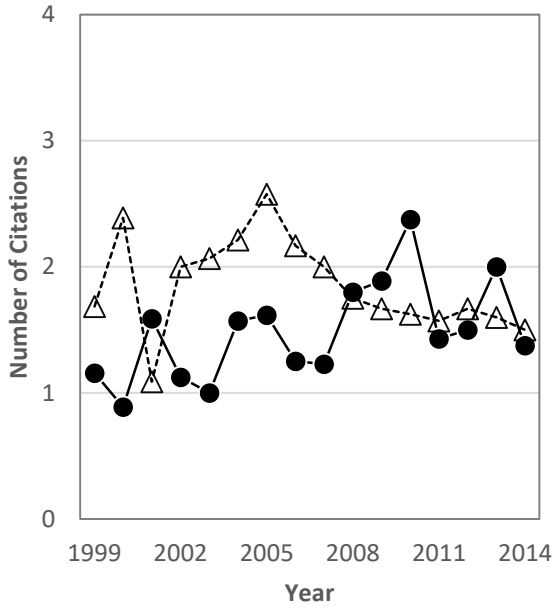
Panel E: Dentistry



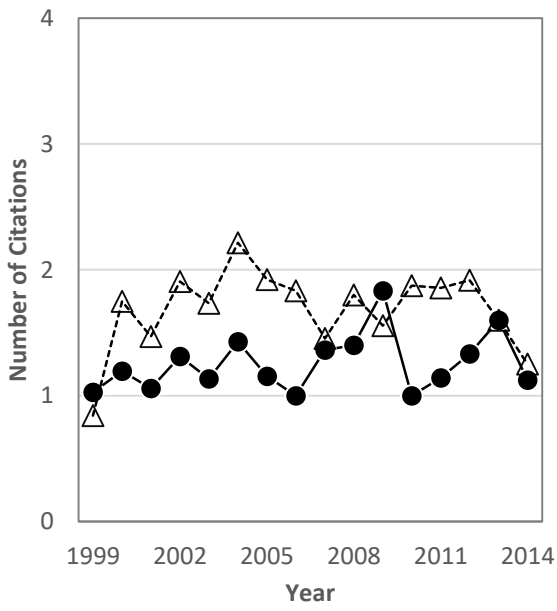
Panel F: Nursing



Panel G: Pharmacy



Panel H: Veterinary



Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

Note: The figure shows the median values for forward citations of inside collaborations and outside collaborations. Year indicates the year of publication of the cited paper.

Figure 3-9

Number of Citations by College

Appendix B

B.1 Number of Original Faculty in Each College

Table B-1

Number of Original Faculty Members in Each College with Scopus IDs

College Name		Faculty	Regular Faculty		Has Scopus IDs for 1999-2014 period	
		(1)	(2)	(2)/(1) %	(3)	(3)/(2) %
College of Biological Sciences (CBS)		273	193	71%	179	93%
College of Food, Agriculture & Natural Resources (CFANS)		752	527	70%	404	77%
Academic Health Center	School of Dentistry	604	183	30%	127	69%
	Medical School, Twin Cities	6,741	3,968	59%	2,717	68%
	Medical School, Duluth	215	105	49%	65	62%
	School of Nursing	432	102	24%	71	70%
	College of Pharmacy	752	334	44%	161	48%
	School of Public Health	422	266	63%	222	84%
	College of Vet. Medicine	419	275	66%	215	78%
Total		10,164	5,896		4,161	

Source: Developed by author using data from UMN Data Warehouse and own calculations.

Note: Faculty in column (1) represents the number of faculty whose employee group identifiers indicates them as faculty and title associated with the job code corresponds to full-time equivalent faculty (no adjuncts, clinical, teaching, and visiting). Among these, we removed faculty who have non-regular appointments with such as supplemental, clinical, visiting, and temporary less than three years in column (2). See the data documentation for the details. Column (3) indicates the number of faculty members who have Scopus author IDs from 1999 to 2014, indicating that they have at least one publication (not limited to articles) during that time. In this column, faculty members who do not have any publications in 1999-2014 are excluded. Total number of faculty at the bottom row represents the number of faculty without duplicates.

B.2 Summary Statistics at College-level

Table B-2 Individual Summary Statistics by College

Panel A: Mean

	CBS	CFANS	MED	PUB	DEN	NUR	PHR	VET
Numbers of								
Publications	2.9	2.8	3.2	4.8	2.6	2.7	3.1	3.3
Coauthors	12.1	9.5	15.7	20.8	9.7	8.2	11.6	13.5
UMN Coauthors	5.3	4.7	7.8	8.9	5.6	5.6	6.5	7.5
Authors per Paper	5.5	4.7	6.8	6.1	5.4	4.5	5.5	5.9
Citations per Paper	4.4	2.4	3.8	3.7	2.7	2.1	2.4	2.2
Professional Age	23.2	20.4	19.9	20.0	21.0	17.0	21.3	21.3

Panel B: Median

	CBS	CFANS	MED	PUB	DEN	NUR	PHR	VET
Numbers of								
Publications	2	2	2	3	2	2	2	3
Coauthors	7	6	10	13	6	6	9	10
UMN Coauthors	4	4	6	7	4	4	5	6
Authors per Paper	5.0	4.0	6.0	5.5	5.0	4.2	5.0	5.4
Citations per Paper	2.9	1.5	2.4	2.4	1.8	1.2	1.8	1.7
Professional Age	22	20	19	19	21	17	20	21

Panel C: Standard Deviation

	CBS	CFANS	MED	PUB	DEN	NUR	PHR	VET
Numbers of								
Publications	2.3	2.5	3.1	4.3	2.2	2.1	2.3	2.7
Coauthors	15.7	12.6	18.6	26.5	10.4	7.6	10.5	13.0
UMN Coauthors	5.0	4.2	7.4	7.6	5.0	4.8	5.3	6.2
Authors per Paper	3.9	3.0	4.7	3.5	4.3	2.6	2.5	2.8
Citations per Paper	5.2	3.1	6.1	6.7	3.1	3.9	2.5	2.1
Professional Age	9.5	9.7	10.8	10.1	9.8	9.1	11.5	9.9

Source: Developed by author using data from Scopus, UMN Data Warehouse and own calculations.

Note: The number of citations per paper is standardized. MED stands for Medical School, PUB for School of Public Health, DEN for School of Dentistry, NUR for School of Nursing, PHR for College of Pharmacy, and VET for College of Veterinary Medicine.

4 Assessing the Propensity to Collaborate in Life Sciences Research

4.1 Introduction

Selecting a coauthor is, in principle, a complex and multidimensional process. However, empirical evidence suggests that selecting a coauthor is, in fact, a matching process in which individuals actually depend on a small set of meaningful and often similar characteristics of potential partners (Azoulay et al. 2017; McPherson et al. 1992 and 2001). The challenge in this matching process is that it is impossible to know all the characteristics of potential coauthors before the match actually takes place, and therefore search costs and friction are inevitable (Boudreau et al. 2017; Goyal 2007). A large body of evidence indicates these costs can be reduced via colocation and face-to-face encounters (Boudreau et al. 2017; Freeman et al. 2014; Catalini 2018; Azoulay et al. 2017; Kabo et al. 2014) which became less costly due to recent developments in communication and transportation technologies (Agrawal and Goldfarb 2008; Catalini et al. 2016).[31] However, much less is known about the factors contributing to team formation among co-located researchers.

In this paper, we examine what information researchers use in selecting research partners from among a large pool of potential collaborators. In particular, we examine three particular types of information signals in selecting potential coauthors: perceived research productivity, knowledge complementarity, and professional familiarity. To answer these questions, we use academic publication data published by life sciences faculty at the University of Minnesota (UMN) from 2010 to 2014, focusing on papers in which all of the authors are affiliated with UMN. We choose this scope because we expect search friction to be less pervasive than in collaborations that transcend institutional boundaries, allowing for cleaner investigation of the role of research productivity and complementary knowledge. To identify causal relationships given the many potential unobserved factors that may influence researchers' decision to collaborate, we control for pairwise fixed effects and add

[31] Freeman et al. (2014) found that the development of communication and transportation technology cannot fully substitute for the role of geographical proximity.

time-varying control variables to control for endogeneity.[32]

We found that perceived research productivity is positively associated with the initiation of first collaboration while complementary knowledge is not. For continuing collaborations, however, knowledge complementarity becomes significant as well as the recency of first collaboration and the frequency of past collaborations. While complementarity is noted as a key driver of repetitive collaborations, researchers are more likely to collaborate with coauthors who have moderate differences in research profile. This suggests that the factors researchers consider when choosing coauthors vary with respect to the types of collaboration.

This study contributes to the empirical literature on research team formation in a number of ways. First, the findings on complementarity is consistent with the view that search costs and friction would be lower for researchers with similar research interests (Boudeau et al. 2017). Given the context of this study, which focuses on collaboration among researchers within one institution, the results imply that individual researchers are more likely to collaborate with coauthors from the same college or department that share similar research interests. While collaboration in one institutional setting is not new (Mairesse and Turner 2005; Catalini 2018; Dhand et al. 2016; Boudeau et al. 2017), we specifically examine the role of complementarity in initiating and continuing collaborations.

To investigate the role of knowledge complementary, we use a novel method that uses journals' subject categories in publication references to compute knowledge attributes and, ultimately, complementarity. By using a richer set of data (i.e., references), we get a more complete and more precise measure of a researcher's knowledge profile than if we used only one subject category given for a publication (i.e., journal title). Although a similar approach has been used to quantify the interdisciplinarity of a paper (see, for example, Leydesdorff and Rafols 2011; Porter and Rafols 2009; Rafols and Meyer 2010) or a laboratory (Catalini 2018), we focus on quantifying researcher knowledge and inter-

[32] Lee and Bozeman (2005), Fafchamps et al. (2010), Ductor (2014), Bourdeau et al. (2017) and Catalini (2018) have also addressed the issue of causal relationship in collaboration.

researcher differences.

Our final contribution is to analyze not only the decision to initiate a collaboration, but also the subsequent decision to continue collaborating, which is relatively new. Fafchamps et al. (2010) studied the formation of subsequent collaborations, but their data was limited to collaboration in economics and did not address the benefits and costs associated with the first collaboration. This paper examines the quality of past collaboration and the frequency of collaboration among different fields in the life sciences, which help us learn more about the unique characteristics of subsequent collaborations.

