Essays on the Production and Commercialization of New Scientific Knowledge

by

Michaël Bikard

Diplôme des Grandes Écoles, MSc in Management, Diplom Kaufmann, ESCP Europe (2005)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

Signature of Author...
MIT Sloan School of Management
July 31, 2012

Certified by...
Fiona E. Murray
SARNOFF ~~Associate~~ Professor of Technological Innovation, Entrepreneurship, and Strategic Management
~~Associate~~ FACULTY Director, Martin Trust Center for MIT Entrepreneurship
Thesis Supervisor

Certified by...
Scott Stern
School of Management Distinguished Professor of Technological
Innovation, Entrepreneurship and Strategic Management
Thesis Supervisor

Accepted by...
Ezra Zuckerman Sivan
Nanyang Technological University Professor
Codirector, Economic Sociology PhD Program; Chair, PhD Program

Dissertation Committee Members

Fiona Murray

Associate Professor of Technological Innovation,
Entrepreneurship, and Strategic Management
Associate Director, Martin Trust Center for MIT Entrepreneurship
MIT Sloan School of Management

Scott Stern

School of Management Distinguished Professor
of Technological Innovation, Entrepreneurship
and Strategic Management
MIT Sloan School of Management

Ezra Zuckerman Sivan

Nanyang Technological University Professor
Codirector, Economic Sociology PhD Program
Chair, PhD Program
MIT Sloan School of Management

Essays on the Production and Commercialization of New Scientific Knowledge

by

Michaël Bikard

Submitted to the MIT Sloan School of Management on
June 1, 2013, in partial fulfillment of the requirements for the
degree of Doctor of Philosophy

ABSTRACT

Scientific research frequently generates tremendous economic value. Yet, this value tends to be elusive and public and private organizations often struggle to obtain returns from their investment in science. This dissertation, composed of three essays, examines persistent challenges to the production and commercialization of new scientific knowledge.

The first essay of the dissertation describes simultaneous discoveries and their potential as a research tool for social science. It also introduces the first systematic and automated method to generate a list of such events. The resulting dataset of 578 recent simultaneous discoveries can be used to investigate a number of questions, including the impact of the discovery environment, by using them to conduct the first "twin studies" of new knowledge. As an example, the second essay investigates the relative impact of universities and firms on science-based invention by examining 39 discoveries made simultaneously in academia and in industry. As compared to universities, the results indicate that firms amplify the technological impact of new scientific knowledge. The third essay of the dissertation, coauthored with Fiona Murray and Joshua Gans, explores tradeoffs associated with collaboration in the production of new scientific knowledge. Specifically, we find that collaboration is not only associated with higher-quality output, it is also associated with lower individual productivity as well as challenges surrounding the allocation of credit. Taken together, the three essays examine important challenges associated with the production and commercialization of new scientific knowledge—thus providing insights about the drivers of economic value from public and private investment in science.

Thesis Supervisor: Fiona E. Murray
Title: Associate Professor of Technological Innovation, Entrepreneurship, and Strategic Management; Associate Director, Martin Trust Center for MIT Entrepreneurship

Thesis Supervisor: Scott Stern
Title: School of Management Distinguished Professor of Technological Innovation, Entrepreneurship and Strategic Management

Acknowledgements

This dissertation would not have been possible without the support of many, and I am immensely grateful to those around me who have given me so much intellectual, emotional and financial support.

I feel extremely lucky that I was able to work as a PhD student under the supervision of Fiona Murray, Scott Stern and Ezra Zuckerman. While each of them provided deep insights about his or her unique approach to research, it is through their conversations that I learned how disciplines can complement one another. I am especially grateful that Scott Stern chose to come to MIT and to take me as his student when I was in the middle of my PhD journey. It is difficult to imagine that I would have had the confidence to invest so much energy on the study of simultaneous discoveries without his encouragement and mentorship. I also received substantial support from other faculty members at MIT and beyond. I particularly benefited from the wisdom and experience of Pierre Azoulay, Lotte Bailyn, Emilio Castilla, Michael Cusumano, Jason Davis, Roberto Fernandez, Jeff Furman, Olenka Kacperczyk, Kate Kellogg, Alan MacCormack, Matt Marx, Elena Obukhova, Ray Reagans, Mike Roach, Ed Roberts, Henry Sauermann, Susan Silbey, Michael Tushman, Chris Wheat and Pai-Ling Yin.

I owe a large debt to the broader MIT community. The assistance and support from Colette Boudreau, Richard Brewer, Sharon Cayley, Alice Downing, Christine Haynes, Keira Horowitz, Pam Liu, Lisa Miyake and Hillary Ross made my life as a PhD student considerably easier. I am also grateful to Anthony Sinskey and Stacy Springs at the MIT Center for Biomedical Innovation for getting me actively engaged with the practitioner side of university-based innovation. Finally, I have benefited enormously from the help and camaraderie of other MIT Sloan PhD students and feel fortunate that we overlapped for a few years. I am particularly indebted to Chuck Eesley, Jason Greenberg, Hiram Samel, and Yanbo Wang for having given me so much advice over the years. I would also like to thank Phil Anderson, Alan Benson, Rodrigo Canales, Kenny Ching, Joelle Evans, Dan Fehder, John-Paul Ferguson, Roman Galperin, Oliver Hall, Ryan Hammond, Benjamin Mako Hill, Jason Jay, Stella Kounelaki, Joshua Krieger, Eun-Suk Lee, John Lyneis, Ethan Mollick, Abhishek Nagaraj, Aruna Ranganathan, Ben Rissing, Eunhee Sohn, Lynn Wu and Enying Zheng.

To complete my dissertation project, I had to acquire knowledge and skills that go far beyond the social sciences. I am immensely grateful to my brother David Bikard for introducing me to computer programing and synthetic biology. Jisi Tang of Boston University and Katherine Romer of MIT helped me implement the algorithm that found the simultaneous discoveries on which my dissertation is based. Xavier Duportet at MIT and Gilad Evrony at Harvard Medical School were not only incredible friends, but also helped me understand daily life in the lab. I am grateful as well to a number of undergraduate and graduate students at MIT who helped me in one way or another in my research; a particular thank-you to Devin Fensterheim, Kelly Peterson and Tendelle Sheu. Finally, I am grateful to the dozens of scientists from around the US and beyond who were willing to answer my questions about the scientific discoveries that they were involved in.

My dissertation has benefited from the generous financial support of various organizations that have seen some promise in its research program. At MIT, in addition to the

Sloan Doctoral Fellowship, the dissertation has been awarded the MIT Energy Fellowship and a Roberts E-Center Fund Award, as well as considerable support from the Center of Biomedical Innovation at the Department of Biology. Outside of MIT, the dissertation has benefited from the Kauffman Dissertation Fellowship, a J. Robert Beyster Fellowship and an NSF SBE Research Doctoral Dissertation Grant.

Last but not least, over the past few years I have received considerable support from a few people without whom I cannot imagine my life now. During my time in Cambridge, I met and married my wife Arielle and am indebted to America forever for having produced her. Also, her family in New York—Elaine, Marty, Daniel and Margot—as well as Ana, provided me with a warm home on this side of the Atlantic ocean as well as considerable help editing the English of a large number of papers and essays. I am particularly grateful to Bijou for our long conversations and her provocative research ideas. Jonathan Kluberg has been supportive of me since my very first days in the US. I also feel particularly lucky to have become close friends with Patrick Gaulé, Laurent Guérin, and Stéphane Chong. Finally, my parents Claire and Patrick and siblings Arnaud, David, Aurélia, Marine, Pierre-Emmanuel, François and Florian were always there for me and they have remained a constant source of love and happiness over the years.

CONTENTS

Essays on the Production and Commercialization of New Scientific Knowledge

Chapter One

Essays on the Production and Commercialization of New Scientific Knowledge: Introduction and Overview

1.1. BACKGROUND

This dissertation investigates the relationship between organizations and science. On the one hand, public and private organizations invest heavily in pushing the scientific frontier with the hope that economic gains will follow (e.g., Cohen and Levinthal 1990; Henderson, Jaffe, and Trajtenberg 1998; Aghion et al. 2010). Some scientific discoveries can indeed open the door to the creation of new technologies, new firms, or even new industries (Rosenberg and Nelson 1994; Fleming and Sorenson 2004). On the other hand, economic returns from investments in scientific research are often disappointing. The process of scientific discovery involves tremendous uncertainty and the appropriate organization of knowledge work is often unclear. In addition, it sometimes takes years before a new piece of scientific knowledge is used to produce a novel technology (Rosenberg 1994; Mokyr 2002). High failure rates make investment in science-based innovation tremendously costly. People, companies, and nations intending to use science as a source of competitive advantage need to understand these tensions.

To address these critical issues, this dissertation explores the process of production and commercialization of new scientific knowledge. In three distinct essays, it investigates the phenomenon of simultaneous discoveries, the process by which organizations use scientific knowledge to produce new technologies, and the organization of scientific work. It therefore examines the production and commercialization of scientific knowledge from a variety of perspectives, at the level of individuals, organizations, and knowledge itself. This research proposes simultaneous discoveries—"knowledge twins"—as a new research tool for social scientists. It also provides insights about the role that organizations play in amplifying or obstructing the technological impact of new scientific discoveries. Finally, it expands our understanding of the tradeoffs associated with collaboration in scientific research. The ambition of this dissertation is to help scholars, managers, and policymakers generate economic value from scientific research.

Micro-economic empirical analysis of the production of scientific knowledge and of its development into new technologies has traditionally relied on large bibliometric and patent datasets (e.g., Henderson, Jaffe, and Trajtenberg 1998; Fleming and Sorenson 2004; Wuchty, Jones, and Uzzi 2007). These datasets have been extremely helpful in uncovering general patterns and trends in scientific discovery and in invention. One limit of such datasets, however, is that the process preceding discovery, invention, or failure, is unobserved. As a result, the drivers of productive efficiency in discovery and in invention remain unclear. In order to get around this difficulty, this dissertation uses various approaches that complement those bibliometric and patent datasets. For instance, the first essay proposes a new research tool—simultaneous discoveries operationalized as "paper twins"—and uses insights from computer science and sociology to generate a large dataset of such events. The second essay studies these knowledge twins in order to analyze science-based invention or the absence thereof as a function of the environment of discovery. Finally, the third essay examines publications but at the level of a scientist's year of work.

1.2. OVERVIEW OF THE DISSERTATION ESSAYS

Scientific knowledge is often seen as a source of competitive advantage for individuals, firms, and nations. Yet, the creation of economic value from scientific research is challenging, raising important questions about the appropriate process of production and commercialization of new knowledge. Contributing to this line of inquiry, this dissertation is composed of three essays investigating respectively (1) the phenomenon of simultaneous discoveries and its potential as a research tool, (2) the use or non-use of new scientific knowledge to produce new technologies, and (3) the organization of scientific work.

1.2.1 Simultaneous Discoveries as a Research Tool: Method and Promise

Half a century after Merton's description of simultaneous discoveries "as a strategic research site" (Merton 1963), they are hardly ever used by social scientists. This essay attempts to unleash the potential of simultaneous discoveries as a research tool. First, they provide a lens into the determinants of creativity in general and scientific advancement in particular (e.g., Merton 1961). Indeed, their frequency constitutes striking evidence that creative ideas, although novel, might not necessarily be unique. Second, simultaneous discoveries provide important insights about the process of social construction of science (Kuhn 1969). These events are often associated with racing and conflict about credit allocation and are therefore revealing of many features of the institutions that contour scientists' behavior. Third, simultaneous discoveries are instances in which the same knowledge emerges around the same time in two different environments. As such, they can be used to conduct "twin studies" of new knowledge and identify the impact of the environment on the utilization (or non-utilization) of that knowledge.

This essay also proposes the first systematic and automated method to build a dataset of simultaneous discoveries. The method goes beyond the old debates about the scientific similarity of two (twin) discoveries. It is based on the insight that teams of scientists that make the same discovery around the same time will share the credit for that discovery—and that credit-sharing will be visible in the citation patterns of scientific papers (Cozzens 1989). This method is further implemented into an algorithm that generates a dataset of 578 recent simultaneous discoveries made by 1,246 teams of scientists working in a variety of settings around the world.

1.2.2 Is Knowledge Trapped Inside the Ivory Tower? Technology Spawning and the Genesis of New Science-Based Inventions

The third essay of the dissertation investigates some of the conditions under which organizations translate —or fail to translate— scientific discoveries into new technologies. Historical examples reveal that, while this development can be very rapid in certain circumstances, scientific knowledge can sometimes remain unexploited for years (Rosenberg 1994; Mokyr 2002). For instance, the first person who purified EPO, Eugene Goldwasser, could not find anyone who would invest in turning his scientific discovery into a new technology. The first firm that did work on this project, five years later, was a start-up named Amgen; and it created a new technology (recombinant EPO) that became one of the most successful drugs ever produced by the biotechnology industry. Discerning the circumstances under which scientific discoveries are developed into new technologies is very difficult because the technological potential of the scientific knowledge is always unobserved. This essay addresses this challenge by using simultaneous discoveries to conduct the first "twin study" of new scientific knowledge.

Specifically, the paper examines the relative impact of universities and firms as discovery environments on science-based invention (e.g., Henderson, Jaffe, and Trajtenberg 1998; Aghion et al. 2010). Analysis of follow-on inventions, based on 39 simultaneous discoveries between academia and industry involving 90 teams, reveals that the team from industry produces more than 3 times more inventions based on its discovery than the co-discoverers from academia. Moreover, third-party inventors are 10-20% more likely to cite the industry publication in their patent than its academic twin. Taken together, these results indicate that new scientific knowledge is more likely to be utilized to produce new technologies if it emerges in firms than if it emerges in the "Ivory Tower."

1.2.3 Exploring Tradeoffs in the Organization of Scientific Work: Collaboration and Scientific Reward

This essay, coauthored with Fiona Murray and Joshua Gans, explores the use of collaboration in scientific research. Prior studies on the topic have been optimistic about this organization of creative work, showing for instance that more collaborative scientific papers (and patents) tend to be of higher quality than those that have fewer authors (e.g., Wuchty, Jones, and Uzzi 2007; Singh and Fleming 2010). This type of evidence does not consider, however, that collaboration is a choice, and that its benefits in terms of output quality might be offset by coordination costs and challenges with regard to credit allocation. This paper explores these tradeoffs in two ways. First, we develop a formal model to structure our understanding of the factors shaping scientists' collaborative choices. Second, we test our model's assumptions empirically by examining the actual choices made by 661 faculty-scientists from one institution – the Massachusetts Institute of Technology – over a thirty-year period from 1976 to 2006.

We find that collaboration is associated with important tradeoffs, including higher-quality publications, lower individual productivity and disproportionate credit attribution—i.e. that credit for a given collaborative paper is shared across coauthors in a way that sums to more than 1. Interestingly, these results suggest that the "net value" of collaboration in creative work might be superior for the credit-seeking worker than it is for the output-focused manager or policy-maker. The type of collaborator has also important consequences. For instance, the benefits of collaboration are particularly high and its costs are particularly low when the collaboration brings together individuals having different skills and perspectives—as in the case of cross-departmental collaborations.

1.3. FUTURE DIRECTIONS AND CONCLUSIONS

This dissertation opens the door to a variety of potential studies of the production and commercialization of new scientific knowledge. Below, I describe four such studies that are currently underway but are not formally part of this dissertation.

One study examines the influence of the geographic location of the discovery team on science-based invention. The project, titled "Geographic Localization of Knowledge Spillovers: Evidence from Knowledge Twins," investigates current debates about the extent to which knowledge spillovers are localized. Using citations of 275 twin papers in the patent literature, it is possible to identify the extent to which inventors are more likely to draw on scientific knowledge that emerges within closer geographic proximity—while keeping the discovery constant. Early results indicate that knowledge spillovers are localized not only at the country level, but also very strongly at the metropolitan-area level.

Another project investigates the division of innovative labor across different types of organizations. This study is titled "In the Shadow of Uncertainty: Entrepreneurial Strategy and the Selection of New Projects," and it explores how entrepreneurs exploit uncertainty (Knight 1921) to compete against incumbents in pharmaceutical R&D. Using instances in which the same discovery is made simultaneously in an entrepreneurial venture and at a large firm, the preliminary results indicate that entrepreneurs tend to disengage from projects involving too little uncertainty for fear of competition with companies that have much greater resources. On the other hand, larger firms tend to reject ideas with high uncertainty, providing space for young firms to grow, "sheltered from competition" by this very uncertainty.

The managerial implications of the division of innovative labor are explored in a working paper titled "Idea-Centered Innovation Management: A Novel Approach to R&D for the

Biopharmaceutical Industry." Most current approaches to innovation management do not consider that R&D ideas have different uncertainty profiles, therefore calling for different organizational structure. This paper proposes a new approach to innovation management that takes this uncertainty into account. The relevance of the new approach is illustrated in the context of the alliance between Sanofi and the Center for Biomedical Innovation at MIT.

Regarding the production of new scientific knowledge, one project, titled "Is Collaborative Diversity Creative or Costly? Group Composition, Inspiration, and Time Wasted" (with Fiona Murray), proposes an analysis of the specific impact of different types of collaborative diversity on creative performance. Our preliminary results indicate that the costs and benefits of collaboration are driven by distinct mechanisms. For a given scientist, the addition of a new collaborator is associated with lower productivity but does not have any significant correlation with output quality. On the other hand, collaborating across disciplines is associated with higher-quality work but is not significantly linked with lower productivity. These findings highlight the complexity of the relationship between collaboration and creativity and provide a more nuanced view of the tradeoffs associated with collaboration in scientific research.

In conclusion, the overarching ambition of this research is to improve our understanding of how individuals, firms and nations can use scientific research as a source of competitive advantage. This agenda will continue to require innovative approaches, using various empirical strategies and drawing insights from a number of disciplines. My hope is that this dissertation and continuing research will provide new insights about the drivers of scientific discovery, science-based innovation, organizational performance, and economic growth.

1.4. REFERENCES

Aghion, Philippe, Mathias Dewatripont, Julian Kolev, Fiona Murray, and Scott Stern. 2010. "The Public and Private Sectors in the Process of Innovation: Theory and Evidence from the Mouse Genetics Revolution." *The American Economic Review* 100 (2) (May 1): 153–158.

Cohen, Wesley M., and Daniel A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation." *Administrative Science Quarterly* 35 (1) (March): 128–152.

Cozzens, Susan E. 1989. *Social Control and Multiple Discovery in Science: The Opiate Receptor Case*. State Univ of New York Press.

Fleming, Lee, and Olav Sorenson. 2004. "Science as a Map in Technological Search." *Strategic Management Journal* 25: 909–928.

Henderson, Rebecca M., Adam B. Jaffe, and Manuel Trajtenberg. 1998. "Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965-1988." *Review of Economics and Statistics* 80 (1) (February): 119–127.

Knight, Frank H. 1921. *Risk, Uncertainty and Profit*. Boston, MA: Hart, Schaffner & Marx.

Kuhn, Thomas S. 1969. "Energy Conservation as an Example of Simultaneous Discovery." In *Critical Problems In The History Of Science: Proceedings Of The Institute For The History Of Science, 1957*, 321.

Merton, Robert K. 1961. "Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science." *Proceedings of the American Philosophical Society* 105 (5) (October 13): 470–486.

———. 1963. "Resistance to the Systematic Study of Multiple Discoveries in Science." *European Journal of Sociology* 4 (August): 237–282.

Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton University Press.

Rosenberg, Nathan. 1994. *Exploring the Black Box: Technology, Economics, and History*. Cambridge Univ Press.

Rosenberg, Nathan, and Richard R. Nelson. 1994. "American Universities and Technical Advance in Industry." *Research Policy* 23 (3) (May): 323–348.

Singh, Jasjit, and Lee Fleming. 2010. "Lone Inventors as Sources of Breakthroughs: Myth or Reality?" *Management Science* 56 (1): 41–56.

Wuchty, Stefan, Benjamin Jones, and Brian Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316 (5827): 1036.

Chapter Two

Simultaneous Discoveries as a Research Tool: Method and Promise

2.1. INTRODUCTION

Simultaneous discoveries are a strategic research setting for a variety of purposes. First, they raise important questions concerning the predictability of scientific advancement and the extent to which it can be influenced by policy (Merton 1961; Simonton 1999). Second, they are a lens into scientific norms and other institutionalized mechanisms used by scientists to deal with priority and conflict (Kuhn 1969; Cozzens 1989). Third, as instances of knowledge "twins," they constitute a quasi-natural experiment allowing researchers to investigate the impact of the context of discovery on the commercialization of that knowledge (Hounshell 1975; Voss 1984).

Despite its potential as a research setting, the study of simultaneous discoveries has been hindered by tremendous difficulties of definition and operationalization. While no author ever claimed that simultaneous discoveries did not exist, heated debates have opposed those who believe that scientific multiples are the norm (Merton 1961) and those who argue that they are the exception (Schmookler 1966). Partisans of multiples as the norm have based their argument on large lists of multiples stemming from historical accounts (Ogburn and Thomas 1922; Merton 1961; Simonton 1979; Niehans 1995). Their opponents have conducted in-depth analyses of alleged simultaneous discoveries and inventions and have argued that the technological or scientific similarity in these instances is superficial at best (Schmookler 1966; Constant 1978; Patinkin 1983; De Marchi 1995).

These debates have not been resolved, and social scientists in the past 15 years have generally turned away from studying simultaneous discoveries. Yet, the phenomenon itself has not disappeared from the scientific discourse. Biomedical researchers continue to write and worry about such events (Troyer 2001; Saudek 2003; Castillo 2008). Every year, the USPTO typically deals with more than 100 cases of interferences—cases in which two or more inventors submit patent applications claiming the same invention at the same time (Merz and Henry 2004). Finally, there is no reason to believe that scientists are less worried today about being forestalled (Marshall 2002; Lawrence 2003) than they were a few decades ago (Hagstrom 1974).

This paper attempts to unleash the tremendous potential of simultaneous discoveries as a research tool for social scientists. First, drawing on prior research, we show that these events can be used to address a number of fascinating theoretical questions. Second, we propose a novel method to identify simultaneous discoveries which provides a way out of the old disputes about technological or scientific similarity. Third, we implement this method into the first systematic and reproducible algorithm to generate a list of simultaneous discoveries. Fourth, we describe the resulting dataset, which includes 578 events and is the first such dataset that is based on recent occurrences.

Although new and creative, scientific discoveries are often not unique. The fact that the same new idea might emerge at the same time in two different places has potentially important implications for research on creativity. As illustrated by the frequency of priority disputes, simultaneous discoveries have also important behavioral consequences for scientists who are rewarded for originality. The ability to observe the same knowledge in several different places ought to provide a useful lens for researchers who are concerned with the economic impact of new ideas. A deep exploration of any one of these research directions is beyond the scope of this paper. Rather, by presenting a new method to collect a dataset of simultaneous discoveries and by pointing to the theoretical promise of such data, our hope is to (re)establish a potentially rich research tool for social science.

2.2. SIMULTANEOUS DISCOVERIES AS A RESEARCH SETTING

2.2.1. The antecedents of scientific discovery

Prior research has mostly used scientific multiples as a strategic research site to investigate the antecedents of scientific discovery. Three broad approaches have been taken that explain the emergence of these events: (1) studies of "culture" (i.e. the state of science and society), (2) studies of simultaneous discoveries as emerging from a stochastic process, and (3) studies of within-multiple differences and similarities. While many have noted that these approaches are not mutually exclusive, the literature has been marked by disputes opposing authors who identify rather clearly with one of these three positions.

Simultaneous discoveries and inventions have traditionally been considered evidence of the crucial role played by culture as opposed to individuals in the advancement of science and technology. Indeed, if many discoveries or inventions are made at the same time by a variety of individuals, then it is probably that such advance was inevitable and would have been made by someone anyway (Kroeber 1917; Ogburn and Thomas 1922). Merton set the basis for a sociological approach to simultaneous discoveries in the most eloquent manner: their occurrence "suggests that discoveries become virtually inevitable when prerequisite kinds of knowledge and tools accumulate in man's cultural store and when the attention of an appreciable number of investigators becomes focused on a problem, by emerging social needs, by developments internal to science, or by both" (Merton 1963, 237). A number of variables have been uncovered from mostly historical inquiries. These sociological variables include the discoverer's education, the social demand for a discovery and the current state of knowledge (Ogburn and Thomas 1922; Merton 1961). For instance, Kuhn conducted an in-depth historical study of the simultaneous discovery of energy conservation (Kuhn 1969). He asks: "Why, in the years 1830-50, did so many of the experiments and concepts required for a full statement of energy conservation lie so close to the surface of scientific consciousness?" and suggests the following three answers: the availability of conversion processes, the concern with engines, and the philosophy of nature. Constant studied the alleged simultaneous inventions of the steam turbine and Pelton water wheels and also concludes from his analysis that three variables account for the evolution of technology: the inventor's paradigmatic commitment, the general engineering standards and ideologies and technological co-evolution. Constant's theory of technological co-evolution considers that "the development of one set of devices may be intimately linked to the development of other devices within a macro-system, and that the two sets of devices may exert powerful mutual selective pressure on each other" (Constant 1978, 184). Instead of exploring the cultural factors leading to simultaneous discoveries, Brannigan and Wanner argue that multiples ought to be an anomaly since no scientist likes to be part of one, and most fear to be forestalled (Brannigan and Wanner 1983a). They developed a communication theory according to which independent discoveries are instances of imperfect communication between scientists. As communication becomes easier, multiples become less frequent and increasingly simultaneous.

The second stream of the literature on simultaneous discoveries has argued that they are evidence of the importance of chance in the evolution of science. Price most notably remarked that there are many discoveries toward which several people are working simultaneously. He further proposes to approximate the emergence of multiples using a Poisson distribution in what he describes as the "ripe apple" model. "If there are 1000 apples on a tree, and 1000 blindfolded men reach up at random to pick an apple, what is the chance of a man getting one to himself, or finding himself grasping as well the hand of another picker, or even more than one?" (de Solla Price 1965, 60) The result is that 368 apples will be left on the tree. Out of the rest, 37% will be picked by a single hand whereas 63% will end in multiple discovery. Price compares this distribution to the list of multiples' frequency collected by Merton and Barber and notes that both are consistent with each other. Simonton elaborated on this idea, arguing that the existence of numerous "nulltons" in this type of model—or discoveries that were not made—shows that discoveries are not inevitable. Indeed, the idea that the discovery process follows a Poisson distribution is supported by the frequency of multiples in the available datasets, and this suggest that scientific evolution is more determined by chance than by sociological factors (Simonton 1978; Simonton 1979; Simonton 1986).