4.2 Data

4.2.1 Sources

We begin with the data from the previous chapter, but filter it for focus, computational tractability, and further outlier cleaning. (For more details, please refer to the previous chapter, section 3.3.1, and the beginning of section 3.4.2.) First, in this chapter, we restrict our attention to multi-author papers for which all the authors have an affiliation with UMN. Doing so enables U.S. to focus our analytical attention more credibly on the role of productivity and complementarity in determining collaboration. Second, we focus on the latest-available five years (from 2010 to 2014) for tractability; using all sixteen years of data would result in a total of more than 2 billion pairwise collaboration possibilities. Finally, we eliminate outliers for non-life sciences faculty in an analogous way to our treatment of life sciences outliers.[33] The effects of outlier filtering for non-life sciences

[33] See section 3.3.3 for details on dealing with extreme outliers for the life sciences faculty. In addition to the source of outliers explained in section 3.3.3, non-life sciences faculty have two additional attributes that could generate problematic outliers. One possibility is errors that arise from name ambiguity. Although Scopus provides a unique ID for each author, a few cases have multiple authors that are grouped together as one author−usually with Asian names. For example, for an author named Li Li, we identified hundreds of papers in a year from more than a dozen different affiliations around the world. In addition, sometimes, authors have different first names but the same initials; for example, all the JH Lee names are grouped together. Another possibility is when these various measured attributes cases occur at the same time. For example, a physicist in Germany is very productive with a large number of papers and coauthors, but he has one exceptional paper with thousands of coauthors, and is sometimes grouped with other authors because his name is fairly common. We were unable to find any other tractable way to clean these data, and, thus, eliminated them from our dataset. Freeman and Huang (2015) eliminated all names with a large number of papers as a robustness check,

faculty are summarized in Panel A (before) and Panel B (after) of Table 4-1. Before filtering, some non-life sciences faculty had more than five-thousand coauthors in a particular year or authored a paper with more than two thousand coauthors. Filtering removes 73 non-life sciences faculty from our dataset (i.e., approximately 1.4 percent of the total of 5,398 non-life sciences faculty in our dataset), who are eliminated entirely from our sample because we construct all possible researcher pairs.

[Table 4-1: Summary Statistics for Non-Life Sciences Faculty]

The final dataset includes 4,498 papers written by 1,715 UMN life sciences faculty and 5,325 coauthors (altogether 7,040 authors) from 2010 to 2014. In our analysis, the unit of observation is a dyad of researchers and we construct balanced panel data by calculating all possible pairs of the 7,040 authors. The final dataset includes 24,777,280 pairs each year (calculated by ${}_{7040}C_2$) and a total of 123,886,400 pairs for the five years.

4.2.2 Variables

For the 7,040 authors examined in this study, we compute variables measuring each researchers' research productivity, professional networks, professional rank, knowledge profile, and other characteristics that may affect the likelihood of coauthoring. For the first set of variables, we calculate lagged three-year moving averages (t-3 to t-1) to measure pre-publication characteristics. Although our panel data are only for 2010 to 2014, we use the data from previous periods to calculate the three-year moving averages.

Number of Publications. To measure productivity, we count the unique number of articles published by a researcher in the past three years (t-3 to t-1) using unique publication IDs in Scopus. Unlike prior work focused on economic researchers (Fafchamps et al. 2010; Ductor 2014; Bosquet and Combes 2013) we use raw counts: we do not discount the number of publications by the number of authors in a paper and adjust the productivity by the number of pages and the journals' quality index.[34] We do so for two reasons. First, our

and found similar results to their main analysis.

[34] Fafchamps et al. (2010) and Ductor (2014) calculated the researcher productivity as follows:

study encompasses more than a dozen different fields with publishing norms (i.e., in which journal to publish, and rules on coauthorship) that differ substantially across disciplines. Second, a reliable journal quality index across all life science journals does not exist, and the number of pages and range of impact factors vary substantially across disciplines. For example, the median impact factor for journals published in the "medicine, research, and experimental" category is 2.460 in 2014 from the Journal Citation Reports (JCR) 2014 of Thomson Reuters, while for economics, the median impact factor is 0.866. Thus, adjusting the number of papers by the number of pages, journal quality, and number of authors in a paper does not necessarily yield a better measure of productivity.

Number of (distinct) coauthors. We count the number of (distinct) coauthors in all articles published in the past three years (t-3 to t-1) using unique author IDs in Scopus.[35] The resulting counts represent the size of each researcher's coauthorship network (Bosquet and Combes 2013).

Number of UMN (distinct) coauthors. Among a researcher's coauthors, we identify those with a UMN affiliation for the past three years (t-3 to t-1). For these coauthors, we also use the same affiliation cleaning mechanism as in the previous chapter. This variable represents the local (within UMN) network size of a researcher.

Average number of authors per paper. This variable measures the average team size that a researcher has for collaboration in the past three years (t-3 to t-1). To calculate this, we first count the number of authors in each paper and estimate a per paper average for the past three years. Since we exclude sole-authored papers in our study, the minimum number of authors per paper is two.

$$\text{productivity} = \sum_{i}^{S} \frac{Number\ of\ Pages_i \times Journal\ quatilty\ index_i}{Number\ of\ authors_i}$$

when i represents an article and S is the total number of articles published by an author in a given period. Bosquet and Combes (2013) took a similar approach to calculate productivity.

[35] Using unique author IDs in Scopus results in very few errors in which authors with similar names are erroneously grouped together. Zeng et al. (2016) used a string of the last name and first name initials instead of author IDs to count the number of distinct coauthors. As they pointed out, this would consider "Jane Linda Smith" and "Jane Smith" as different names. For our analysis, we do not have these issues, so the error rate is substantially lower.

Average annual citation rate. The average annual citation rate for researcher i at year t, c_{it}, is defined as follows:

$$c_{it} = \frac{1}{3}\sum_{s=t-3}^{t-1}\frac{1}{n_{is}}\left(\sum_{j=1}^{n_{is}}\frac{Number\ of\ Citation_{ijs}}{2019-s}\right)$$

for researcher i's paper j $(1 \leq \mathrm{j} \leq n_s)$ published in year s $(\mathrm{t}-3 \leq \mathrm{s} \leq \mathrm{t}-1)$. For each paper published in year t, we first obtain the total number of forward citations in 2019 as listed in Scopus. We divide them by the number of years since publication to obtain the per year average number of citations for each paper.[36] Then, we take the yearly average if there is more than one collaboration in a given year. Finally, we compute the three-year moving average. Self-citations are not excluded. Delays between the time of publication and citation accumulation (Ductor 2014) are of limited concern because all papers in our dataset have at least three years of accumulated citations. In addition, we use the average annual citation rate instead of total citations to control for variations across disciplines for the peak of annual citations, time needed to reach a peak, and annual citation values after the peak. (Galiani and Galvez 2017).

We use the average annual citation rate as a proxy for research quality, appealing to prior work as precedent (Bidault and Hildebrand 2014; Chung et al. 2009). Although the number of citations increases as the number of authors in a paper increases (Wuchty et al. 2007; Chung et al. 2009), citation data are still a useful proxy for quality of publication. The number of citations reflects the market value in the academic community (Chung et al. 2009), and, thus, is closely intertwined with decisions regarding recruitment, promotion, and remuneration as well as researchers' scientific reputation and networking capabilities (Hilmer and Hilmer 2005; Laband and Piette 1994; Ductor et al. 2014).

Professional Age. By definition, a researcher has professional age of one when his or her first paper is published (i.e., the publication year of an article minus the year of the researcher's first publication plus one). To calculate this professional age, we find the first paper published for both life sciences faculty and non-life sciences faculty from Scopus.

[36] Chung et al. (2009) also adjusted the number of citations by the number of years since publication.

This definition is widely used in prior work (Hampton and Parker 2011; Lariviere et al. 2016; Bu et al. 2017b).

Knowledge attributes vector. Because commonality of research interest is an important factor in determining collaboration (Fafchamps et al. 2010), we construct a vector measure of each researcher's knowledge and use it to quantify differences between two researchers' interests. We construct that vector using subject categories of the publications a researcher cites, which contrasts with prior approaches relying on classifications of a researcher's own output (Fafchamps et al. 2010; Ductor 2014; Azoulay et al. 2017; Bu et al. 2017a; Tang et al. 2008; see also Jaffe 1986 in the patenting literature).

Using cited references to estimate the knowledge attributes of each researcher has many advantages. Cited references provide an indicator of the authors fields of study by way of "the shoulders" of previous researchers on which their research stands (Freeman and Huang 2015). Specifically, the subject categories of cited references have been widely used to analyze if a publication is interdisciplinary or not (Leydesdorff and Rafols 2011; Porter and Rafols 2009; Rafols and Meyer 2010) by using the diversity measures defined by Stirling (2007) and Rao (1982).[37] In this paper, however, we use the subject categories of cited references to reveal the research diversity of each researcher, or knowledge attributes, instead of whether or not a paper is interdisciplinary. In addition, having multiple categories to classify a paper (rather than a single category associated with the journal in which the paper is published, for example) allows for a more refined measure of a researcher's overall knowledge profile. This will, in turn, improve our measures of inter-researcher distances.

[37] Stirling (2007) proposed a formula for diversity D:

$$\mathrm{D} = \sum_{i \neq j} p_i \, p_j d_{ij}$$

when p_i and p_j are the probability distribution of elements i and j in the system and d_{ij} is the degree of difference between i and j. Using this and the initially introduced derivation in Rao (1982), Porter and Rafols (2009) defined a Rao-Stirling diversity measure as $\mathrm{I} = 1 - \sum_{i,j} s_{ij} p_i p_j$ when p_i is the proportion of references citing the subject categories, i, in a given paper. s_{ij} is the cosine measure of similarity between subject categories i and j. This Rao-Stirling diversity index takes into account variety, balance, and distance (or disparity) between categories when defining the diversity of disciplines in interdisciplinary research.

Formally, we define knowledge attributes and inter-researcher knowledge differences as follows. Given K subject categories, individual researcher i's knowledge attributes[38] at period t are defined using a K-dimensional vector whose **k**th element is

$$s_{itk} = \frac{\sum_{j \in A_t} r_{ijk}}{\sum_{j \in A_t} R_{ij}}$$

where A_t = {j: paper j was published in year y where $t - b \leq y \leq t - 1$} for some choice of b (for this analysis, it is given as three), and r_{ijk} is the number of references in category k that paper j cites k $(1 \leq k \leq 218)$. When there are multiple subject categories for a given journal, r_{ijk} can be a fraction where each category is assumed to have an equal weight within a journal.[39] Finally, $R_{ij} = \sum_k r_{ijk}$ is the total number of references that paper j cites. Information on subject categories is obtained from the Journal Citation Reports (JCR) 2014 of Thomson Reuters. Among all the references that we collected, we only considered articles included in either the Science Citation Index (SCI) or the Social Sciences Citation Index (SSCI). The journal titles in the references provided in Scopus are not cleaned, so we cleaned them to improve matching with the lists of SSCI and SCI.[40] If the cleaned journal names do not match the lists of SCI and SSCI, we exclude them from our analysis. Given the fact that some of the subject categories can be highly correlated with each other, we use the Mahalanobis distance to calculate the dyadic distance, which uses the variance-covariance matrix to orthnormalize the vectors.

For the analyses on pairs who have collaborated in the past, we add variables for the outcome of past collaboration. These include the average annual citation rate from past collaboration, time since the first publication, and number of skipped years without collaboration. The average annual citation rate from past collaboration is computed as a

[38] Notably, the knowledge attributes being measured here are not the same as knowledge attributes discussed in the business literature. From a knowledge management view, knowledge attributes are defined as a dimension where there is a range of values, types (descriptive, procedural, or reasoning) and categories (tacit or explicit) (Holsapple 2004).

[39] To give an example, *Review of Economics and Statistics* is given two subject categories, "economics" and "social sciences – mathematical methods", and, therefore, r_{ijk} is 1/2 for each category.

[40] See the Appendix for the entire list of subject categories in the JCR.

three-year moving average (t-3 to t-1).

Average annual citation rate from past collaboration. The average annual citation rate from past collaboration for researcher *i* at year *t*, cp_{it}, is defined as follows:

$$cp_{it} = \frac{1}{3}\sum_{s=t-3}^{t-1}\frac{1}{n_{is}}\left(\sum_{j=1}^{n_{is}}\frac{Number\ of\ Citation_{ijs}}{2019-s}\right)$$

where for researcher *i*'s paper *j* ($1 \leq j \leq n_s$) published in year *s* ($t-3 \leq s \leq t-1$), we first compute the per year average of citations for each paper. Then, we take the yearly average if there is more than one collaboration in a given year, and, finally form a three-year moving average.

Time since the first publication. This variable measures the time that has elapsed since the first collaboration from 1999 to 2014.

Number of skipped years without collaboration. This variable measures whether the researchers have repeatedly, or persistently collaborated for the past years (Bu et al. 2017b). We calculate the number of years a pair of researchers has no collaboration after the first collaboration in the period 1999-2014.[41]

4.2.3 Summary Statistics

Table 4-2 provides summary statistics for 122,615,527 pairs. Among pairs of researchers in the life sciences at UMN, each researcher averages 4.2 papers per year, with the average difference in output between researchers averaging 5.6 papers per year. All the variables are right-skewed even after eliminating outliers. (Figure 4-1).

[Table 4-2: Summary Statistics of All Pairs]

[Figure 4-1: Kernel Density Estimation of the Differences in Knowledge Attributes]

When examining factors influencing initial collaboration, we first parse out 24,063

[41] See the Appendix for an example of these concepts.

pairs of researchers who have their first joint publication from 2010 to 2014. For continuing collaborations, we limit our data to those who have at least one joint publication prior to 2014, yielding 26,962 pairs of researchers.[42] In both datasets, we eliminate pairs that never collaborate.

Researchers that collaborate for the first time have more publications and numbers of coauthors on average and greater differences in those variables than pairs who do not yet collaborate (Table 4-3). Differences in knowledge attributes, however, are greater for pairs who do not collaborate, largely due to a long tail in its distribution (Figure 4-2). Researchers that collaborate again also have less differences in knowledge complementarily but greater differences in average number of publications and citations pairs who do not collaborate again (Table 4-4). Pairs who collaborate more than once have smaller differences in knowledge attributes than pairs who collaborate for the first time, indicating that a larger overlap in knowledge attributes than pairs who do not collaborate more than once (Figure 4-3).

[Table 4-3: Summary Statistics for First-time Collaboration]

[Figure 4-2: Kernel Density Estimation of the Differences in Knowledge Attributes for the First Collaboration]

[Table 4-4: Summary Statistics for Continuing Collaborations]

[Figure 4-3: Kernel Density Estimation of the Differences in Knowledge Attributes for Continuing Collaborations]

4.3 Estimation Strategy

We estimate a series of models at the researcher pair by year level to investigate the role that researcher attributes play in determining collaboration. Our first set of models focuses on why researchers collaborate for the first time. In particular, using the sample of pairs

[42] Pairs who collaborate for the first time in 2014 are eliminated because we cannot examine the likelihood of continuing collaboration after the first collaboration for these pairs.

who have not collaborated as of a given year, we estimate the following linear probability model:

$$Y_{\mathrm{ijt}} = \alpha + \beta \boldsymbol{DIST}_{ijt} + \gamma X_{ijt} + \theta_{ij} + \delta_t + e_{ijt} \quad \text{--- (1)}$$

where Y_{ijt} is a dummy variable for whether pair *ij* collaborate in a given year *t*. The vector $\boldsymbol{DIST}_{ijt}$ includes measures of differences in knowledge attributes, productivity, network size, quality of research, and professional age. We add a quadratic term of difference in knowledge attributes here, too, because collaboration is unlikely to occur when researchers lack any common research interests or have identical research interests, so that in either case, there would be no complementary skills to exchange (Fafchamps et al. 2010). Our vector of pair-level controls, X_{ij} , contains the average values within a pair for the variables included in $\boldsymbol{DIST}_{ijt}$, except for knowledge attributes. Both the differences of interest and control variables reflect three-year moving averages for the past three years, from t-3 to t-1. We also introduce several dummy variables to flexibly account for other determinants of collaboration: an indicator variable reflecting whether both members of the pair are life sciences faculty, plus pair effects, θ_{ij}, and year effects, δ_t. All standard errors are clustered at the researcher pair level.

We focus on a linear model because nonlinear alternatives tend to underestimate the probability of non-zeroes when there are excessive zeroes in the dataset (King and Zheng 2001). However, for robustness we also estimate non-linear models; the resulting marginal effects are very similar to those from our linear model.[43]

Estimating equation (1) requires some care. Selecting a co-author is a complex process, and we may not observe all the characteristics a researcher considers. Therefore, we employ a pair fixed effect, which controls for all time-invariant factors that may affect the propensity to collaborate, and introduce time-varying controls such as the average number of publications, coauthors, average team size, average annual citation rate, and age.[44] Further, the model is inherently dynamic: past collaboration decisions affect several

[43] See the Appendix for the results of the probit and logit estimation.

[44] This approach is similar to Fafchamps et al. (2010). However, they used fixed effects logit instead of GMM.

right hand side variables, such that regressors are predetermined and not strictly exogenous. For these reasons, both pooled ordinary least squares (OLS) and standard fixed effects models are likely to be biased and inconsistent. Thus, we use a generalized method of moments (GMM) estimator by Arellano and Bond (1991).[45] The Arellano and Bond estimator first takes the first difference of all the variables and eliminates the pair effects and time-invariant variables. Then it uses lagged levels of the endogenous variables as instruments in a GMM procedure.[46] We evaluate the excludability of these instruments using standard Sargan/Hansen tests of over-identifying restrictions.

We augment this model when examining how researcher attributes affect the likelihood of repeat collaboration. Once researchers have collaborated, they have more information about each other as potential coauthors, and also incur different benefits and costs of collaboration compared with their first collaboration. For example, researchers might encounter new attributes of coauthors that they had not considered up front. Alternatively, benefits may increase from a much finer division of labor as researchers have more information about their coauthors' expertise. Finally, costs could decrease due to lower coordination and communication costs.

To examine these different incentives for subsequent collaborations, we estimate the following model:

$$Y_{\text{ijt}} = \alpha + \beta \boldsymbol{DIST}_{ijt} + \gamma X_{ij} + \delta Z_{ij} + \theta_{ij} + \delta_t + e_{ijt}. \qquad \text{--- (2)}$$

This model is identical to (1) except, notably, for the addition of a vector of variables Z_{ij} that measures the quality and persistency of past collaborations between researcher *i* and *j*. Specifically, this includes the average annual citation rate from past collaborations, time since first co-authored publication, and number of years since that first collaboration without a co-authored publication (i.e., "skipped years"). We include the average annual citation rate from past collaborations as a proxy for the expected quality of the next collaboration; the number of citations a paper receives is typically considered a measure of

[45] We used Roodman (2009)'s Stata command for the GMM estimations.

[46] The moment condition for this estimator is:

$$\mathrm{E}\left(y_{ij,t-\rho}\Delta e_{ijt}\right) = 0 \; for \; all \; \rho = 2, \ldots, t-1$$

a paper's quality (Bidault and Hildebrand 2014).[47] Time since the first collaboration indicates the "age" of partnership between researchers, and as the age of the partnership increases, the less the returns of coauthorship for both researchers, and, thus, the likelihood of collaborating again (Bidault and Hildebrand 2014).[48] The number of skipped years without collaboration measures whether or not collaboration between researchers has been frequent through time.

4.4 Results

4.4.1 First-time collaborations

The results in Table 4-5 suggest that researchers initiate a collaboration with others who have similar knowledge attributes. The differences in knowledge attributes do not have an Inverted-U shape and the likelihood of forming a collaboration decreases when the differences in knowledge attributes increases.[49] Thus the similar skills and knowledge are perceived as benefits of first-time collaborations for the researchers at UMN.

[Table 4-5: Estimation Results for the First-time Collaborations]

Benefits of collaboration seem higher for researchers who are highly productive in terms of the number of publications but not the number of citations. If the average number of publications between two researchers increases by one, the likelihood of collaboration increases by 27 percent, holding other variables constant. For citations, however, researchers are more likely to initiate collaboration when the pair-average citation rates are smaller and the differences are greater. This suggests that benefits of first-time collaboration with coauthors at the same institution is lower for researchers who are already well-established. These researchers, who have higher citations rates, would be more widely

[47] Uzzi et al. (2013) and Lee et al. (2015) measured creativity in a different way by calculating the rareness of the references' pairwise combination.

[48] Bidault and Hildebrand (2014) also noted that even if the time since the first publication is "aged", the returns from coauthorship can be positive if the researchers have collaborated repeatedly over the years.

[49] The likelihood of forming a new collaboration is lowest when the difference in knowledge attributes is 59.5 We used four decimal points of the coefficients of knowledge attributes to calculate the lowest point, which was 0.0001 for the quadratic term, and 0.0119 for the first-order term.

known and better connected in the network of academia than those with less citations, and thus, have more opportunities to collaborate coauthors from other institutions.

In our analysis, the numbers of (all and UMN affiliated) coauthors and team size do not have significant effects on the probability of forming a new collaboration. When the knowledge complementarity and research productivity are controlled, the network and team size do not have much explanatory power.

Lastly, Table 4-5 also presents the Hansen test of overidentifying restrictions and the tests for the first order and second order serial correlation tests of the first-differenced residuals. Our results do not reject the null hypothesis of the Hansen test, which is that the instruments are valid. We find no evidence of serial correlation in the residuals in levels.

4.4.2 Continuing Collaboration

In this section, we examine the propensity to collaborate again among pairs that have collaborated in the past (Table 4-6).

[Table 4-6: Estimation Results for Continuing Collaborations]

The results in Table 4-6 suggest that researchers collaborate again with others who have complementary knowledge, but choose those with moderate intellectual distance. Unlike the first-time collaborations, the differences in knowledge attributes have a shallow Inverted-U shape and researchers are more likely to collaborate as knowledge differences increase. The likelihood of continuing collaboration is highest when the difference in knowledge attributes is 52.8 (column 3) or 46.7 (column 4), which encompasses almost all the pairs of researchers in the dataset.[50] This suggests that while complementary knowledge is not perceived as benefits of collaboration to initiate a new collaboration, different skills and knowledge eventually becomes a key to collaborate subsequently.

While complementary knowledge is sought by researchers, indicators of

[50] We used four decimal points of the coefficients of knowledge attributes to calculate the highest point, which was -0.0002 (column 3) or -0.0003 (column 4) for the quadratic terms, and 0.0211 (column 3) or 0.0280 (column 4) for the first-order terms.

productivity (both the number of publications and citations) decline in importance. Citation rates from past collaborations have no statistically significant effect. These public measures of productivity may no longer be useful for researcher pairs that have collaborated before, because they already have private information on the relevant productivity and quality attributes of their past collaborators.[51]

On the other hand, the coefficients for the average number of UMN coauthors become significant and negative, indicating that researchers with greater number of UMN coauthors are less likely to collaborate again. These researchers may have more opportunities to collaborate with other researchers at the same institution, and, thus, collaborating again with the past coauthors does not seem beneficial to them.