The third stream of literature argues that a large proportion of the alleged scientific multiples are in fact not multiples at all. As a result, the frequency of multiples is being over-estimated, which suggests that researchers fail to appreciate the importance of individuals in the evolution of science. Schmookler, for instance, attacks the largest list of scientific multiples collected and published at the time (by Ogburn and Thomas): "the list in question is based largely on a failure to distinguish between the genus and the individual. Whatever the term 'the' electric telegraph may mean, the telegraphs of Henry, Morse, Cooke and Wheatstone, and Steinheil were not the same telegraphs. (...) The Ogburn-Thomas list of 'duplicates' consists in fact of inventions with similar generic terms (...) Those who regard inventions bearing such titles as identical are like tourists to whom all Chinamen look alike" (Schmookler 1966, 191). Using in-depth analyses, researchers in this tradition have argued that many alleged multiples are in fact quite different, be it substantially (e.g. nature of the discovery or invention) or functionally (e.g. the interpretation of the discovery and its purpose) or both (Constant 1978; Elkana 1971; Patinkin 1983). In short, "scientific research is less redundant—and, as a corollary, that the individual scientist is more important—than the by-now-familiar long lists of alleged multiple discoveries lead us to believe" (Patinkin 1983, 320).

These studies have in common that they considered scientific multiples as outcome variable. They have proposed various factors that might explain the variation in the number of scientists taking part to the same discovery or invention independently (i.e whether it is a singleton, a doubleton, a tripleton, a quadrupleton, etc…or even a nullton). In their enterprise, they have all been severely limited by the scarcity of data available since barely three quantitative dataset of multiples have been developed and only one of them (Ogburn-Thomas) has been published. Besides, no predicting variable has been developed in a way that could allow quantitative analyses.

2.2.2. Other uses of simultaneous discoveries and inventions

Besides exploring the antecedents of discovery, scientific multiples have been used for two other purposes in the social science literature: exploring (1) the process of social construction of scientific discoveries and (2) the process of commercialization of new inventions.

Simultaneous discoveries have been used as a strategic research site to explore the social construction of discoveries and the attribution of credit. Indeed, they are cases in which these processes are often uneasy, to the extent that they often involve open conflict. Building on Brannigan's call to "consider not what make discoveries happen but what makes them discoveries" (Brannigan 1981, 152), Cozzens proposes a change in perspective in the study of multiples from the viewpoint of the historian to the viewpoint of the scientific community: "We began with the assumption that no two contributions to science are never identical in all respects, and raised the question: 'How do scientists decide when two or more contributions to science are similar enough to be grouped together as a multiple discovery?'" (Cozzens 1989, 163). Cozzens finds that scientific multiples do not exist per se but are socially constructed by scientists *after* the discoveries have been made. "An after-the-fact process is also needed, a fine-tuning device which accomplishes the thorough homogenization of the contributions into a recognized multiple discovery." (Cozzens 1989, 170) Using interviews and archival data, she digs into this process of social construction of multiples in the case of the discovery of the opiate receptor. She finds that scientific multiples emerge from decisions of credit attribution made by third parties and are visible through patterns of citations. Such decision of citation attribution is a moral one and is loosely based on standards of justice such as public evidence of simultaneity and independence. In Cozzens' case, the simultaneous discovery emerged as a "social moral convention adopted to help solve the problem of social conflict" (Cozzens 1989, 161). However, she also suggests that the simultaneous discovery could also have existed had there been no conflict—i.e. had all the co-discoverers agreed to share the credit for the discovery.

Because they are instances in which two or more individuals have developed the same new piece of knowledge or technology, scientific multiples are also an interesting setting to explore the process of commercialization of new ideas. Hounshell, for instance, studied the simultaneous invention of the telephone by Alexander Graham Bell and Elisha Gray. He asks: "why is Bell so widely if not universally known as the inventor of the telephone and Gray, who envisioned the same device at the same time, known to few except historians of technology?" (Hounshell 1981, 157). Using historical data, Hounshell offers as explanation that Gray was disadvantaged because he was an expert and a member of a community of experts of the telegraph industry. His deep knowledge of the needs of this industry as well as his social network blinded him from the commercial potential of the talking telegraph. While he had the opportunity to contest Bell's patent (both invention disclosures arrived on the same day--February 14 1876--at the US Patent Office), he decided to focus instead on multiplex telegraphy, an invention for which he saw much larger commercial promises (Hounshell 1975). In another such study Voss considers the origin and commercialization outcome of 17 independent inventions of "real-time software on small computer systems to produce shipping documentation and to support the activities of shipping agents and freight forwarders." The author finds a multitude of ways in which the innovation was precipitated, developed and commercialized and argues that this case shows the benefits of studying the emergence of technology and its diffusion jointly (Voss 1984).

Robert Merton's call to use scientific multiples as a strategic research site was heard for about twenty years. Attention to these events decreased in the mid-1980s and no publication in a top tier journal has used this research setting since. Subsequently, questions about the determinants of scientific evolution, the social construction of discoveries and the process of commercialization of new ideas have all been explored in ways that did not involve the use of simultaneous discoveries or inventions. Yet, their tremendous potential as a research setting has been established. In order to go beyond the fierce debates about the collection of dataset of simultaneous discoveries in a reproducible manner, we describe below a new systematic method to generate a database of these twins of new scientific knowledge.

2.3. METHOD OF BUILDING A DATASET OF SIMULTANEOUS DISCOVERIES

2.3.1. Prior art

A number of authors have built lists of scientific multiples. Three large lists have received particular attention. The oldest one brings together 148 multiples (Ogburn and Thomas 1922), the second 264 multiples (Merton 1961) and the third 579 multiples (Simonton 1979)[1]. One important commonality of these three lists is that they were compiled based on the accounts of historians of science. In the footnote accompanying their list of multiples, Ogburn and Thomas mention that developing such a list is sometimes difficult because disagreements often exist about the similarity and independence of multiples. Unfortunately, they do not propose any criteria of inclusion or exclusion of multiples: “Our guides have been the histories of science, and where there are differences in the historical accounts, we have followed the general practice.” Similarly, Merton and Barber’s list of 264 multiples was obtained through historical inquiry but the authors have not published their list and we could not find any detailed account of the method used. Lastly, Simonton built a large list of 579 cases made of “all multiples mentioned in at least one source as long as no contradictory evidence could be found in any other source” and tested his results on a shorter, exclusive list including “only those multiples mentioned in two separate sources, without contradiction by any other source” (Simonton 1979, 609). More recent lists also exist. Niehans generated a list of 40 cases of scientific multiples in economic theory. Although he discusses the difficulty associated with building such a list, he

[1] Only Ogburn and Thomas have made their list public

does not detail the process through which his list was built (Niehans 1995). In contrast, Voss's study of the multiple inventions of freight industry software provides a detailed description of the manner in which the dataset was generated. A first list of potential multiples was created using literature search and various interviews. Similarity was ensured using strict definition about the specifications of the software invented and independence was verified using interviews and archival research in the journals that the inventors read at the time of invention (Voss 1984).

As noted above, these lists–and the frequency of multiples–have typically been hotly debated. On the one hand, a number of authors have criticized these studies based on the assertion that two discoveries or inventions are always somehow different, be it substantially or functionally (Constant 1978; Schmookler 1966; Elkana 1971; Brannigan and Wanner 1983b; Patinkin 1983). This difficulty was in fact noted by Merton himself: "It is no easy matter to establish the degree of similarity between independently developed ideas. Even in the more exact disciplines, such as mathematics, claims of independent multiple inventions are vigorously debated. The question is, how much overlap should be taken to constitute 'identity'?" (Merton 1968, 9–10). On the other hand, other authors have documented the fact that researchers often admit their fear of being forestalled. Hagstrom, for instance, found that more than 60% of the 1,718 US scientists he surveyed declared having been anticipated by another scientist in the publication of a discovery at least once in their career (Hagstrom 1974). Merton noted that "another kind of evidence seems presumptive if not compelling evidence of identity or equivalence: the report of a *later* discoverer that another had arrived there before him. Presumably, these reports are truthful since the modern age of science puts a premium on originality" (Merton 1968, 10).

Two authors have proposed specific criteria for the identification of multiples and noted that few of the instances recorded in multiple lists would meet these criteria. Elkana focuses on the case of the conservation of energy principle and argues that the alleged co-discoverers brought different answers to different problems, and that it is only with the hindsight of time that their discoveries seem identical. Although he concedes that his criterion might not be suitable to all types of multiples, Elkana proposes that "such discoveries should be considered as simultaneous which give related answers to similar problems" (Elkana 1974, 178). Patinkin studies Keynes' General Theory with its alleged co-discoverers the Polish economist Kalecki and the Stockholm School of Economics. Although elements of the General Theory were indeed anticipated by other economists, Patinkin claims, they were typically not identically strictly speaking and were not part of these economists' central message. Patinkin argues that two criteria should be adopted when constituting a list of multiples: a precise definition of the discovery and an examination of "the extent to which the alleged co-discoverers 'really meant it'" (Patinkin 1983, 306).

The state of the art in the construction of datasets of scientific multiples is unsatisfactory. Multiples tend to be matched by historians of science with no clear criteria about which discoveries ought to be considered. To the extent that this approach is based upon subjective decisions about relevance by the individuals compiling the list (and/or the historians that they read), the analysis is open to criticism for possible bias and lack of reproducibility. It is not surprising, then, that this unsystematic method led to severe skepticism. As Niehans puts it: "Multiple discoveries are a 'fuzzy set', and the harder one tries to delineate it, the fuzzier it looks" (Niehans 1995, 7).

2.3.2. Using adjacent citations to detect multiples: Theoretical basis

Independently developed discoveries and inventions are never exactly identical. In fact, total replication in science is believed to be impossible (Collins 1992). Yet, scientists continue to speak about simultaneous discoveries and inventions and continue to be involved in priority disputes. How can we reconcile these two facts?

One way to resolve this contradiction is to observe that discoveries are socially constructed by the scientific community (Cozzens 1989). Simultaneous discoveries, therefore, do not exist per se but are instead the result of a process of "homogenization" by the scientific community which perceives the results obtained by several independent discoverers as equivalent. Most interestingly for our purpose, Cozzens' study shows that the result of this homogenization process is apparent in the scientific literature through the citation patterns of those that build on the new knowledge. Indeed, when mentioning a discovery that was a multiple, scientists typically cite all the co-discoverers, and in so doing split the credit between them. Two important dimensions of citations in academic work explain such a practice.

First, citations are symbols for particular ideas, methods, and experimental data in the scientific literature. As they are being cited, complex documents become concept-symbols, i.e. they acquire "a standard or conventional interpretation that is crucial for the social determination of scientific ideas" (Small 1978, 338). The process of attribution of meaning is a collective and emergent process which "condenses or 'capsulizes' a complex original text into a few standard statements." As symbols, citations link specific discoveries (i.e. ideas) with specific publications. In so doing, they provide the reader with a source of the idea, potentially also acknowledging intellectual debt and invoking authority to legitimate a new knowledge claim. This perspective however does not explain Cozzens' finding that independent co-discoverers tend to be cited at the same time by the literature. Because citations are used as symbols for a particular idea, the use of several adjacent citations (e.g. in the same parenthesis) can be perceived as redundant since "from a strictly scientific point of view, reference to one single paper would be sufficient" (Moravcsik and Murugesan 1975, 90).

Second, citations are not only symbols but are also a critical currency in the cycles of scientific credit (Latour and Woolgar 1986) . Because credit needs to be split between co-discoverers, the scientific community typically tends to cite all the discovery papers (when there is more than one) for a specific discovery. Cozzens's interviews show that the position of the community on whether or not a discovery is a multiple can be read in the way it is cited by the subsequent literature: "the interviews with third parties indicated that a close examination of what was said about the co-discoveries when they were cited would say something about the extent of consensus." In fact, Cozzens's interviewees described proper citation as a "moral obligation"—as well as the "a way of expressing one's own viewpoint" on the debate of who deserves credit (Cozzens 1989, 120–121). In the case of the opiate receptor, citations can therefore be used as indicator of the "votes" of the scientific community on whether or not the discovery ought to be considered a simultaneous discovery.

Two dimensions of citation in the academic literature thus collide and make citation attribution an ideal setting to detect simultaneous discoveries. As symbols, citations homogenize discoveries by linking publications with specific ideas. As currency in the cycles of scientific credit, several equivalent publications tend to be cited together when referring to a multiple discovery. Co-citation proximity is therefore a good measure of whether or not two publications embed the same specific idea. Co-citation frequency is a good indicator of whether or not credit splitting takes place between two publications for the same general idea. As illustrated in Figure 1, these two dimensions are orthogonal to each other: it is at the intersection between high co-citation frequency and high co-citation proximity that we are likely to find simultaneous discoveries.

[Insert Figure 1 here]

Of course, citations are also used for other purposes than concept-symbols and credit attribution. Such purpose includes "window dressing", marking membership in social groups, protection of "property rights," etc. A number of citation typologies have been developed and recent work has highlighted that citations often fulfill more than one purpose at a time (for a recent review of the citation literature, see Callahan, Hockema, and Eysenbach 2010). In order to detect simultaneous discoveries, the honorific use of citations (Biagioli and Galison 2003) could be especially problematic since it would suggest that prominent scientists are more likely to be given undeserved credit as co-discoverers whereas less prominent ones might be left out. While we cannot completely exclude this possibility we can nonetheless note that this interpretation would not be consistent with Cozzens's findings. Her analysis shows indeed clearly that in the case of the opiate receptor, credit splitting was taken very seriously by the scientific community and that the latter split the credit among the four co-discoverers regardless of their social status, centrality or claims.

2.3.3. Using adjacent citations to detect multiples: Empirical basis

Since the independent discovery of co-citation analysis nearly 40 years ago (Marshakova 1973; Small 1973), citations have been widely used to quantify the scientific similarity. Co-citation refers to the share of forward citations that two documents have in common (Small 1973) and ought to be distinguished from bibliographic coupling, which refers to the share of references that two documents have in common (Kessler 1963). Co-citation studies have traditionally been taking place at the level of the document and rest on the observation that more frequently co-cited documents tend to be more similar. However, this approach has also been applied to measure the scientific proximity at other levels of analysis such as the academic journal (Narin 1976) or the individual scientist (White and Griffith 1981).