Researchers are also less likely to continue collaborating when time since the first collaboration increases. While continuing collaboration may reduce the communication and coordination costs, the benefits may also decrease as the age of collaboration grows. Researchers may see the older collaborations less beneficial as the number of citations earned decline over time.[52]

In our analysis, repetitive collaboration with coauthors at the same institution does not seem beneficial to researchers. The results indicate that researchers are more likely to collaborate when they have more years of no collaboration. One possible reason for this negative effect of repetitive collaboration is that the number of skipped years does not truly address the years of no collaboration. Researchers may have worked on research projects even during the years of no publication because collaboration over research projects and publishing the outcome often require a longer and synchronous period of time. Another possible reason is that we have not considered the mobility of researchers. Because our dataset only deals with papers written by all authors with UMN affiliation, we do not observe continuing collaborations even after a coauthor has moved to another institution.

[51] Another possible reason for these insignificant coefficients for citations is that researchers may have started a subsequent collaboration before they can observe the ultimate number of citations because it takes considerable time to receive citations after publication.

[52] For the discussions on the life-cycle of citations see Galiani and Galvez (2017).

4.5 Conclusion

Although collaboration among researchers is generally increasing among all possible pairs of researchers, few actually work together. It is thus natural to ask what information researchers use in selecting research partners from among a large pool of potential collaborators. This study has addressed that question for UMN researchers for three particular types of information signals: perceived research productivity, knowledge complementarity, and professional familiarity. We investigate these signals by using academic publication data published by life sciences faculty at UMN from 1999 to 2014, focusing on papers in which all of the authors are affiliated with UMN.

We have found that researchers initiate a collaboration with others who have similar knowledge. Researchers also seem to find benefit in working with highly and similarly productive coauthors. However, for research pairs who have collaborated before, the pairs choose coauthors with complementary knowledge while productivity no longer remains significant. Researchers are more likely to collaborate again when the first collaboration was relatively recent and when they have worked with past collaborators less frequently. Citation rates from past collaboration do not bring any significant effects to continuing collaborations.

These findings suggest that researchers use different information when selecting their coauthors when they have worked with the coauthors in the past. When researchers collaborate for the first time, they tend to rely on information that is publicly available, such as the number of publications, and citations that potential coauthors have. However, once researchers have collaborated, this publicly available information is no longer useful, because they now have private information on the relevant productivity and quality attributes of their past collaborators. Instead, other factors, including the age of the partnership and frequency of prior outputs, become more important in deciding whether or not to continue collaboration.

These findings offer important contributions to the existing literature on team formation. First, this study is among a few studies that empirically tested the role of knowledge complementarity in research collaboration. In doing so, we employ a relatively

new way of constructing knowledge differences, by using the subject categories that are assigned to the journals in which the cited references in the publications are published. This helps us quantify researcher knowledge attributes and inter-researcher differences more precisely. Second, the findings on complementarity corroborate that search costs and friction appear to be substantially lower for researchers with similar research interests (Boudeau et al. 2017). By using data that encompass larger groups of researchers for a longer time span, our analysis implies that that individual researchers are more likely to start collaborating with coauthors from the same college or department that share similar research interests. Thus, our findings are not only of interest to the academic community but also to university administrators seeking to promote more collaboration among researchers.

Our final contribution is to analyze not only the decision to initiate a collaboration, but also the subsequent decision to continue collaborating, which is relatively new. Our work builds on Fafchamps et al. (2010), who studied the role of network effects in determining subsequent collaboration among economists. Our analysis extends the study of continuing collaboration to the life sciences, and examines the impact of past collaboration quality and frequency on the likelihood of continuation.

Table 4-1

Summary Statistics for Non-Life Sciences Faculty

Panel A. All Dataset

	Mean	Median	SD	Min	p99	Max
Number of Publications per Researcher	2.8	1.0	12.8	1.0	15.0	830.0
Number of Coauthors per Researcher	14.2	7.0	96.1	2.0	76.0	5,041.0
Average Number of Authors per Paper per Researcher	7.2	6.3	27.1	2.5	15.0	2,295.9
Average Number of Citations per Paper per Researcher	3.3	2.3	4.2	0.0	20.3	128.2

Panel B. Data Excluding Outliers

	Mean	Median	SD	Min	p99	Max
Number of Publications per Researcher	2.3	1.0	2.6	1.0	13.0	47.0
Number of Coauthors per Researcher	10.5	7.0	11.5	2.0	60.0	156.0
Average Number of Authors per Paper per Researcher	6.7	6.3	2.5	2.5	14.5	38.3
Average Number of Citations per Paper per Researcher	3.3	2.3	4.1	0.0	19.5	128.2

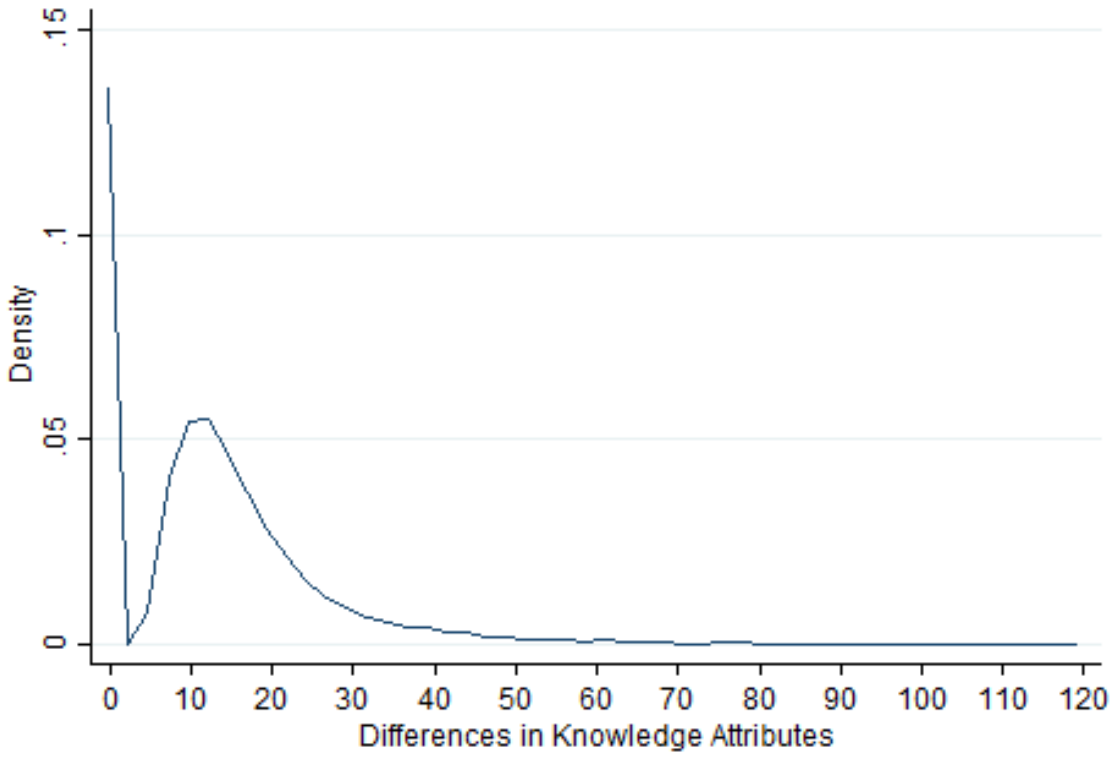
Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.

Table 4-2

Summary Statistics for All Pairs

	Mean	Median	Std.D
Differences (between a researcher i and j, 3 year moving average, t-3 to t-1)			
Knowledge Attributes	16.4	13.6	12.8
# of Publications	5.6	3.0	8.0
# of Coauthors	22.4	11.0	45.2
# of UMN Coauthors	8.8	5.0	10.9
# Authors per Paper	3.0	2.1	3.5
# of Citations per Paper	4.3	2.5	10.6
Professional Age	11.4	8.0	10.7
Control Variables (averaged between a researcher i and j, 3 year moving average, t-3 to t-1)			
Average # of Publications	4.2	2.5	4.9
Average # of Coauthors	16.6	10.0	25.4
Average # of UMN Coauthors	9.3	7.5	7.0
Average # of Authors per Paper	2.8	2.5	2.4
Average # of Citations per Paper	3.4	2.4	5.8
Average Professional Age	10.3	8.5	7.9

Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.



Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.

Note: Kernel density is derived by using Epanechnikov (1969) with the bandwidth of 0.1827; Among 123,886,400 pairs in the dataset, 9.2 percent of these pairs (11,429,907 pairs) have zero knowledge attributes, because some researchers had no publications for the period 1999-2014 or did not have a publication that cites sources from SCI or SSCI.

Figure 4-1

Kernel Density Estimation of the Differences in Knowledge Attributes

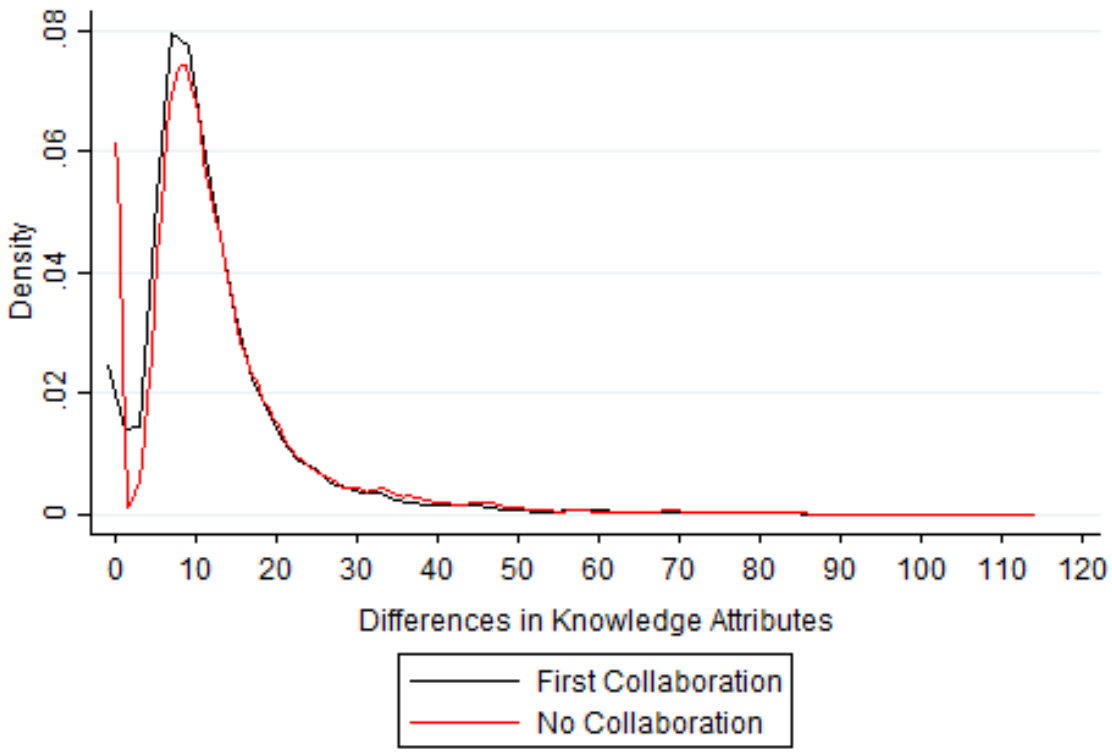
Table 4-3

Summary Statistics for First-time Collaborations

	No Collaboration	First Collaboration	Differences
Differences (between a researcher i and j)			
Knowledge Attributes	12.9	12.6	0.3**
# of Publication	7.9	8.7	-0.8**
# of Coauthors	31.9	35.3	-3.4**
# of UMN Coauthors	13.1	14.3	-1.2**
# Authors per Paper	3.8	3.9	-0.1**
# of Citation per Paper	4.8	4.7	0.1
Age	12.5	12.6	-0.1**
Control Variables (averaged between a researcher i and j)			
Average # of Publication	5.6	6.6	-1.0**
Average # of Coauthors	22.7	27.1	-4.4**
Average # of UMN Coauthors	12.4	14.2	-1.8**
Average # of Authors per Paper	3.1	3.6	-0.5**
Average # of Citation per Paper	3.7	4.1	-0.4**
Average Age	10.6	12.3	-1.7**
Number of Pairs	19,508	24,063	
Number of Observation	49,346	24,063	

Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.