Co-citation analyses have broadly been endeavored for two distinct purposes. The first use of such a metric has aimed at the establishment of "maps of science." Indeed, answering De Solla Price's call to map science (de Solla Price 1965) became one of the first applications of co-citation analysis and a program to do so was launched as early as 1974. While this approach has not been exempt from criticism (e.g. Leydesdorff 1987), the production of maps of science has thrived (Small 1998) and is still common today (Tsai and Wu 2010). The second use of co-citations has been the development of similarity metrics to relate documents. Such a metric has proved particularly useful with the development of large databases of scientific publications. Indeed, a number of search engines such as *CiteSeer*[2] have been developed that use co-citations to compute the relatedness between academic articles (Giles, Bollacker, and Lawrence 1998).

Recent work has proposed to refine the co-citation metric with an analysis not only of the frequency of co-citations but also with the analysis of the proximity of the citations in the co-citing papers. Using parsing algorithms, several authors have shown that the measure of similarity between scientific papers can be improved significantly by observing co-citations at the level of the sentence (Gipp and Beel 2009; Tran et al. 2009).

The above-mentioned literature provides a helpful basis for the conception of an algorithm detecting simultaneous discoveries. Indeed, in many respects, the discoveries can be operationalized as closely related papers written around the same time and having no author in common. In order to ensure that two papers are not only related but are in fact instantiations of the same new knowledge, it is important to ensure that citations patterns reflect credit sharing and symbolic equivalence. We use consistent adjacent citations (e.g. in the same parentheses) in order to identify simultaneous discoveries—or paper twins.

2.3.4. Algorithm

[2] http://citeseer.ist.psu.edu

We describe below the first systematic and automated method to build a dataset of simultaneous discoveries – here operationalized as "paper twins". Unlike prior lists, it is entirely transparent and reproducible. The algorithm matches papers based on their date, authors and citation patterns. Specifically considered paper twins are all the pairs of research articles having no author in common, written no more than one calendar year apart, being frequently cited by the same papers and being consistently cited adjacent to each other (e.g. in the same parenthesis). The algorithm involves five steps described in Figure 2.

1) Collection of a sample of citing articles
2) Attribution of a unique identifier to each of their references
3) Generation of a sample of candidate twins based on these references
4) Computation of a measure of co-citation frequency at the paper level
5) Selection of the pairs of papers that consistently cited adjacently

Practically, the method is composed of two important computational efforts. The first (step 1-4 in Figure 2) consists in generating a list of "potential twins" and does not require observing the actual text of the paper. Instead, it focuses on information generally available in publication databases such as the authors' names, the publication year and the references list. The second large computational effort (step 5 in Figure 2) requires an analysis of the text of a large number of citing papers in order to test whether the co-citations are or are not systematically adjacent.

[Insert Figure 2 about here]

We started by building a dataset of all the articles published between 2000 and 2010 by the 15 journals with the highest impact factor in 2009 from ISI Web of Science[3]. The 15 journals considered were the following: Nature, Science, Cell, New England Journal of Medicine, JAMA, Lancet, CA: A Cancer Journal for Clinicians, Nature Genetics, Nature Materials, Nature Medicine, Nature Immunology, Nature Nanotechnology, Nature Biotechnology, Cancer Cell and

[3] I excluded review journals because review articles contain less information for my purpose since they tend to cite a very large number of articles at the same time.

Cell Stem Cell. The analysis focused on the information available in the reference section of these 42,106 articles.

41,008 articles had a total of 1,294,357 references, or 744,583 unique references. Unfortunately, reference data in Web of Science only includes information about the first author, the year, the journal and sometimes the volume, page and DOI (Digital Object Identifier) of the paper. It was therefore necessary to build a web crawler that could get more information online such as the complete list of authors and whether or not the referred paper is a journal article or another type of document. Pubmed and Crossref were used for that purpose. Out of the 744,583 references, 142,509 (or19%) were excluded: 7,210 because the referenced document has not been published, 14,682 because the referred publication was not a journal article, 5,656 because of missing information on the author or on the DOI or Pubmed ID, and 114,961 could simply not be retrieved neither in Pubmed or in Crossref.

The remaining 602,134 referenced journal articles were then assigned a unique identifier based on the author list, the journal, and the volume and page of each of them. Using this reference data, a dataset of pairs of references was built. In order to be considered, a pair had to meet the following criteria:

- The two papers are cited at least once by a common third paper
- The two papers were published at most 1 calendar year apart
- The two papers have no author in common

17,050,914 pairs were co-cited at least once. Out of them, 13,212,649 were more than a calendar year apart and were therefore excluded. 297,802 additional pairs were excluded because they had an author in common.

Finally, out of the 3,540,463 pairs left, 3,091,046 had one or both references that were cited fewer than five times by our 42,106 citing articles. We decided to exclude them in order to be able to study forward co-citation frequencies on a large enough sample of citations. We were left with 449,417 pairs providing a sample of potential "paper twins" made out of 597,306 unique journal articles co-cited in 26,798 articles.

The oldest–and one of the most popular–measure of co-citation proximity is a set-theoretic measure consisting in the fraction of the intersection of the two sets of citations divided by their union (see figure 3). It is often called Jaccard index and is expressed by the following formula:

$$S(i,j) = \frac{coc(i,j)}{cit(i) + cit(j) - coc(i,j)}$$

where $coc(i,j)$ is the intersection between $cit(i)$the citations of $publication_i$ and $cit(j)$ the citations of $publication_j$. In the denominator, $cit(i) + cit(j) - coc(i,j)$ is the union of both sets of citations.

[Insert Figure 3 about here]

Most of our 449,417 pairs of journal articles had a very small Jaccard index (Figure 4). The average of the index is 7.3%. These pairs are unlikely to be simultaneous discoveries. Based on Cozzens's results, we defined as potential twins only those pairs that had a Jaccard index superior to 50% (these are paper pairs situated in the right column of the table in Figure 1). In the case of the opiate receptor, Cozzens finds that 269 out of the 510 citations to at least one of the four discovery articles did not include any citations to the other co-discovering teams, suggesting a co-citation rate of only 47% on average for a discovery that is broadly considered a simultaneous discovery. Using a Jaccard index of 50% as our threshold, we obtained 2,320 paper pairs. This threshold is very conservative and therefore likely to involve a high number of type 1 errors (false negative) but few type 2 errors (false positives). This conservative approach was chosen because our goal here is not to uncover every instance of simultaneous discovery. Rather, we are aiming at generating a sound list of "paper twins." In the future we are hoping that this method will be improved so as to decrease the amount of type 1 errors while keeping the number of type 2 errors low.

[Insert Figure 4 about here]

In step 5, we ran a parsing algorithm on our papers co-citing each of these candidate twins in order to distinguish those pairs that are consistently cited adjacently—i.e. simultaneous discoveries—from those who are cited in the same literature but not in the same parenthesis—i.e.

distinct but complementary ideas. Due to issues with the formatting of references in the text of the paper, the algorithm could analyze 3 co-citing publications or more for 1,825 pairs. Figure 5 presents the percentage of forward citing papers in which the 1,825 pairs were co-cited adjacently. Confirming that our 50% Jaccard Index threshold is very conservative, we find that a large number of the 1,825 pairs are consistently cited in the same parenthesis. In fact, 720 pairs of papers were co-cited adjacently in 100% of the papers that co-cited them. In order to remain conservative, we consider that only these pairs are instances of "paper twins."

[Insert Figure 5 about here]

2.4. A DATASET OF PAPER TWINS

2.4.1. Robustness Analysis

Prior theory would suggest that our method is conservative and should generate a list of simultaneous discoveries. In order to test this prediction empirically, we devised three different tests.

First, we examined the number of months separating the publications of two twin papers. As noted above, the algorithm did not match articles on simultaneity beyond ensuring that the difference between the calendar years of publications was not greater than one. If our alleged paper twins were not really the same, one would expect them to be on average six months apart or more. The time lag between discovery and publication is likely to vary across papers and is likely to add considerable noise in the data since some papers might be quickly published whereas their twins might be rejected by a few journals before they finally get accepted. Variance might depend on many factors including the journal to which the submission was made, its editor, and the reviewers that were selected. Strikingly, despite this noise, our 720 paper twins were published on average only 1.8 months apart—a lag considerably shorter than the average time between paper submission and publication (assuming there is no rejection). In fact, 373 pairs of twins were published the exact same month and 267 of them were published in the same issue of the same journal. The distribution of publication month difference for the 720 twins in the dataset is shown in Figure 6.

[Insert Figure 6 about here]

Second, we examine semantic similarity between twin papers. This measure is likely to be noisy. Indeed, as in Kuhn's case of the discovery of energy conservation or in Cozzens's opiate receptor case, the same discovery is commonly simultaneously made by individuals working in different disciplines. Semantic similarity between twins in these cases is likely to be moderate at best. By definition, semantic closeness is more likely to provide insights about the authors' language and approach rather than about their findings, which they routinely call with different names at first. Still, this measure can provide some reassurance that the twin papers use at least similar language. In order to test for semantic similarity, we used the Pubmed related citation algorithm. While not all the 720 twins in the dataset are disclosing life science discoveries, most are. We could find the two twin papers in Pubmed for 93% of our data—i.e. 669 twins. For each paper, the algorithm ranks other papers that are semantically related, starting with the semantically closest paper. The results are shown in Figure 7. Pubmed ranks two papers of the same twin right next to each other in 42% of observations. The rank difference is inferior to 10 for 90% of the twins[4].

[Insert Figure 7 about here]

Finally, we also collected the opinion of the discoverers themselves. We selected randomly 10 discoverers and asked them to describe the discovery process. Nine of them told us about the twin paper(s) without us asking. After we asked the tenth person why he did not mention the twin paper, he asserted angrily that he deserved all the credit and that his idea had been stolen. Of course, the fact that two teams have published twin papers around the same time does not mean that they conducted the exact same experiment or that they interpreted their results in the exact same way. Our interviews reveal that differences sometimes exist in the approach taken, with the instrument used, in the number of robustness checks that the teams had realized, or concerning the interpretation of the results. While all of our interviewees acknowledged the existence of the paper twin, a few of them highlighted that their paper was superior in some way or that they got there first. The validity of twin papers also does not mean

[4] Rank difference calculated after dropping articles that are published more than a calendar year apart and that have an author in common

that both discoveries were independent of one another. In fact, interviews with discoverers uncovered a number of cases in which one team accused the other of idea theft. However, the fact that the community of experts ruled that credit ought to be shared between two teams does indicate (1) that both teams are widely believed to have had the capability to make the discovery and (2) that each team has provided convincing evidence supporting their claim to priority.

2.4.2. Dataset Description

The 720 paper twins are composed of 1,246 unique papers. A few papers are part of more than one pairs. We observe 578 unique discoveries including 505 twins, 63 triplets, 12 quadruplets and 1 quintuplet.

The journals in which the papers were published as well as the subfield to which ISI assigned them appear in Figure 8. Most of our papers were published in prominent journals such as *Nature* (254 papers), *Science* (136 papers) and *Cell* (130 papers). The list of scientific publications in which the 1,246 papers were published is long, however, and includes 103 journals. The prominence of the journals in which our twins were published should not come as a surprise. Because it is based on citation patterns, our approach excluded poorly cited papers. Our dataset is mostly composed of life science discoveries. Out of the 790 papers to which ISI could assign a subfield,[5] 96% were in life sciences. Thirty-four papers were assigned to a subfield relating to physics and materials science. In line with the apparent prevalence of life science simultaneous discoveries, 1,226 of our papers are listed in Pubmed. This amounts to 98% of our twin papers. The extent to which this result indicates that simultaneous discoveries are particularly common in life sciences is unclear, however. Indeed, it should be noted that the journals from which our observations were drawn publish primarily life sciences results. We would have certainly drawn a different list of simultaneous discoveries, had we used the same algorithm on a different set of journals.

[Insert Figure 8 and 9 about here]

[5] 457 of our papers appeared in a multidisciplinary journal such as Nature and Science. For these journals, ISI does not assign a subfield to the paper.

The descriptive statistics for the dataset of scientific papers are displayed in Table 1 and Figure 9 shows the distribution of the papers by year of publication. The 1,246 papers were published between 1970 and 2009. However, the pair of 1970 twins is an outlier in the dataset. The next oldest twin papers date from 1988 and the large majority of our observations (99%) appear between 1992 and 2009. The average publication year for our papers is 2001. 96% of our papers were published by academics. The corresponding author has an academic affiliation in 1195 instances and a private sector affiliation in 51 cases. Out of the 720 twins, 664 involve two academic teams, 49 involve one team from academia and the other from industry, and 7 paper twins were both published by firms. The 1,246 papers stem from 985 different organizations. The most common addresses are University of California (79 papers), Harvard University (65 papers), University of Texas (40 papers) and MIT (27 papers). The average paper in our dataset involved 7.5 authors and 3.7 addresses but the distribution of authorship (and addresses) is highly skewed with one paper involving 216 authors and 73 addresses. Most teams were based in the US. 61% of our papers originated from a US address, 26% from a European address and 6% from Japan. Within the US, the papers originated primarily from Massachusetts (21%), California (20%), New York (12%), Maryland (7%) and Texas (7%).