Note: The equality of the mean is tested. * $p<0.10$, ** $p<0.05$.



Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.

Note: Kernel density is derived by using Epanechnikov (1969) with the bandwidth of 0.7122.

Figure 4-2

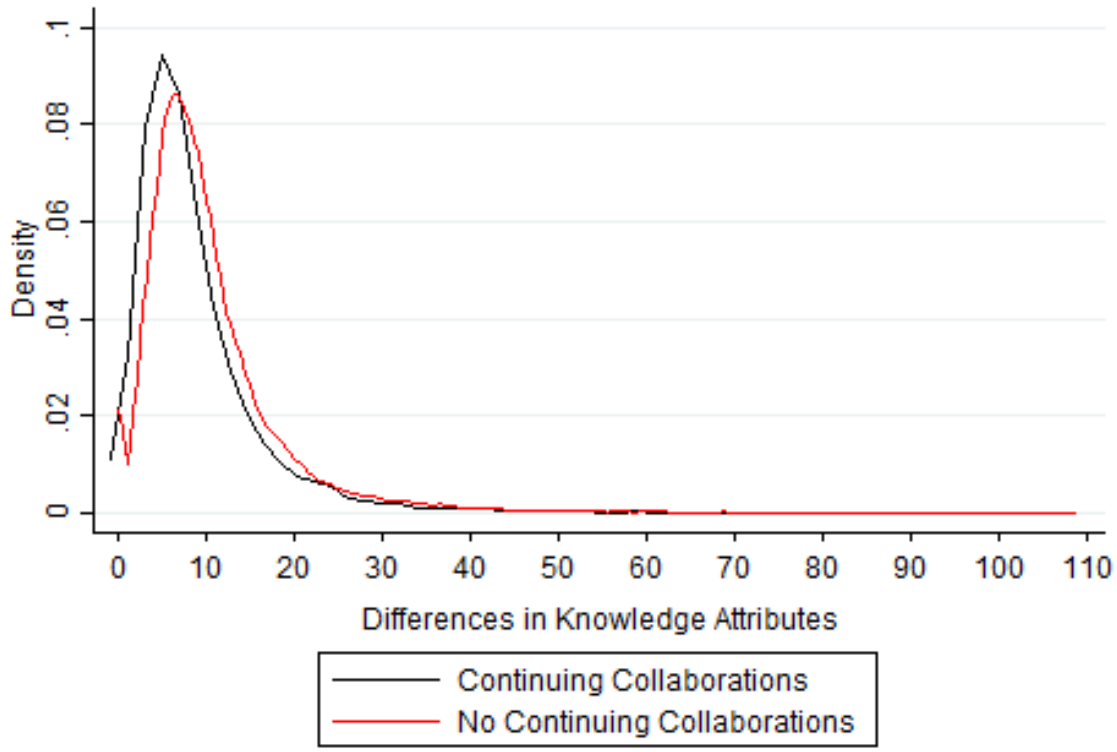
Kernel Density Estimation of the Differences in Knowledge Attributes for the First Collaboration

Table 4-4

Summary Statistics for Continuing Collaborations

	No Continuing Collaboration	Continuing Collaboration	Differences
Differences (between a researcher i and j)			
Knowledge Attributes	10.5	9.0	1.5**
# of Publication	9.8	10.9	-1.1**
# of Coauthors	39.7	38.2	1.5**
# of UMN Coauthors	15.8	16.0	-0.2
# Authors per Paper	3.1	2.8	0.3**
# of Citation per Paper	4.1	3.5	0.6**
Age	12.2	13.1	-0.9**
Control Variables (averaged between a researcher i and j)			
Average # of Publication	9.6	11.7	-2.1**
Average # of Coauthors	38.8	40.6	-1.8**
Average # of UMN Coauthors	19.7	21.7	-2.0**
Average # of Authors per Paper	5.1	5.2	-0.1**
Average # of Citation per Paper	5.2	5.5	-0.3**
Average Age	17.3	17.6	-0.3**
Past Collaboration Outcome			
Average Annual Citation Rate from Past Collaborations	3.2	3.3	-0.1**
Persistency of Collaboration (at t)			
Time since the First Collaboration	3.9	3.6	0.3**
# of Skipped Years without Collaboration	2.4	1.6	0.8**
Number of Pairs	26,052	6,731	
Number of Observation	75,984	9,792	

Note: The equality of the mean is tested. * $p<0.10$, ** $p<0.05$.



Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.

Note: Kernel density is derived by using Epanechnikov (1969) with the bandwidth of 0.7078.

Figure 4-3

Kernel Density Estimation of the Differences in Knowledge Attributes for Continuing Collaborations

Table 4-5

Estimation Results for First-time Collaboration (1)

	(1) Pooled OLS	(2) Pooled OLS All Colleges	(3) FD. GMM All Colleges
Differences			
Knowledge Attributes	0.003**	0.002**	-0.012*
	(0.000)	(0.000)	(0.006)
Knowledge Attributes^2	-0.000**	-0.000**	0.000*
	(0.000)	(0.000)	(0.000)
# of Publication	0.001**	-0.001**	-0.115**
	(0.000)	(0.001)	(0.049)
# of Coauthors	0.000	0.000**	0.002
	(0.000)	(0.000)	(0.011)
# of UMN Coauthors	0.000	-0.001**	0.002
	(0.000)	(0.000)	(0.007)
Average Team Size	-0.001**	-0.011**	-0.059
	(0.000)	(0.001)	(0.040)
Average Annual Citation	-0.001**	-0.001**	0.164**
	(0.000)	(0.001)	(0.070)
Controls			
Average # of Publications		0.005**	0.270**
		(0.001)	(0.114)
Average # of Coauthors		-0.001**	-0.006
		(0.000)	(0.023)
Average # of UMN Coauthors		0.003**	-0.019
		(0.000)	(0.013)
Average Team Size		0.025**	0.130
		(0.002)	(0.082)
Average Annual Citations		0.001	-0.292**
		(0.001)	(0.129)
Both Faculty		-0.104**	
		(0.007)	
Constant	0.236**	0.131**	
	(0.006)	(0.004)	
R-Squared	0.194	0.203	
Hansen			0.451
AR(1)			0.097
AR(2)			0.151
# of Pairs			19,405
Observations	73,041	73,041	49,118

Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.

Note: Hansen is Hansen J test for overidentifying restrictions (p-values reported); AR(1) and AR(2) are tests for first and second-order serial correlation (p-values reported); Standard errors in parentheses; * p<0.10, ** p<0.05.

Table 4-6

Estimation Results for Continuing Collaborations

	(1) FD GMM	(2) FD GMM	(3) FD GMM	(4) FD GMM
Differences				
Knowledge Attributes	-0.019**	-0.018**	0.021**	0.028**
	(0.009)	(0.009)	(0.010)	(0.007)
Knowledge Attributes^2	0.000**	0.000**	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)
# of Publications	0.008*	0.008*	0.008	0.002
	(0.005)	(0.005)	(0.006)	(0.004)
# of Coauthors	-0.004**	-0.004**	0.003	-0.002**
	(0.001)	(0.001)	(0.003)	(0.001)
# of UMN coauthors	0.003*	0.003	-0.002	0.001
	(0.002)	(0.002)	(0.002)	(0.001)
Average Team Size	-0.020**	-0.021**	-0.002	-0.003
	(0.010)	(0.010)	(0.011)	(0.008)
Average Annual Citations	0.001	0.001	-0.005*	-0.005*
	(0.003)	(0.004)	(0.003)	(0.003)
Controls				
Average # of Publications	0.000	0.000	0.024*	0.002
	(0.009)	(0.009)	(0.013)	(0.007)
Average # of Coauthors	0.011**	0.011**	-0.011	0.002
	(0.003)	(0.003)	(0.007)	(0.002)
Average # of UMN Coauthors	-0.013**	-0.013**	-0.009**	-0.006**
	(0.003)	(0.003)	(0.004)	(0.002)
Average Team Size	-0.065**	-0.064**	0.033	-0.001
	(0.018)	(0.018)	(0.026)	(0.013)
Average Annual Citations	-0.006	-0.007	0.003	0.003
	(0.004)	(0.005)	(0.004)	(0.003)
Past Citation		0.002	0.001	0.011
		(0.006)	(0.012)	(0.009)
First Collaboration			-0.025**	-0.322**
			(0.007)	(0.022)
# of Skipped Years				0.348**
				(0.027)
Hansen	0.580	0.632	0.064	0.410
AR(1)	0.000	0.000	0.000	0.000
AR(2)	0.065	0.065	0.706	0.357
# of Pairs	21,445	21,445	21,445	21,445
Observations	58,555	58,555	58,555	58,555

Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.

Note: Two-step system GMM employed; Year effects included; Hansen is Hansen J test for overidentifying restrictions (p-values reported); AR(1) and AR(2) are tests for first and second-order serial correlation (p-values reported); Standard errors in parentheses; * p<0.10, ** p<0.05.