[Insert Table 1 about here]

2.4.3. Example

The discovery of the importance of CD4+ T cells for secondary expansion and memory of CD8 T cells provides a compelling example of the type of observations present in our dataset. This specific discovery was an instance of "paper triplets" since it involved three teams. In our data, this triplet appears as three pairs of paper twins. Like most of our observations, this case involved academic teams working in life sciences in the US and having published their work in top-tier journals. In this case, the three teams were based in the La Jolla Institute for Allergy and Immunology, University of Pennsylvania, and University of Washington in Seattle respectively. This triplet involves two papers that appeared in the same issue of the same journal, here *Science*. This is one of 267 twins published back-to-back in the dataset. The *Nature* paper was published 2 months before the two *Science* papers. As we saw earlier, this time lag is typical and the average month difference between the publications of two twins in the dataset is 1.8 months. Importantly, the two *Science* papers were both sent out for publication before the *Nature* paper was published. The Janssen et al. paper was first sent for publication on December 3 2002 and was published on February 20 2003. The two *Science* papers were respectively sent on January 13 2003 and on February 11 2003 and were published back-to-back in the 11 April 2003 issue of the journal. The titles and abstracts of the three papers are the following:

Janssen et al. (February 2003) "CD4+ T cells are required for secondary expansion and memory in CD8+ T lymphocytes" ***Nature***
A long-standing paradox in cellular immunology concerns the conditional requirement for CD4+ T-helper (TH) cells in the priming of cytotoxic CD8+ T lymphocyte (CTL) responses in vivo. Whereas CTL responses against certain viruses can be primed in the absence of CD4+ T cells, others, such as those mediated through 'cross-priming' by host antigen-presenting cells, are dependent on TH cells. A clearer understanding of the contribution of TH cells to CTL development has been hampered by the fact that most TH-independent responses have been demonstrated ex vivo as primary cytotoxic effectors, whereas TH-dependent responses generally require secondary in vitro re-stimulation for their detection. Here, we have monitored the primary and secondary responses of TH-dependent and TH-independent CTLs and find in both cases that CD4+ T cells are dispensable for primary expansion of CD8+ T cells and their differentiation into cytotoxic effectors. However, secondary CTL expansion (that is, a secondary response upon re-encounter with antigen) is wholly dependent on the presence of TH cells during, but not after, priming. Our results demonstrate that T-cell help is 'programmed' into CD8+ T cells during priming, conferring on these cells a hallmark of immune response memory: the capacity for functional expansion on re-encounter with antigen.

Shedlock et al. (April 2003) "Requirement for CD4 T Cell Help in Generating Functional CD8 T Cell Memory." ***Science***
Although primary CD8 responses to acute infections are independent of CD4 help, it is unknown whether a similar situation applies to secondary responses. We show that depletion of CD4 cells during the recall response has minimal effect, whereas depletion during the priming phase leads to reduced responses by memory CD8 cells to reinfection. Memory CD8 cells generated in CD4+/+ mice responded normally when transferred into CD4/ hosts, whereas memory CD8 cells generated in CD4/ mice mounted defective recall responses in CD4+/+ adoptive hosts. These results demonstrate a previously undescribed role for CD4 help in the development of functional CD8 memory.

Sun et al. (April 2003) "Defective CD8 T Cell Memory Following Acute Infection Without CD4 T Cell Help." ***Science***

> The CD8+ cytotoxic T cell response to pathogens is thought to be CD4+ helper T cell independent because infectious agents provide their own inflammatory signals. Mice that lack CD4+ T cells mount a primary CD8 response to Listeria monocytogenes equal to that of wild-type mice and rapidly clear the infection. However, protective memory to a challenge is gradually lost in the former animals. Memory CD8+ T cells from normal mice can respond rapidly, but memory CD8+ T cells that are generated without CD4 help are defective in their ability to respond to secondary encounters with antigen. The results highlight a previously undescribed role for CD4 help in promoting protective CD8 memory development.

As apparent from these excerpts, at the time of the discovery, it was known that CD4 cells were sometimes but not always important to trigger a response to viruses from CD8 cells. Yet the circumstances under which CD4 cells were required were unknown. The three teams have used mice models to show that while CD4 cells are not necessary for primary expansion of CD8 cells (i.e. so that the CD8 cells respond when they encounter the virus for the first time), these CD4 cells are necessary for a secondary response (i.e. so that the CD8 cell respond when they re-encounter the same virus). Hence, CD4 cells have an important role in CD8 memory development.

2.5. DISCUSSION

This paper attempts to unleash the potential of simultaneous discoveries as a research tool. Our proposed contribution is threefold. In the first place, we reviewed the literature on scientific multiples in order to describe the theoretical promise of this tool. Next, using the literature on the social construction of science, we described a way out of traditional debates about the sufficient level of similarity of distinct discoveries. Lastly, we implemented and tested this method with an algorithm that enables the generation of large datasets of recent simultaneous discoveries. We described our dataset, which includes 578 discoveries made by 1,246 teams of scientists.

For decades, simultaneous discoveries have been considered a strategic research setting, holding great promise for a number of purposes. First, their frequency provides striking evidence of the fact that creative insights, although novel, might not necessarily be unique. The potential redundancy of new ideas in turn highlights the importance of macro-level rather than local drivers of scientific creativity. If individuals located in San Diego, Seattle and Philadelphia can see the same idea around the same time, it is probably because they are standing on the same shoulders—and are looking in the same direction. While a number of studies of scientific creativity have emphasized the role of more micro variables such as individual ability or colleagues' quality (Azoulay, Graff Zivin, and Wang 2010), the evidence presented here that simultaneous discoveries are frequent occurrences should serve as a reminder of the important role played by more macro-level drivers.

Second, our approach provides important insights about the process of social construction of science. The method presented in this paper is based on the inference that teams of scientists that make the same discovery around the same time will share the credit for that discovery—and that this will be visible in the citation patterns of scientific papers. By implementing this insight into an algorithm that successfully built a list of simultaneous discoveries, we have tested and validated our proposition inspired from Cozzens's qualitative analysis. Furthermore, the process of credit sharing in science is often a contentious one. In our interviews, we have uncovered several instances in which scientists accused one another of espionage, idea theft and scientific fraud. Therefore, paper twins constitute a potentially rich setting to explore the norms and practices through which policing and conflict management take place in science.

Third, simultaneous discoveries constitute twins of new scientific knowledge. Twin studies (of humans) have been used for decades in genetics in order to disentangle the impact of the genes from the impact of the environment on individuals' characteristics and behaviors. We propose that the same approach can be used with these "knowledge twins." Using paper twins, it is possible to observe the same knowledge that emerged at the same time in two different environments. Through qualitative or quantitative methods, it is therefore possible to identify the influence of environmental variables on knowledge dissemination and commercialization. We are currently pursuing this line of research and are investigating the impact of the academic environment and of geographic isolation on science-based invention (Bikard 2012).

The method proposed here is not without limitation. Our algorithm uses very conservative criteria. The explicit goal was to limit the number of false positive at the cost of excluding numerous false negatives. Our interviews have uncovered a number of cases of simultaneous discoveries that did not appear in our dataset. At the same time, we cannot be sure that all our observations were really simultaneous discoveries. Our results are not totally exempt from the criticism concerning discovery identity that was raised in prior literature. Although our papers are broadly regarded as making the same contribution by the community of experts, we have found instances in which the discovery was made in different animal models, or using different instruments, or in which some of the authors had misinterpreted their results, or one author might have gone further in his or her discovery than the other ones.

In his landmark 1961 article on scientific multiples, Merton quipped that the idea of independent discovery is confirmed by its own history since it has been "periodically rediscovered over a span of centuries" (Merton 1961, 475). This essay is not a rediscovery of the idea of simultaneous discoveries. Rather, it is an attempt to answer Merton's fifty-year-old call to recognize that these events offer tremendous research opportunities for social scientists. By describing some of these opportunities and presenting a method to acquire the data to explore them, our hope is that we are making it easier to stand on Merton's shoulders.

2.6. REFERENCES

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang. 2010. "Superstar Extinction." *The Quarterly journal of economics* 125 (2): 549–589.

Biagioli, Mario, and P. Galison. 2003. *Scientific Authorship: Credit and Intellectual Property in Science.* Routledge New York.

Bikard, Michaël. 2012. "Is Knowledge Trapped Inside the Ivory Tower? Technology Spawning and the Genesis of New Science-Based Inventions." *MIT Sloan Working Paper.*

Brannigan, Augustine. 1981. *The Social Basis of Scientific Discoveries.* Cambridge Univ Pr.

Brannigan, Augustine, and Richard A. Wanner. 1983a. "Multiple Discoveries in Science: A Test of the Communication Theory." *The Canadian Journal of Sociology / Cahiers Canadiens De Sociologie* 8 (2): 135–151.

———. 1983b. "Historical Distributions of Multiple Discoveries and Theories of Scientific Change." *Social Studies of Science* 13 (3) (August): 417–435.

Callahan, Alison, Stephen Hockema, and Gunther Eysenbach. 2010. "Contextual Cocitation: Augmenting Cocitation Analysis and Its Applications." *Journal of the American Society for Information Science and Technology*: n/a–n/a. doi:10.1002/asi.21313.

Castillo, M. 2008. "Scientific Multiples, Neuroradiology, and the American Journal of Neuroradiology." *AJNR Am J Neuroradiol* 29 (8) (September 1): 1423–1424. doi:10.3174/ajnr.A1206.

Collins, Harry M. 1992. *Changing Order: Replication and Induction in Scientific Practice.* University of Chicago Press.

Constant, Edward W. 1978. "On the Diversity and Co-Evolution of Technological Multiples: Steam Turbines and Pelton Water Wheels." *Social Studies of Science* 8 (2) (May): 183–210.

Cozzens, Susan E. 1989. *Social Control and Multiple Discovery in Science: The Opiate Receptor Case.* State Univ of New York Pr.

Elkana, Y. 1971. "«The Conservation of Energy: a Case of Simultaneous Discovery?»." *Archives Internationales d'Histoire Des Sciences* 24: 31–60.

———. 1974. *The Discovery of the Conservation of Energy.* Harvard University Press.

Giles, C. L, K. D Bollacker, and S. Lawrence. 1998. "CiteSeer: An Automatic Citation Indexing System." In *Proceedings of the Third ACM Conference on Digital Libraries*, 89–98.

Gipp, B., and J. Beel. 2009. "Citation Proximity Analysis (CPA)-A New Approach for Identifying Related Work Based on Co-Citation Analysis." In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, 571–575.

Hagstrom, Warren O. 1974. “Competition in Science.” *American Sociological Review* 39 (1) (February 1): 1–18. doi:10.2307/2094272.

Hounshell, David A. 1975. “Elisha Gray and the Telephone: On the Disadvantages of Being an Expert.” *Technology and Culture* 16 (2) (April): 133–161.

———. 1981. “Two Paths to the Telephone.” *Scientific American* 244 (1): 156–164.

Kessler, M. M. 1963. “Bibliographic Coupling Between Scientific Papers.” *American Documentation* 14 (1) (January): 10–25. doi:10.1002/asi.5090140103.

Kroeber, A. L. 1917. “The Superorganic.” *American Anthropologist* 19 (2). New Series (June): 163–213.

Kuhn, Thomas S. 1969. “Energy Conservation as an Example of Simultaneous Discovery.” In *Critical Problems In The History Of Science: Proceedings Of The Institute For The History Of Science, 1957*, 321.

Latour, B., and S. Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts.* Princeton University Press.

Lawrence, Peter A. 2003. “The Politics of Publication.” *Nature* 422 (6929) (March 20): 259–261. doi:10.1038/422259a.

Leydesdorff, L. 1987. “Various Methods for the Mapping of Science.” *Scientometrics* 11 (5-6) (May): 295–324. doi:10.1007/BF02279351.

De Marchi, Neil. 1995. “Comment on Niehans, ‘Multiple Discoveries’.” *European Journal of the History of Economic Thought* 2 (2): 275. doi:Article.

Marshakova, I. V. 1973. “System of Document Connections Based on References.” *Nauchno-Teknicheskaia Informatsiia* 2 (1): 3–8.

Marshall, Eliot. 2002. “DNA Sequencer Protests Being Scooped With His Own Data.” *Science* 295 (5558) (February 15): 1206 –1207. doi:10.1126/science.295.5558.1206.

Merton, Robert K. 1961. “Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science.” *Proceedings of the American Philosophical Society* 105 (5) (October 13): 470–486.

———. 1963. “Resistance to the Systematic Study of Multiple Discoveries in Science.” *European Journal of Sociology* 4 (August): 237–282.

———. 1968. *Social Theory and Social Structure.* Simon and Schuster.

Merz, Jon F, and Michelle R Henry. 2004. “The Prevalence of Patent Interferences in Gene Technology.” *Nat Biotech* 22 (2) (February): 153–154. doi:10.1038/nbt0204-153.

Moravcsik, Michael J., and Poovanalingam Murugesan. 1975. “Some Results on the Function and Quality of Citations.” *Social Studies of Science* 5 (1) (February 1): 86–92.

Narin, F. 1976. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Computer Horizons Cherry Hill, NJ.

Niehans, Jurg. 1995. “Multiple Discoveries in Economic Theory.” *European Journal of the History of Economic Thought* 2 (1): 1. doi:Article.

Ogburn, William F., and Dorothy Thomas. 1922. “Are Inventions Inevitable? A Note on Social Evolution.” *Political Science Quarterly* 37 (1) (March): 83–98.

Patinkin, Don. 1983. “Multiple Discoveries and the Central Message.” *The American Journal of Sociology* 89 (2) (September): 306–323.

Saudek, Christopher D. 2003. “2002 Presidential Address: A Tide in the Affairs of Medicine.” *Diabetes Care* 26 (2) (February 1): 520 –525. doi:10.2337/diacare.26.2.520.

Schmookler, J. 1966. *Invention and Economic Growth*. Harvard University Press Cambridge, MA.

Simonton, Dean K. 1978. “Independent Discovery in Science and Technology: A Closer Look at the Poisson Distribution.” *Social Studies of Science* 8 (4) (November 1): 521 –532. doi:10.1177/030631277800800405.

———. 1979. “Multiple Discovery and Invention: Zeitgeist, Genius, or Chance?” *Journal of Personality and Social Psychology* 37 (9) (September): 1603–1616. doi:10.1037/0022-3514.37.9.1603.

———. 1986. “Multiple Discovery: Some Monte Carlo Simulations and Gedanken Experiments.” *Scientometrics* 9 (5-6) (May): 269–280. doi:10.1007/BF02017248.

———. 1999. *Origins of Genius: Darwinian Perspectives on Creativity*. Oxford University Press, USA.

Small, Henry. 1973. “Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents.” *Journal of the American Society for Information Science* 24 (4) (July): 265–269.

———. 1978. “Cited Documents as Concept Symbols.” *Social Studies of Science* 8 (3) (August): 327–340.

———. 1998. “A General Framework for Creating Large-scale Maps of Science in Two or Three Dimensions: The SciViz System.” *Scientometrics* 41 (1): 125–133.

de Solla Price, D. J. 1965. *Little Science, Big Science... and Beyond*. Columbia University Press.

Tran, Nam, Pedro Alves, Shuangge Ma, and Michael Krauthammer. 2009. “Enriching PubMed Related Article Search with Sentence Level Co-citations” 2009: 650–654.

Troyer, J. 2001. “In the Beginning: The Multiple Discovery of the First Hormone Herbicides.” *Weed Science* 49 (2): 290–297.

Tsai, Wenpin, and Chia-hung Wu. 2010. "Knowledge Combination: A Cocitation Analysis." *Academy of Management Journal*, June.

Voss, C. A. 1984. "Multiple Independent Invention and the Process of Technological Innovation." *Technovation* 2 (3) (June): 169–184. doi:10.1016/0166-4972(84)90002-6.

White, Howard D., and Belver C. Griffith. 1981. "Author Cocitation: A Literature Measure of Intellectual Structure." *Journal of the American Society for Information Science* 32 (3) (May): 163–171. doi:10.1002/asi.4630320302.

2.7. TABLES & FIGURES

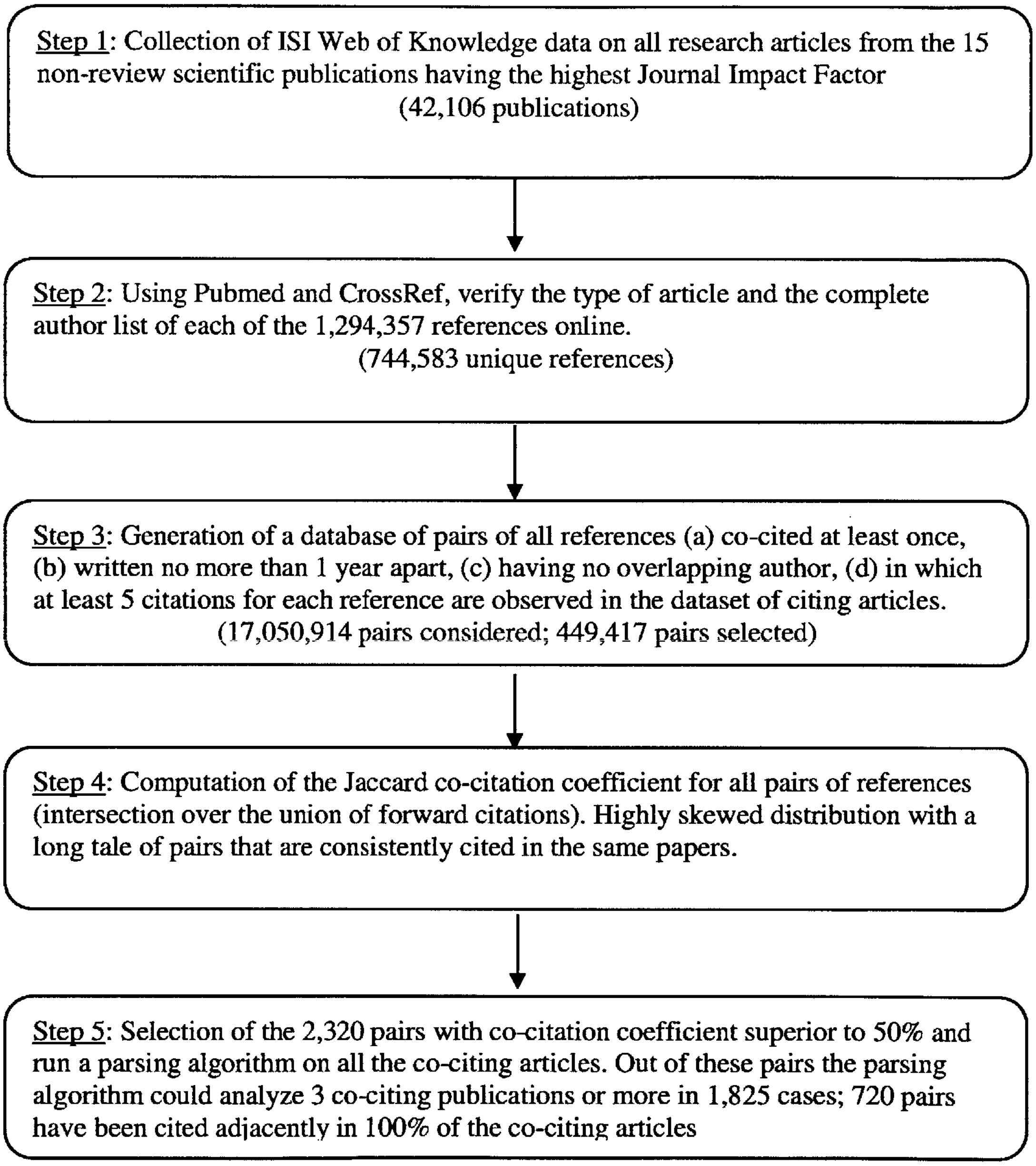
Table 1. Main Descriptive Statistics: the 1,246 papers

Variable	Mean	Std. Dev.	Min	Max
Paper Publication Year	2000.95	4.22	1970	2009
Number of Authors	7.47	8.02	1	216
Number of Addresses	3.7	4.38	0	73
Academic Paper	0.96	0.2	0	1
US address	0.61	0.44	0	1

Figure 1: The two dimensions of co-citation (articles written about the same year and sharing no author)

		COCITATION	
		Low	High
COCITATION	Low	Unrelated publications	Distinct but complementary ideas
	High	Several good sources for a specific idea / discovery	Multiple discovery

Figure 2: Diagram of the method used to generate the list of paper twins



Step 1: Collection of ISI Web of Knowledge data on all research articles from the 15 non-review scientific publications having the highest Journal Impact Factor
(42,106 publications)

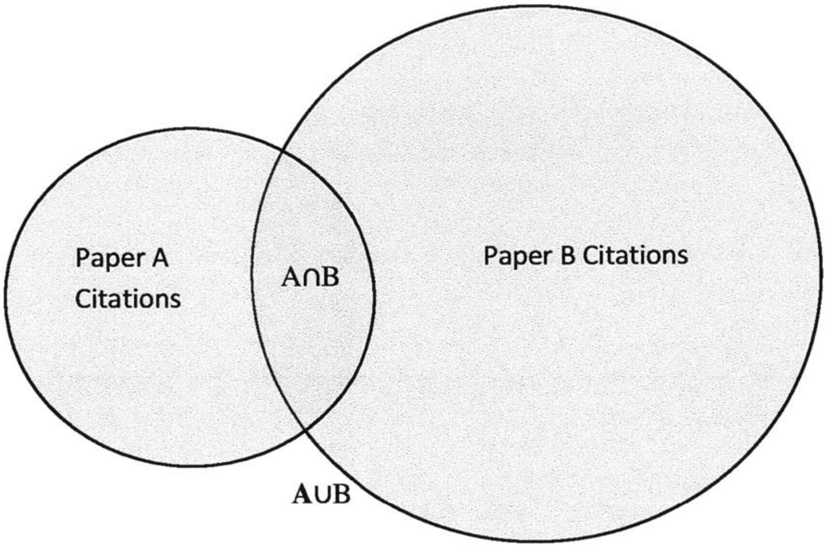
Step 2: Using Pubmed and CrossRef, verify the type of article and the complete author list of each of the 1,294,357 references online.
(744,583 unique references)

Step 3: Generation of a database of pairs of all references (a) co-cited at least once, (b) written no more than 1 year apart, (c) having no overlapping author, (d) in which at least 5 citations for each reference are observed in the dataset of citing articles.
(17,050,914 pairs considered; 449,417 pairs selected)

Step 4: Computation of the Jaccard co-citation coefficient for all pairs of references (intersection over the union of forward citations). Highly skewed distribution with a long tale of pairs that are consistently cited in the same papers.

Step 5: Selection of the 2,320 pairs with co-citation coefficient superior to 50% and run a parsing algorithm on all the co-citing articles. Out of these pairs the parsing algorithm could analyze 3 co-citing publications or more in 1,825 cases; 720 pairs have been cited adjacently in 100% of the co-citing articles

Figure 3: Strength of association between co-cited papers



$$\text{Jaccard co-citation Index} = \frac{\mathbf{A \cap B}}{\mathbf{A \cup B}}$$