Appendix C

C.1 Subject Categories in Journal Citation Report (JCR)

Table C-1

List of Subject Categories

	Subject Categories		Subject Categories
1	Acoustics	110	Logic
2	Agricultural Economics Policy	111	Management
3	Agricultural Engineering	112	Marine Freshwater Biology
4	Agriculture Dairy Animal Science	113	Materials Science Biomaterials
5	Agriculture Multidisciplinary	114	Materials Science Ceramics
6	Agronomy	115	Materials Science Characterization Testing
7	Allergy	116	Materials Science Coatings Films
8	Anatomy Morphology	117	Materials Science Composites
9	Andrology	118	Materials Science Multidisciplinary
10	Anesthesiology	119	Materials Science Paper Wood
11	Anthropology	120	Materials Science Textiles
12	Astronomy Astrophysics	121	Mathematical Computational Biology
13	Audiology Speech-Language Pathology	122	Mathematics
14	Automation Control Systems	123	Mathematics Applied
15	Behavioral Sciences	124	Mathematics Interdisciplinary App.
16	Biochemical Research Methods	125	Mechanics
17	Biochemistry Molecular Biology	126	Medical Ethics
18	Biodiversity Conservation	127	Medical Informatics
19	Biology	128	Medical Laboratory Technology
20	Biophysics	129	Medicine Legal
21	Biotechnology Applied Microbiology	130	Medicine Research Experimental
22	Business	131	Metallurgy Metallurgical Engineering
23	Business Finance	132	Meteorology Atmospheric Sciences
24	Cardiac Cardiovascular Systems	133	Microbiology
25	Cell Biology	134	Microscopy
26	Cell Tissue Engineering	135	Mineralogy
27	Chemistry Analytical	136	Mining Mineral Processing
28	Chemistry Applied	137	Multidisciplinary Sciences
29	Chemistry Inorganic Nuclear	138	Mycology
30	Chemistry Medicinal	139	Nanoscience Nanotechnology
31	Chemistry Multidisciplinary	140	Neuroimaging
32	Chemistry Organic	141	Neurosciences
33	Clinical Neurology	142	Nuclear Science Technology
34	Communication	143	Nursing
35	Computer Science Artificial Intelligence	144	Nutrition Dietetics
36	Computer Science Cybernetics	145	Obstetrics Gynecology
37	Computer Science Hardware Architecture	146	Oceanography
38	Computer Science Information Systems	147	Oncology
39	Computer Science Interdisciplinary App.	148	Operations Research Management Science
40	Computer Science Software Engineering	149	Ophthalmology
41	Computer Science Theory Methods	150	Optics
42	Construction Building Technology	151	Ornithology
43	Criminology Penology	152	Orthopedics
44	Critical Care Medicine	153	Otorhinolaryngology
45	Crystallography	154	Paleontology
46	Cultural Studies	155	Parasitology
47	Demography	156	Pathology

48 Dentistry Oral Surgery Medicine
49 Dermatology
50 Developmental Biology
51 Ecology
52 Economics
53 Education Educational Research
54 Education Scientific Disciplines
55 Educational Special
56 Electrochemistry
57 Emergency Medicine
58 Endocrinology Metabolism
59 Energy Fuels
60 Engineering Aerospace
61 Engineering Biomedical
62 Engineering Chemical
63 Engineering Civil
64 Engineering Electrical Electronic
65 Engineering Environmental
66 Engineering Geological
67 Engineering Industrial
68 Engineering Manufacturing
69 Engineering Marine
70 Engineering Mechanical
71 Engineering Multidisciplinary
72 Engineering Ocean
73 Engineering Petroleum
74 Entomology
75 Environmental Sciences
76 Environmental Studies
77 Ergonomics
78 Ethnic Studies
79 Evolutionary Biology
80 Family Studies
81 Fisheries
82 Food Science Technology
83 Gastroenterology Hematology
84 Genetics Heredity
85 Geochemistry Geophysics
86 Geography
87 Geography Physical
88 Geology
89 Geosciences Multidisciplinary
90 Geriatrics Gerontology
91 Gerontology
92 Health Care Sciences Services
93 Health Policy Services
94 History
95 History Philosophy Science
96 History Social Sciences
97 Horticulture
98 Hospitality Leisure Sport Tourism
99 Imaging Science Photographic Tech.
100 Immunology
101 Industrial Relations Labor
102 Infectious Diseases
103 Information Science Library Science
157 Pediatrics
158 Peripheral Vascular Disease
159 Pharmacology Pharmacy
160 Physics Applied
161 Physics Atomic Molecular Chemical
162 Physics Fluids Plasmas
163 Physics Mathematical
164 Physics Multidisciplinary
165 Physics Nuclear
166 Physics Particles Fields
167 Physiology
168 Planning Development
169 Plant Sciences
170 Political Science
171 Polymer Science
172 Primary Health Care
173 Psychiatry
174 Psychology
175 Psychology Biological
176 Psychology Clinical
177 Psychology Experimental
178 Psychology Mathematical
179 Psychology Multidisciplinary
180 Psychology Psychoanalysis
181 Psychology Applied
182 Psychology Educational
183 Psychology Social
184 Public Administration
185 Public Environmental Occupational Health
186 Radiology Nuclear Medicine Medical Imaging
187 Rehabilitation
188 Remote Sensing
189 Reproductive Biology
190 Respiratory System
191 Rheumatology
192 Robotics
193 Social Issues
194 Social Sciences Biomedical
195 Social Sciences Interdisciplinary
196 Social Sciences Mathematical Methods
197 Social Work
198 Sociology
199 Soil Science
200 Spectroscopy
201 Sport Sciences
202 Statistics Probability
203 Substance Abuse
204 Surgery
205 Telecommunications
206 Thermodynamics
207 Toxicology
208 Transplantation
209 Transportation
210 Transportation Science Technology
211 Tropical Medicine
212 Urban Studies

104	Instruments Instrumentation	213	Urology Nephrology
105	Integrative Complementary Medicine	214	Veterinary Sciences
106	International Relations	215	Virology
107	Law	216	Water Resources
108	Limnology	217	Women Studies
109	Linguistics	218	Zoology

Source: Journal Citation Reports (JCR) 2014.

C.2 An Example of Number of Skipped Years without Collaboration

The following table shows an example of two collaboration pairs to help understand the concept of the number of skipped years without collaboration. Ioannidis et al. (2014) and Bu et al. (2017b) introduced this concept as one method to measure persistent scientific collaboration.

In the example, both pairs have collaborated from 2002 to 2010, and the total number of papers published, time since first paper (as of 2010), and the number of skipped years without collaboration are identical. However, the nature of collaboration differs between pair 1 and pair 2, because pair 1 has more dispersed years without collaboration while pair 2 has a longer stretch consecutive years of no publication. This suggests that for pair 2, collaboration is likely to be (temporarily, at least) terminated. For pair 1, however, a few years of no collaboration suggests that researchers may still collaborated in those periods although the results have not been published yet (Bu et al. 2017b).

Table C-2

An Example of Number of Skipped Years without Collaboration

	2002	2003	2004	2005	2006	2007	2008	2009	2010	Total	Time since First Paper in 2010	NSY
pair 1	1	0	1	0	0	1	0	0	1	4	8	5
pair 2	1	1	0	0	0	0	0	1	1	4	8	5

Source: Developed by author.

Note: Numbers in each cell represents the numbers of papers published in each year.

C.3 Non-Linear Forms

Table C-3

Estimation Results using Non-Linear Functional Forms

	(1) Logit	(2) Marginal Effects	(3) Probit	(4) Marginal Effects
Differences				
Knowledge Attributes	0.014***	0.003***	0.008***	0.003***
	(0.002)	(0.000)	(0.001)	(0.000)
Knowledge Attributes^2	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
# of Publication	-0.008**	-0.001**	-0.005**	-0.001**
	(0.003)	(0.000)	(0.002)	(0.000)
# of Coauthors	0.002**	0.000**	0.001**	0.000**
	(0.001)	(0.000)	(0.000)	(0.000)
# of UMN Coauthors	-0.008***	-0.001***	-0.005***	-0.001***
	(0.001)	(0.000)	(0.001)	(0.000)
Average Team Size	-0.062***	-0.012***	-0.037***	-0.012***
	(0.005)	(0.001)	(0.003)	(0.001)
Average Annual Citation	-0.008*	-0.002*	-0.004*	-0.001*
	(0.003)	(0.001)	(0.002)	(0.001)
Controls				
Average # of Publications	0.027***	0.005***	0.016***	0.005***
	(0.005)	(0.001)	(0.003)	(0.001)
Average # of Coauthors	-0.005***	-0.001***	-0.003***	-0.001***
	(0.001)	(0.000)	(0.001)	(0.000)
Average # of UMN Coauthors	0.014***	0.003***	0.009***	0.003***
	(0.002)	(0.000)	(0.001)	(0.000)
Average Team Size	0.139***	0.026***	0.082***	0.026***
	(0.008)	(0.002)	(0.005)	(0.002)
Average Annual Citations	0.007	0.001	0.004	0.001
	(0.005)	(0.001)	(0.003)	(0.001)
Both Faculty	-0.614***	-0.114***	-0.356***	-0.112***
	(0.041)	(0.007)	(0.023)	(0.007)
Observations	68,180	68,180	68,180	68,180

Source: Developed by author using data from Scopus, UMN Data Warehouse, Journal Citation Reports (JCR) 2014, and own calculations.

Note: Marginal effects in columns (2) and (4); Standard errors in parentheses; ; * p<0.10, ** p<0.05, *** p<0.01.

5 Conclusion

The objective of this dissertation is to examine the increasing role of R&D investment among private firms in the U.S. food and agricultural sectors (Chapter 2) and research collaborations among the life scientists at UMN (Chapter 3 and 4).

In Chapter 2, I first present and discuss the details of newly updated firm-level data pertaining to R&D conducted over the period 1950-2014 by food or agriculturally related businesses located in the United States. I identify the shifting structure of investments in machinery, agriculture and chemicals, and food and beverage processing R&D, emphasizing changes in the portfolio of firms conducting this research amongst other details. I also present the results of empirical analysis that examines the stylized association between firm-level R&D expenditures and a firm's financial position, such as sales growth and profit rate. I find that firms invest more in R&D when expectations of future sales are high. Firms not only adjust their contemporaneous R&D spending when past sales increase but also invest more in capital which eventually increases the R&D investment. These results hold even if I eliminated firms with high variance of R&D growth and little R&D spending. The findings of this chapter contribute to understand the increasing dominance of private R&D in the food and agricultural sectors in the United States and around the world.

Chapter 3 and Chapter 4 are inter-related and examines the research collaboration among the life sciences at UMN. Chapter 3 examines the patterns of (internal and external) research collaborations by UMN life sciences researchers and the factors that might drive those partnerships. By using a panel dataset of publications and coauthorship information drawn from Scopus for the period 1999-2014, I find that researchers have become more inclined to collaborate, and most notably, with researchers outside of their own institutions. This suggests that researchers increasingly perceive that benefits from outside collaborations outweigh the costs that arise from collaborating across institutions. Not only do outside collaborations incur higher citation numbers than collaboration within the same institution, but also the cost of cross-institution collaborations is declining with more new UMN faculty hires who received their post-graduate degrees from non-UMN institutions.

Using the same data, Chapter 4 focuses on the papers that are produced internally, without coauthors outside of UMN. This paper estimates the likelihood of collaboration for three particular types of information signals: perceived research productivity, knowledge complementarity, and professional familiarity. To measure the knowledge complementarity in research collaboration, I use the subject categories that are assigned to the journals in which the cited references in the publications are published. This helps us quantity a researcher's overall knowledge profile and the knowledge complementarity in a more refined way. Using system generalized methods of moments (GMM), I find that researchers initiate a collaboration with others who have similar knowledge and high productivity. For researcher pairs who collaborate again, however, productivity no longer remains significant while complementary knowledge is positively associated with collaboration. Researchers are more likely to collaborate again when the first collaboration was relatively recent and when they have worked with past collaborators less frequently.

The findings of Chapter 3 and Chapter 4 not only contribute to the empirical literature on research collaboration and team formation, but also has an important policy implication for policy makers, funding agencies, and research institutes who strongly favor research collaboration. The findings may also shed light on collaboration within and across firms, especially when that collaboration targets the acquisition or leveraging of knowledge.