Figure 4: Distribution of the pairs by Jaccard index value

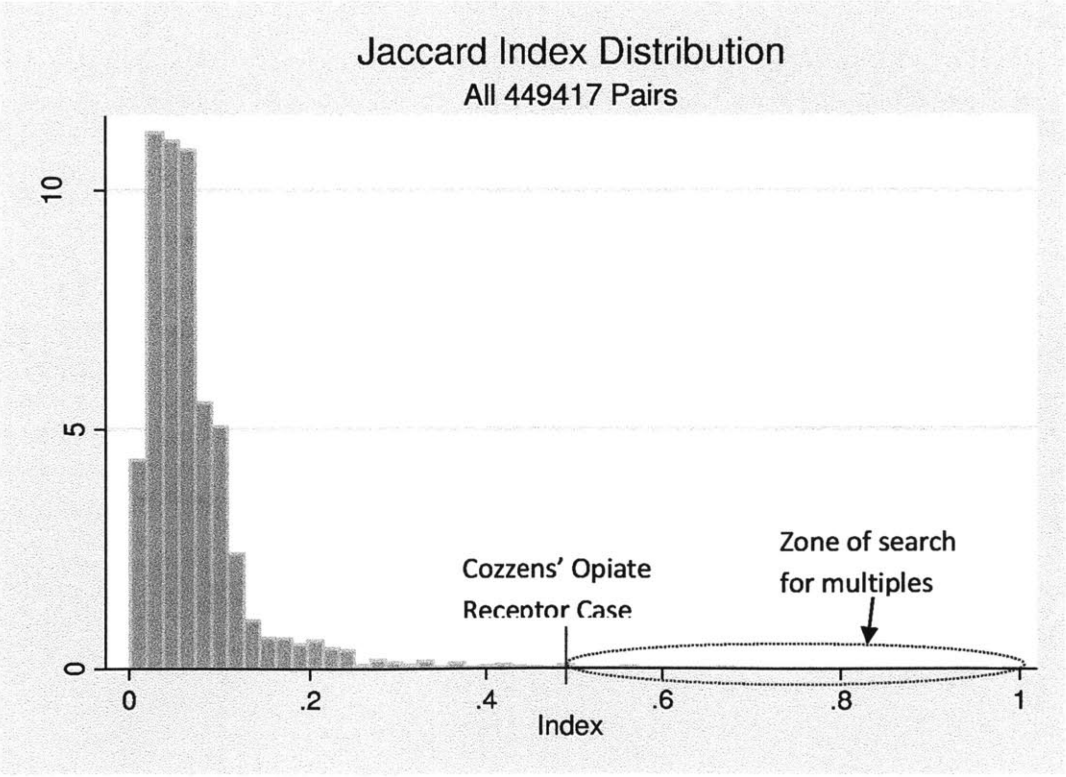


Figure 5. Adjacent Citation Frequency Among Co-Cited Paper pairs

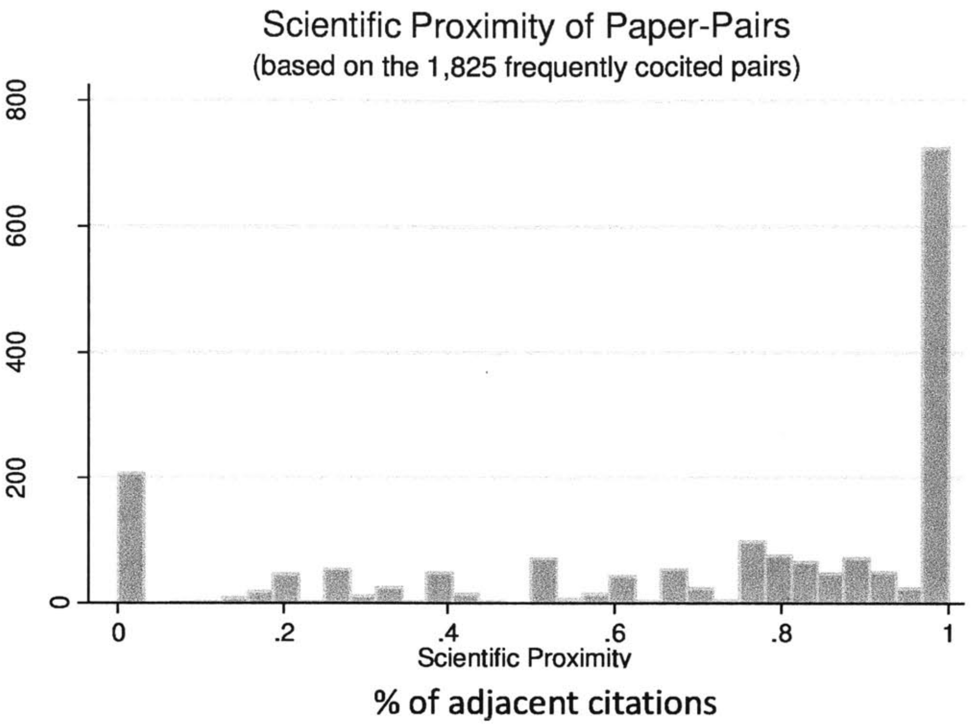


Figure 6. Number of months separating the publication of the 720 twin papers

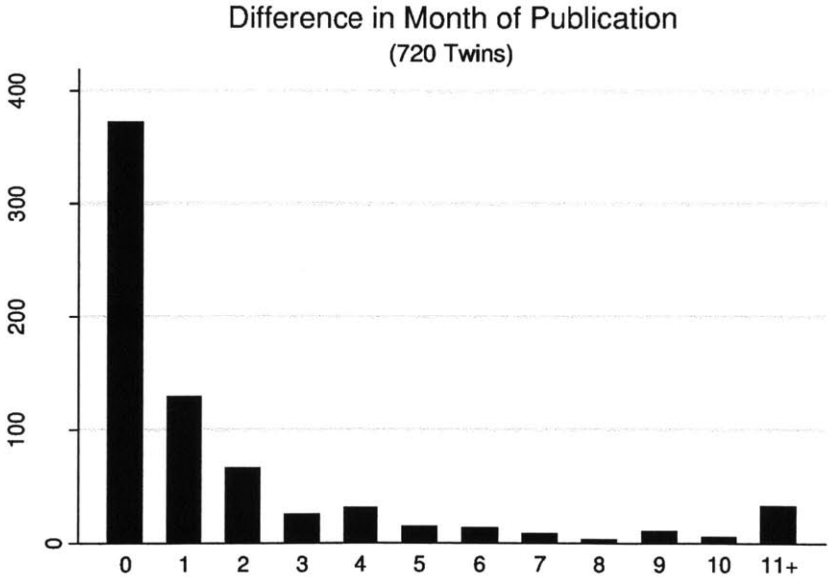
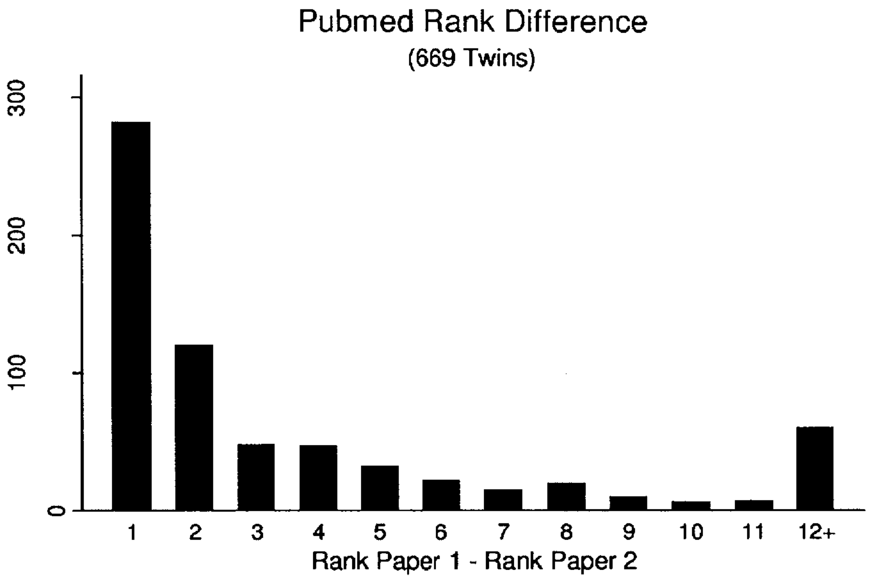
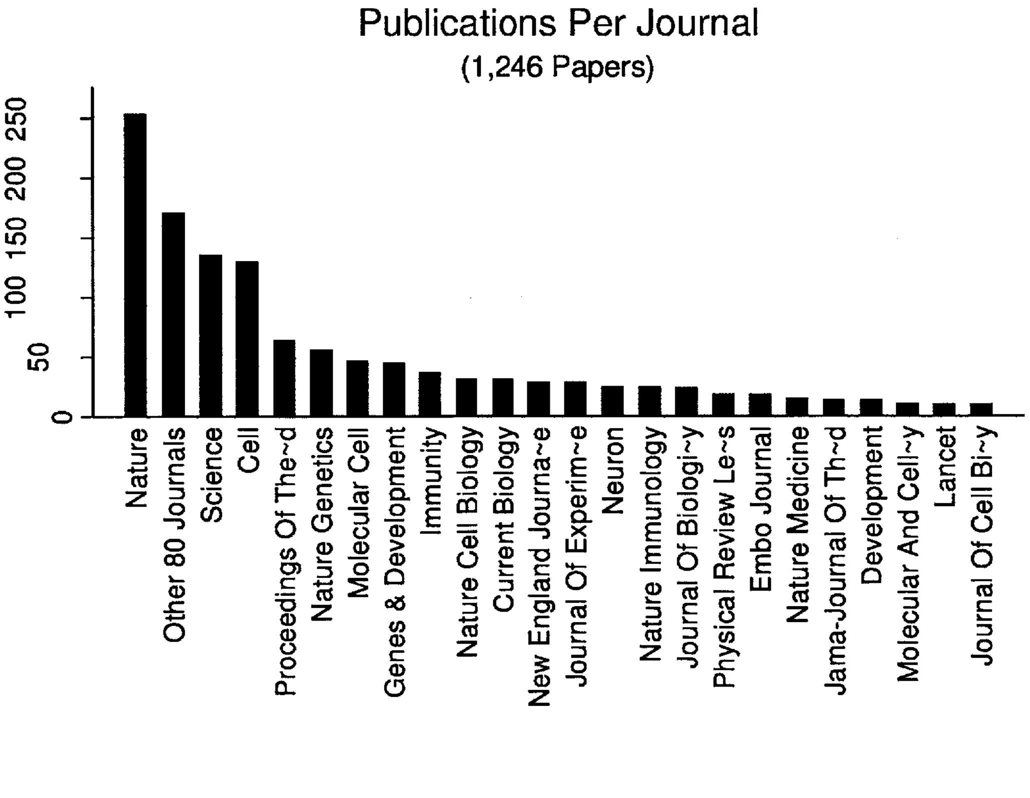


Figure 7. Pubmed rank separating the publication of the 669 twin papers



Note: Excludes paper written more than one year apart

Figure 8. Journals and Subfields in which the 1,246 papers were published



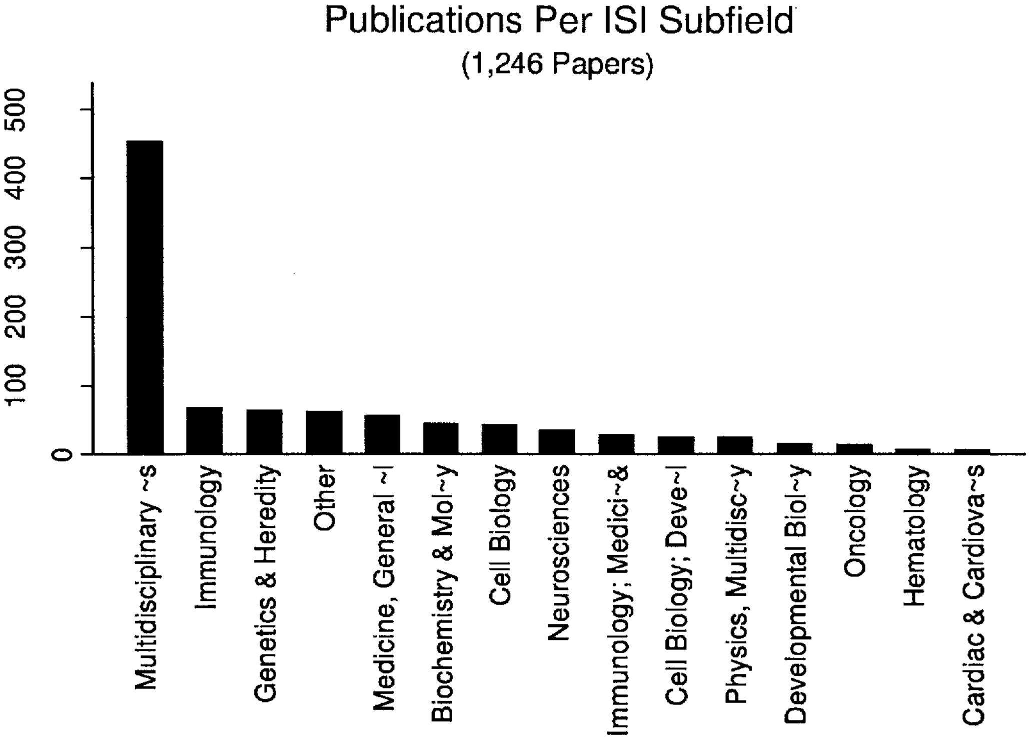
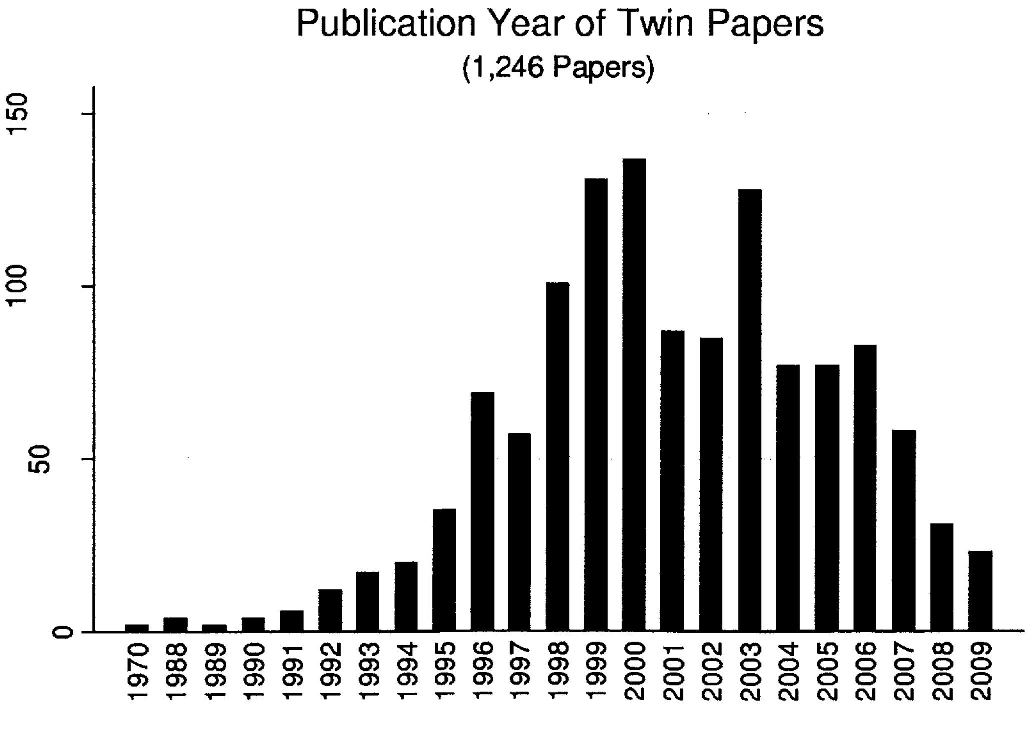


Figure 9. Number of twins by calendar year of publication



Is Knowledge Trapped Inside the Ivory Tower? Technology Spawning and the Genesis of New Science-Based Inventions

3.1. INTRODUCTION

A large and growing literature has argued that scientific research has a positive impact on technological innovation whether it is conducted in firms (Cohen and Levinthal 1990; Gambardella 1992; Gittelman and Kogut 2003) or in universities (Mansfield 1995; Henderson, Jaffe, and Trajtenberg 1998; Zucker, Darby, and Armstrong 2002). However, much less is known of the conditions under which the scientific knowledge produced in firms and universities is recombined into novel technologies. Historical examples reveal that while this translation can be very rapid in some circumstances, scientific knowledge can also remain unexploited for decades before an inventor finally uses it—if at all[6] (Rosenberg 1994; Mokyr 2002). This paper attempts to fill this gap by examining the factors that lead to the translation of scientific discoveries into new technologies, a process we term "technology spawning."

We examine two views of this process. In one view, technology spawning is driven by the ease of access to scientific knowledge. Easy access drives invention because it reduces its cost (Dasgupta and David 1994; Sorenson and Fleming 2004). Another view argues that technology spawning is driven by control over the produced knowledge. Invention requires large investments that will not be undertaken unless scientists can capture economic benefits from those investments (Jensen and Thursby 2001; Lach and Schankerman 2008). These two perspectives are not inherently contradictory. Ease of access to scientific knowledge might decrease the cost of invention, and, at the same time, control over the knowledge produced might encourage investment in this costly process.

[6] For instance, the Hellenistic civilization produced Ptolemaic astronomy but never used it for navigation, they understood optics but did not translate that knowledge into making binoculars or glasses (Mokyr 2002, 262)

These views do offer competing economic implications, however, concerning the type of discovery organization that encourages technology spawning. Those emphasizing access argue that universities might encourage science-based invention (Owen-Smith and Powell 2004; Furman et al. 2005). Not only is the diffusion of knowledge from academic labs a consequence of the educational mission of universities, it also results from the academic norms of sharing and openness (Merton 1973; Murray 2010). In contrast, those emphasizing control have suggested that firms foster technology spawning because industry scientists have clearer incentives and can draw on more resources to develop new technologies (Nelson 1959; Aghion, Dewatripont, and Stein 2008). This paper presents an empirical test of the relative impact of the academic and corporate discovery environments on the emergence of new technologies.