In conclusion, this dissertation provides a deeper understanding on the increasing importance of private firms in the U.S. food and agricultural sectors and life scientists who collaborate across disciplines and institutions. The findings of this dissertation will allow U.S. to unveil the "black box" of innovation activities the U.S. food, health and agricultural sectors.

Bibliography

Adams, J. D., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999. *Research Policy*, *34*(3), 259–285. https://doi.org/10.1016/j.respol.2005.01.014

Adams, J. (2013). Collaborations: The fourth age of research. *Nature*, *497*(7451), 557–560. https://doi.org/10.1038/497557a

Agrawal, A., & Goldfarb, A. (2008). Restructuring research: Communication costs and the democratization of university innovation. *American Economic Review*, *98*(4), 1578–1590. https://doi.org/10.1257/aer.98.4.1578

Alston, J. M., Andersen, M. A., James, J. S., & Pardey, P. G. (2009). *Persistence pays: U.S. agricultural productivity growth and the benefits from public R&D spending (Vol. 34)*. Springer Science & Business Media.

Alston, J. M., Andersen, M. A., James, J. S., & Pardey, P. G. (2011). The economic returns to U.S. Public agricultural research. *American Journal of Agricultural Economics*, *93*(5), 1257–1277. https://doi.org/10.1093/ajae/aar044

Alston, J. M., & Pardey, P. G. (2006). Agricultural Productivity. In S. B. Carter, S. S. Gartner, M. R. Haines, A. L. Olmstead, R. Sutch, & G. Wright (Eds.), *Historical Statistics of the United States, Earliest Times to the Present—Millennial Edition, Volume 4, Part D, Economic Sectors*. Cambridge: Cambridge University Press.

Alston, J. M., & Pardey, P. G. (2014). Agriculture in the Global Economy. *Journal of Economic Perspectives*, *28*(1), 121–146. https://doi.org/10.1257/jep.28.1.121

Anderson, T. W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, *18*(1), 47–82. https://doi.org/10.1016/0304-4076(82)90095-1

Arellano, M., & Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, *58*(2), 277–297. https://doi.org/10.2307/2297968

Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, *68*(1), 29–51. https://doi.org/10.1016/0304-4076(94)01642-D

Azoulay, P., Liu, C. C., & Stuart, T. E. (2017). Social Influence Given (Partially) Deliberate Matching: Career Imprints in the Creation of Academic Entrepreneurs. *American Journal of Sociology*, *122*(4), 1223–1271. https://doi.org/10.1086/689890

Azoulay, P., Zivin, J. S. G., & Wang, J. (2010). Superstar Extinction. *Quarterly Journal of Economics Journal of Economics*, *125*(2), 549–589.

Barnett, A. H., Ault, R. W., & Kaserman, D. L. (1988). The rising incidence of co-authorship in economics: Further evidence. *The Review of Economics and Statistics*, *70*(3), 539–543. https://doi.org/10.2307/1926798

Bidault, F., & Hildebrand, T. (2014). The distribution of partnership returns: Evidence from co-authorships in economics journals. *Research Policy*, *43*(6), 1002–1013. https://doi.org/10.1016/j.respol.2014.01.008

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*(1), 115–143. https://doi.org/10.1016/S0304-4076(98)00009-8

Bond, S., Harhoff, D., & Reenen, J. Van. (2005). Investment, R&D and Financial Constraints in Britain and Germany. *Annales d'Économie et de Statistique*, (79/80), 433–460. https://doi.org/10.2307/20777584

Bosquet, C., & Combes, P. P. (2013). Are academics who publish more also more cited? Individual determinants of publication and citation records. *Scientometrics*, *97*(3), 831–857. https://doi.org/10.1007/s11192-013-0996-6

Boudreau, B. K. J., Brady, T., Ganguli, I., Gaule, P., Guinan, E., Hollenberg, A., & Lakhani, K. R. (2017). A Field Experiment on Search Costs and the Formation of Scientific Collaborations. *Review of Economics and Statistics*, *99*(4), 565–576. https://doi.org/10.1162/REST

Brown, J. R., Fazzari, S. M., & Petersen, B. C. (2009). Financing innovation and growth: Cash flow, external equity, and the 1990s r&d boom. *Journal of Finance*, *64*(1), 151–185. https://doi.org/10.1111/j.1540-6261.2008.01431.x

Brown, J. R., & Petersen, B. C. (2009). Why has the investment-cash flow sensitivity declined so sharply? Rising R&D and equity market developments. *Journal of Banking and Finance*, *33*(5), 971–984. https://doi.org/10.1016/j.jbankfin.2008.10.009

Bu, Y., Ding, Y., Xu, J., Liang, X., Gao, G., & Zhao, Y. (2017a). Understanding success through the diversity of collaborators and the milestone of career. *Journal of the Association for Information Science and Technology*, *69*(1), 87-97. https://doi.org/10.1002/asi.23911

Bu, Y., Ding, Y., Liang, X., & Murray, D. S. (2017b). Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology*, *69*(3), 438-448. https://doi.org/10.1002/asi.23966

Bureau of Economic Analysis. (2018). Current-Dollar and "Real" Gross Domestic Product.

Castaldi, C., & Los, B. (2012). Are New "Silicon Valleys" Emerging? The Distribution of Superstar Patents across US States. In *DRUID Society Conference*. Copenhagen, Denmark.

Caswell, M., & Day-Rubenstein, K. (2006). Agricultural research and development. In *Agricultural Resources and Environmental Indicators* (pp. 59–65). Economic Research Service/USDA.

Catalini, C., Fons-Rosen, C., & Gaulé, P. (2016). Did Cheaper Flights Change the Geography of Scientific Collaboration? *Ssrn*, (9897). https://doi.org/10.2139/ssrn.2764219

Catalini, C. (2018). Microgeography and the Direction of Inventive Activity. *Management Science*, (December). https://doi.org/10.2139/ssrn.2126890

Chai, Y., Pardey, P. G., Chan-Kang, C., Huang, J., Lee, K., & Dong, W. (2019). *Passing the Food and Agricultural R&D Buck? The United States and China*. Manuscript submitted for publication.

Chung, K. H., Cox, R. A. K., & Kim, K. A. (2009). On the relation between intellectual collaboration and intellectual output: Evidence from the finance academe. *Quarterly Review of Economics and Finance*, *49*(3), 893–916. https://doi.org/10.1016/j.qref.2008.08.001

Coccia, M., & Wang, L. (2016). Evolution and convergence of the patterns of international scientific collaboration. *Proceedings of the National Academy of Sciences*. *113*(8), 2057-2061. https://doi.org/10.1073/pnas.1510820113

Dhand, A., Luke, D. A., Carothers, B. J., & Evanoff, B. A. (2016). Academic cross-pollination: The role of disciplinary affiliation in research collaboration. *PLoS ONE*, *11*(1), 1–13. https://doi.org/10.1371/journal.pone.0145916

Ductor, L. (2014). Does Co-authorship lead to higher academic productivity? *Oxford Bulletin of Economics and Statistics*, *77*(3), 385. https://doi.org/10.1111/obes.12070

Ductor, L., Fafchamps, M., Goyal, S., & van de Leij, M. J. (2014). Social Networks and Research Output. *The Review of Economics and Statistics*, *96*(5), 936–948.

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, *55*(2), 251–76. https://doi.org/http://dx.doi.org/10.2307/1913236

Epanechnikov, V. (1969). Nonparametric estimation of a multidimensional probability density. *Theory of Probability & Its Applications*, *14*(1), 153–158.

Fafchamps, M., & Gubert, F. (2007). The formation of risk sharing networks. *Journal of Development Economics*, *83*(2), 326–350. https://doi.org/10.1016/j.jdeveco.2006.05.005

Fafchamps, M., van der Leij, M. J., & Goyal, S. (2010). Matching and network effects. *Journal of the European Economic Association*, *8*(1), 203–231. https://doi.org/10.1111/j.1542-4774.2010.tb00500.x

Finholt, T. A. (2002). Collaboratories. *Annual Review of Information Science and Technology*, *36*(1), 73–107.

Freeman, R. B. (2015). Immigration, International Collaboration, and Innovation: Science and Technology policy in the global economy. In W. R. Kerr, J. Lerner, & S. Stern (Eds.), *Innovation Policy and the Economy* (Vol. 15, pp. 1–29). University of Chicago Press. https://doi.org/10.1086/680062

Freeman, R. B., Ganguli, I., & Murciano-Goroff, R. (2014). Why and Wherefore of Increased Scientific Collaboration. *NBER Working Paper*, (No. 19819). Retrieved from http://papers.nber.org/tmp/99857-w19819.pdf

Freeman, R. B., & Huang, W. (2015). Collaborating with People Like Me: Ethnic Co-authorship within the U.S. *Journal of Labor Economics*, *33*(S1), S289–S318. https://doi.org/10.3386/w19905

Fuglie, K. O., Heisey, P. W., King, J. L., Pray, C. E., Day-Rubenstein, K., Schimmelpfennig, D., … Karmarkar-Deshmukh, R. (2011). *Research Investments and Market Structure in the Food Processing, Agricultural Input, and Biofuel Industries Worldwide*. *USDA-ERS Economic Research Report*, (No. 130).

Fuglie, K. O., & Toole, A. A. (2014). The evolving institutional structure of public and private agricultural research. *American Journal of Agricultural Economics*, *96*(3), 862–883. https://doi.org/10.1093/ajae/aat107

Fuglie, K. (2016). The growing role of the private sector in agricultural research and development world-wide. *Global Food Security*, *10*, 29–38. https://doi.org/10.1016/j.gfs.2016.07.005

Galiani, S., & Galvez, R. H. (2017). The Life Cycle of Scholarly Articles Across Fields of Research. *NBER Working Paper*, (No. 23447).

Geroski, P. A., Machin, S., & Walters, C. (1997). Corporate Growth and Profitability. *Global Journal of Business Research*, *7*(1), 43–59. Retrieved from http://discovery.ucl.ac.uk/16987/

Gibbons, R. (2005). Four formal(izable) theories of the firm? *Journal of Economic Behavior and Organization*, *58*(2), 200–245. https://doi.org/10.1016/j.jebo.2004.09.010

Gilbert, R. J., & Newbery, D. M. G. (1982). Preemptive patenting and the persistence of monopoly. *American Economic Review*, *72*(3), 514–526. https://doi.org/10.1126/science.151.3712.867-a

Goyal, S. (2007). *Connections: An Introduction to the Economics of Networks*. *Journal of Artificial Societies and Social Simulation*. Princeton University Press. https://doi.org/10.1007/s00712-008-0036-9

Griliches, Z. (1988). Productivity Puzzles and R & D: Another Nonexplanation. *The Journal of Economic Perspectives*, *2*(4), 9–21.

Hagedoorn, J. (1993). Understanding the rationale of strategic technology partnering: Interorganizational modes of cooperation and sectoral differences. *Strategic Management Journal*, *14*(5), 371-385. https://doi.org/10.1002/smj.4250140505

Hagedoorn, J., & Schakenraad, J. (1994). The effect of strategic technology alliances on company performance. *Strategic Management Journal*, *15*(4), 291-309. https://doi.org/10.1002/smj.4250150404

Hall, B. H. (1992). Investment and Research and Development at the Firm Level: Does the Source of Financing Matter? *NBER Working Paper*, (No. 4096). Retrieved from http://products.sanofi.us/praluent/PRALUENT_Prefilled_Pen_IFU_150 mg.pdf

Hall, B. H., & Lerner, J. (2010). *The financing of R&D and innovation*. *Handbook of the Economics of Innovation* (Vol. 1). Elsevier B.V. https://doi.org/10.1016/S0169-7218(10)01014-2

Hamel, G. (1991). Competition for competence and interpartner learning within international strategic alliances. *Strategic Management Journal*, *12*(S1), 83-103. https://doi.org/10.1002/smj.4250120908

Hampton, S. E., & Parker, J. N. (2011). Collaboration and Productivity in Scientific Synthesis. *BioScience*, *61*(11), 900–910. https://doi.org/10.1525/bio.2011.61.11.9

Harhoff, D. (1998). Are There Financing Constraints for R&D and Investment in German Manufacturing Firms? *Annals of Economics and Statistics*, *49/50*(Jan-June), 421–456. https://doi.org/10.1007/978-1-4757-3194-1_16

Hemlin, S., Allwood, C. M., & Martin, B. R. (Eds.). (2004). *The influences on creativity in research and innovation*. Edward Elgar Publishing.

Hilmer, C. E., & Hilmer, M. J. (2005). How Do Journal Quality, Co-Authorship, and Author Order Affect Agricultural Economists' Salaries? *American Journal of Agricultural Economics*, *87*(2), 509–523.

Himmelberg, C. P., & Petersen, B. C. (1994). R & D and Internal Finance : A Panel Study of Small Firms in High-Tech Industries. *The Review of Economics and Statistics*, *76*(1), 38–51.

Holsapple, C. W. (2004). Knowledge and its attributes. In *Handbook on Knowledge Management 1* (pp. 165–188). Berlin, Heidelberg.: Springer.

Hsiehchen, D., Espinoza, M., & Hsieh, A. (2015). Multinational teams and diseconomies of scale in collaborative research. *Science Advances*, *1*(8), e1500211. https://doi.org/10.1126/sciadv.1500211

Hudson, J. (1996). Trends in multi-authored papers in economics. *Journal of Economic Perspectives*, *10*(3), 153–158.

Ioannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PLoS ONE*, *9*(7), e101698. https://doi.org/10.1371/journal.pone.0101698

Jaffe, A. B. (1986). Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value. *American Economic Review*, *76*(5), 984–1001.

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

Kabo, F. W., Cotton-Nessler, N., Hwang, Y., Levenstein, M. C., & Owen-Smith, J. (2014). Proximity effects on the dynamics and outcomes of scientific collaborations. *Research Policy*, *43*(9), 1469–1485. https://doi.org/10.1016/j.respol.2014.04.007

Kaplan, S. N., & Zingales, L. (1997). Do Investment-Cash Flow Sensitivities Provide Useful Measures of Financing Constraints? *Quarterly Journal of Economics*, *112*(1), 169–215. https://doi.org/10.1162/003355397555163

Katz, J. S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, *40*(3), 541–554. https://doi.org/10.1007/BF02459299

Katz, J. S. (1993). *Bibliometric Assessment of Intranational University–University Collaboration*. University of Sussex, Brighton.

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

Kim, E. H., Morse, A., & Zingales, L. (2009). Are elite universities losing their competitive edge? *Journal of Financial Economics*, *93*(3), 353–381. https://doi.org/10.1016/j.jfineco.2008.09.007

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137-163. https://doi.org/10.1093/oxfordjournals.pan.a004868

King, J. L., & Schimmelpfennig, D. (2005). Mergers, acquisitions, and stocks of agricultural biotechnology intellectual property. *AgBioForum*, *8*(2–3), 83–88.

Kogut, B. (1988). Joint ventures: Theoretical and empirical perspectives. *Strategic Management Journal*, *9*(4), 319-332. https://doi.org/10.1002/smj.4250090403

Laband, D. N., & Tollison, R. D. (2000). Intellectual collaboration. *Journal of Political Economy*, *108*(3), 632–662.

Laband, D., & Piette, M. (1994). Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *The Journal of Political Economy*, *102*(1), 194-203. https://doi.org/10.1086/261927

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., & Sugimoto, C. R. (2016). Contributorship and division of labor in knowledge production. *Social Studies of Science*, *46*(3), 417–435. https://doi.org/10.1177/0306312716650046

Larivière, V., Gingras, Y., & Archambault, É. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, *68*(3), 519–533. https://doi.org/10.1007/s11192-006-0127-8

Lazear, E. P. (2000). The Power of Incentives. *American Economic Review*, *90*(2), 410–414. https://doi.org/10.1257/aer.90.2.410

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

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*(3), 684–697. https://doi.org/10.1016/j.respol.2014.10.007

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, *5*(1), 87–100. https://doi.org/10.1016/j.joi.2010.09.002

Li, G. C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., … Lee, F. (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010). *Research Policy*, *43*(6), 941–955. https://doi.org/10.1016/j.respol.2014.01.012

McPherson, M., Popielarz, P., & Drobnic, S. (1992). Social Networks and Organizational Dynamics. *American Sociological Review,* 153-172. https://doi.org/10.2307/2096202

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415-444. https://doi.org/10.1146/annurev.soc.27.1.415

Miguel, J. G. S. (1977). The Reliability of R&D Data in COMPUSTAT and 10-K Reports. *The Accounting Review*, *52*(3), 638–641.

Millar, M. M. (2013). Interdisciplinary research and the early career: The effect of interdisciplinary dissertation research on career placement and publication productivity of doctoral graduates in the sciences. *Research Policy*, *42*(5), 1152–1164. https://doi.org/10.1016/j.respol.2013.02.004

Mulkay, B., Hall, B. H., & Mairesse, J. (2001). Firm Level Investment and R&D in France and the United States: A Comparison. In *Investing Today for the World of Tomorrow* (pp. 229–273). Springer, Berlin, Heidelberg.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5200–5205.

Nickell, S. J. (1978). *The Investment Decisions of Firms*. Cambridge: Cambridge University Press.

Nohria, N., & Garcia-Pont, C. (1991). Global strategic linkages and industry structure. *Strategic Management Journal, 12*(S1), 105-124. https://doi.org/10.1002/smj.4250120909

O'Brien, T. L. (2012). Change in Academic Coauthorship, 1953-2003. *Science Technology & Human Values*, *37*(3), 210–234. https://doi.org/10.1177/0162243911406744

Ottonello, P., & Winberry, T. (2018). Financial Heterogeneity and the Investment Channel of Monetary Policy. *NBER Working Paper*, (No. 24221).

Pardey, P. G., Chan-Kang, C., Dehmer, S. P., & Beddow, J. M. (2016). Agricultural R&D is on the move. *Nature*, *537*(September), 301–303. https://doi.org/10.1038/537301a

Pardey, P. G., Chan-Kang, C., Dehmer, S. P., & Beddow, J. M. (2016b). Documentation — InSTePP International Innovation Accounts: Research and Development Spending, version 3.5 (Food and Agricultural R&D Ser.). *InSTePP Center*.

Pardey, P. G., & Craig, B. (1989). Causal Relationship between Public Sector Agricultural Research Expenditures and Output. *American Journal of Agricultural Economics*, *71*(1), 9–19. https://doi.org/10.2307/1241770

Pardey, P. G., & Graff, G. D. (2013). Efficacy of translating genomic research to innovations in a global IP environment. Unpublished manuscript.

Pfeffer, J., & Nowak, P. (1976). Joint Ventures and Interorganizational Interdependence. *Administrative Science Quarterly*, 398-418. https://doi.org/10.2307/2391851

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745. https://doi.org/10.1007/s11192-008-2197-2

Powell, W. W., Koput, K. W., & Smith-Doerr, L. (1996). Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotechnology. *Administrative Science Quarterly*, *41*(1), 116–145. https://doi.org/10.2307/2393988

Pray, C. E., & Fuglie, K. O. (2015). Agricultural Research by the Private Sector. *Annual Review of Resource Economics*, *7*(1), 399–424. https://doi.org/10.1146/annurev-resource-100814-125115

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, *82*(2), 263–287. https://doi.org/10.1007/s11192-009-0041-y

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, *21*(1), 24-43. https://doi.org/10.1016/0040-5809(82)90004-1

Reinganum, J. F. (1983). Uncertain Innovation and the Persistence of Monopoly. *American Economic Review*, *73*(4), 741–748. https://doi.org/10.1257/jep.6.3.79

Rosenberg, N. (1982). *Inside the Black Box: Technology and Economics*. Cambridge University Press.

Roodman, D. (2006). How to Do xtabond2: An Introduction to "Difference" and "System" GMM in Stata. *The Stata Journal*, *9*(1), 86-136. https://doi.org/The Stata Journal

School of Public Health. (n.d.). Research - School of Public Health - University of Minnesota. Retrieved January 29, 2019, from http://www.sph.umn.edu/research/research

Securities and Exchange Commission. (1972). Adoption of Amendments to Regulation S-X (Accounting Series Release No. 125).

Silverberg, G., & Verspagen, B. (2007). The size distribution of innovations revisited: An application of extreme value statistics to citation and value measures of patent significance. *Journal of Econometrics*, *139*(2), 318-339. https://doi.org/10.1016/j.jeconom.2006.10.017

Spender, J.-C. (1996). Making Knowledge the Basis of a Dynamic Theory of the Firm. *Strategic Management Journal*, *17*(Winter), 45–62. https://doi.org/10.1002/smj.4250171106

Stephan, P. (2012). *How Economics Shapes Science*. Cambridge, MA: Harvard University Press.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707-719. https://doi.org/10.1098/rsif.2007.0213

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. https://doi.org/10.1109/ICDM.2008.71

U.S. Securities and Exchange Commission. (n.d.). Division of Corporation Finance: Standard Industrial Classification (SIC) Code List. Retrieved March 21, 2019, from https://www.sec.gov/info/edgar/siccodes.htm

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

Valderas, J. M. (2007). Why Do Team-Authored Papers Get Cited More ? *Science*, *317*(5844), 1496–1498.

Van Reenen, J. (1997). Employment and Technological Innovation: Evidence from U.K. Manufacturing Firms. *Journal of Labor Economics*, *15*(2), 255. https://doi.org/10.1086/209833

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, *34*(10), 1608–1618.

Wang, J., & Hicks, D. (2015). Scientific teams: Self-assembly, fluidness, and interdependence. *Journal of Informetrics*, *9*(1), 197–207. https://doi.org/10.1016/j.joi.2014.12.006

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

Zeng, X. H. T., Duch, J., Sales-Pardo, M., Moreira, J. A. G., Radicchi, F., Ribeiro, H. V., … Amaral, L. A. N. (2016). Differences in Collaboration Patterns across Discipline, Career Stage, and Gender. *PLOS Biology*, *14*(11), e1002573. https://doi.org/10.1371/journal.pbio.1002573

Zhu, M., Huang, Y., & Contractor, N. S. (2013). Motivations for self-assembling into project teams. *Social Networks*, *35*(2), 251–264. https://doi.org/10.1016/j.socnet.2013.03.001