The challenge in exploring technology spawning empirically is considerable. The technological potential of a given piece of new scientific knowledge is always unobserved. Hence, measured rates of cumulative inventions might result from the environment in which a given discovery was made, but it might also be a consequence of the nature and promise of that discovery. For instance, new knowledge produced in universities is likely to be more fundamental on average than scientific discoveries made by firms. In order to examine the two views of technology spawning, it is crucial to account for the technological potential of the scientific discoveries. This paper reports a novel empirical strategy to tackle this challenge.

In the winter of 1999, two teams of scientists simultaneously discovered VR1 (vanilloid receptor-1), the receptor for the pain caused by excessive heat or capsaicin, the pungent component of chili peppers. The first team, led by Dr. John B Davis, sent its results to *Nature* on December 20, 1999 and the paper was published on May 11, 2000. The second team, led by Prof. David Julius, sent its results to *Science* on January 18, 2000 and the paper was published on April 14, 2000. The new knowledge had important implications for the development of pain therapeutics. Yet, both discoveries were made in very different organizations. Julius is an academic based at UC San Francisco. In contrast, Davis is an industrial scientist working at SmithKline Beecham. Simultaneous discoveries are a fascinating and relatively frequent phenomenon (Merton 1961). When the discoverers submit their findings for publication at almost the same time, two or more papers disclosing the same discovery can be accepted, thus leading to the publication of "paper-twins." Paper-twins are scientific articles that disclose the same underlying piece of knowledge. They are thus more than closely related or complementary discoveries. Rather, by embodying the same piece of knowledge that emerged in two distinct environments, paper-twins are a natural consequence of the duplication of effort in science, and a potentially rich setting to study the determinants of science-based invention.

The "experiment" afforded by the observation of discoveries occurring simultaneously in a university and in a firm will allow for a set of precise tests motivated by the two competing views on technology spawning. The citations of each twin paper in the patent literature provide a convenient (though noisy) measure of follow-on invention. We explore whether patents that build on the new knowledge are more likely to cite the academic paper or its twin from industry. In order to get further insight into the mechanism at play, we then distinguish between inventions patented by the discoverers themselves and inventions originating from third parties. Because it enables the observation of the *non-occurrence* of patents (or at least of patent-to-paper citations) that could have occurred, paper-twins are a setting particularly suited to investigating the impact of the research environment on follow-on invention.

The analysis centers on 39 simultaneous discoveries made by 90 teams and that involved at least one team from a university and another team from a firm. These 90 papers are cited in 533 patents, therefore allowing for a quantitative comparison of the two views on technology spawning. The chosen setting is narrower in scope than massive data-based efforts analyzing tens of thousands of academic publications and patents (Narin, Hamilton, and Olivastro 1997; Henderson, Jaffe, and Trajtenberg 1998) or large-scale survey data (Klevorick et al. 1995; Cohen, Nelson, and Walsh 2002) but larger than qualitative small-scale investigations (Colyvas et al. 2002). The results indicate that a type of discovery organization emphasizing control, such as firms, leads to higher rates of technology spawning than a type of organization emphasizing openness. A scientific publication originating from a firm is 20-30% more likely to be cited in follow-on patents than its academic twin. Confirming the importance of control, we find that discoverers working in industry generate far more follow-on patents than their co-discoverers in academia. Moreover, contrary to the idea that ease of access to scientific knowledge plays a crucial role, we find that inventors that did not take part in the discovery are significantly more likely to cite industry papers than their twins from within the "Ivory Tower."

3.2. SCIENCE-BASED CUMULATIVE INVENTION

3.2.1. Technology Spawning

We define technology spawning as the translation of scientific knowledge into new technologies. The idea that scientific knowledge fosters technological innovation is widespread and has received considerable empirical support (Jaffe 1989; Rosenberg and Nelson 1994; Mansfield 1995; Cohen, Nelson, and Walsh 2002). The correlate to this idea is naturally that the absence of scientific knowledge constrains the emergence of inventions. Each incremental addition of new knowledge has therefore the potential to "open doors hitherto closed" (Mokyr 2002, 9). Specifically, a scientific discovery is the addition of a new piece of knowledge to society's aggregate understanding of natural phenomena and regularities. Since technologies are instructions or devices enabling the purposeful manipulation of these regularities, new scientific knowledge can at times allow the spawning of new technologies. The new knowledge might provide guidance in the process of invention, thereby vastly decreasing its cost (Nelson 1982; Fleming and Sorenson 2004). At other times, scientific discoveries can be directly instantiated as new technologies (Stokes 1997; Murray 2002).

Yet, the creation of new scientific knowledge is not sufficient to ensure its development into novel inventions (Nelson 1959). Historians have uncovered numerous instances in which scientific knowledge could have led to the development of new technologies but did not. Mokyr writes: "Opening such doors does not guarantee that anyone will choose to walk through them." Surprisingly, outside the work of a few economic historians, prior research has not explored the conditions under which given pieces of new scientific knowledge might remain underutilized as compared to their technological potential. This omission has important consequences since new scientific knowledge might have a very different technological impact depending on the environment in which it emerges. Prior findings about the impact of scientific research on technological innovation might be biased by the unobserved ability of the chosen empirical setting to spawn new technologies.

The case of inhalation anesthesia—certainly one of the most important medical technologies ever developed—provides a fascinating illustration of the difficulties associated with technology spawning. Nearly half a century separates the first documented discovery of inhalation anesthesia and its development into a technology. As early as 1800, Humphry Davy used nitrous oxide to reduce pain he felt in his wisdom tooth and suggested that inhalation anesthesia could be used in surgery. However, the superintendent of the Pneumatic Institution[7] stopped short of developing this idea. In 1823, an English doctor called Henry Hill Hickman successfully operated on dogs, mice and rabbits after he had made them lose consciousness by inhalation of carbon dioxide. He wrote his findings in a pamphlet but the socially isolated doctor failed to interest the English Royal Society and the French Académie Royale de Médecine[8]. In 1844, Connecticut dentist Horace Wells successfully used nitrous oxide to reduce pain in tooth extraction. However, the demonstration of his technique at the Massachusetts General Hospital was only partially successful and the new technique was rejected. A Boston-based doctor, William Morton, was present at Wells' demonstration and endeavored to develop the idea using ether together with a device that he had had produced by a local instrument maker. On October 16 1846, a demonstration that would make history took place in the operating theater of the Massachusetts General Hospital. That morning, a 20-year-old was operated on for a tumor in the neck—and felt no pain. After forty-six years of independent re-discovery and failed development, the technology of inhalation anesthesia was finally born, marking a great step forward in the reduction of human suffering (Youngson 1979).

This paper explores empirically the conditions under which scientific discoveries with high technological potential are abandoned—or pursued. We focus on the impact of the organizational context of discovery on follow-on invention. Two streams of research have emphasized distinct economic drivers of technology spawning, both highlighting its high cost. Some researchers have emphasized the importance of access to the scientific knowledge, whereas others have stressed the importance of control over the research outcomes.

3.2.2. Access to the new scientific knowledge

[7] This medical research facility was established in Bristol, UK in 1799 to study of the use of gases in medicine.

[8] Other attempts are likely to have been made that were not publicized. For instance, Crawford Long, an American doctor working in Georgia, claimed in 1852 that he had used the inhalation of ether for surgery as early as 1842.

Because technological innovation is a process of recombination, access to scientific knowledge is often critical (Fleming and Sorenson 2004). Not all scientific discoveries are published (Moon 2011) and publication, when it occurs, is rarely sufficient (Murray and O'Mahony 2007). Follow-on innovation requires some understanding of how the new knowledge was developed as well as access to the necessary tools, materials, information and techniques.

Access to new scientific knowledge might be difficult for three reasons. First, knowledge tends to emerge in tacit form, making it highly personal and costly to transfer (Polanyi 1966). This "stickiness" (Von Hippel 1994) or "natural excludability" (Zucker, Darby, and Armstrong 2002) can be difficult to overcome for follow-on inventors. Second, scientific and technological communities tend to be distinct social networks (Murray 2002) shaped by different, sometimes conflicting, institutional logics (Merton 1973; Dasgupta and David 1994). The different sets of values and norms in both communities can create tensions in the process of turning scientific insights into new technologies (Argyres and Liebeskind 1998; Murray 2010). Third, the individuals who made the discovery and/or their organizations might have economic and strategic incentives to make access to the new knowledge difficult for others (Rosell and Agrawal 2009). Scientists at times refuse to share their knowledge with others whom they view as competitors. Similarly, patents can be used as "tollbooth[s] on the road to product development" (Heller and Eisenberg 1998, 699).

This difficulty in accessing scientific knowledge has important consequences. For instance, it is an important underlying mechanism to the geographic localization of knowledge flows (Jaffe, Trajtenberg, and Henderson 1993; Zucker, Darby, and Brewer 1998). Naturally, access to scientific knowledge is also a crucial strategic concern for innovative firms. A large literature has uncovered various mechanisms that innovators use to gain access, including investment in internal R&D (Cohen and Levinthal 1990; Rosenberg 1990), research collaborations (Arora and Gambardella 1994; Cockburn and Henderson 1998), and labor markets (Gittelman and Kogut 2003; Singh and Agrawal 2011). Access is therefore an important determinant of follow-on research and development. Factors increasing the cost of access, such as intellectual property, can decrease follow-on R&D (Murray and Stern 2007; Williams 2012). On the other hand, factors that decrease the access cost, such as biological resource centers, can increase follow-on R&D (Furman and Stern 2011).

Transposing this argument at the level of the knowledge-producing organization, a number of researchers have argued that scientific knowledge conducted in open organizations such as universities and other non-profit research institutions will generate more technologies than if it were produced in corporate laboratories. This view is probably best summarized in the following words: “since universities are in principle dedicated to the widespread dissemination of the results of their research, university spillovers are likely to be disproportionately large and may thus be disproportionately important” (Henderson, Jaffe, and Trajtenberg 1998, 119). Because the mission of universities is to spread new knowledge, ease of access—and thus technology spawning—should be a lot greater for the discoveries that were made in academic labs (Owen-Smith and Powell 2004; Furman et al. 2005).

3.2.3. Control over the new scientific knowledge

In contrast, another line of research has highlighted the importance of control rather than access. Because technology spawning is costly, it will not occur unless scientists have the ability to appropriate some return from their investment.

The emphasis on control in technology spawning is based on the premise that knowledge is easily stolen. The capture of economic rent based on new scientific knowledge is crucially dependent on the extent to which competitors will be able to use the new knowledge without having to incur its development cost. If the knowledge cannot be appropriated–or controlled—by its producer, the incentive for private investment in that knowledge will be weak. In order to prevent such knowledge theft and to preserve their competitive advantage, firms use a variety of strategies such as patenting (Arrow 1962) or secrecy (Nelson 1959). Control over the research produced is therefore an important driver of innovation. Economists have found evidence that this mechanism is even at play within not-for-profit institutions such as universities. Using patents —relative control over the invention— it is possible to increase the chances of commercialization by designing license agreements to induce development work by the academic scientists (Jensen and Thursby 2001; Lach and Schankerman 2008). Similarly, a solid body of evidence confirms that the Bayh-Dole Act did increase the economic impact of academic science (Sampat, Mowery, and Ziedonis 2003; N. Hausman 2010).

Universities and firms tend to have different control structures. While differences should not be overstated (Sauermann and Stephan 2012), firms tend to emphasize control over the research output whereas universities put greater emphasis on control over the research direction. The contrast between industry focus and academic freedom has important consequences on scientists' wages (Stern 2004) and on the division of labor between academia and industry. This insight has been recently developed independently and contemporaneously in two closely related formal models (Aghion, Dewatripont, and Stein 2008; Lacetera 2009). Both models study the impact of the distinct control structures in academia and industry on the division of labor between these two types of organizations. However, while Aghion et al. emphasize the multiple-stage nature of the R&D process and focus on the social welfare implications of the wage-freedom tradeoff, Lacetera takes the point of view of the manager and focuses on the tradeoff between keeping control over a specific project or relinquishing control by collaborating with universities, in order to gain more motivated scientists. In both cases, a division of labor emerges between firms and academia in which firms focus on more applied research (Aghion, Dewatripont, and Stein 2008) or on projects of longer duration and more narrow applicability (Lacetera 2009). Interestingly, for our purpose, the Aghion et al. model suggests that the control structure of firms is more adapted to technology spawning—i.e., typically more applied research.

3.3. EMPIRICAL APPROACH

3.3.1. The Challenge

The empirical challenge in testing the two hypotheses of technology spawning is considerable. When observing the emergence of science-based invention, how can we gauge whether these stem from the intrinsic potential of the knowledge itself or from the characteristics of the environment of discovery? Universities are widely believed to conduct much more basic research than firms. As a consequence, the relevance of university research for invention tends to be more indirect. A large number of studies have described the division of labor between firms and universities (Nelson 1986; Rosenberg and Nelson 1994; Rosenberg 1994; Klevorick et al. 1995; Mansfield 1995; Cohen, Nelson, and Walsh 2002). The fundamental empirical challenge is therefore an identification problem. The risk is to conflate the marginal impact of the environment of discovery with the selection effect of knowledge into this environment. A simple comparison between different types of environments (e.g., university vs. industry) might therefore lead to biased results due to unobserved differences in technological potential. Ideally, the researcher would like to observe the potential of the new knowledge and to compare it to realized technology spawning.

3.3.2. Paper Twins

This paper proposes a novel empirical approach exploiting the existence of simultaneous discoveries operationalized as paper twins. Paper twins are the dual instantiation of the same piece of new scientific knowledge in two distinct environments. The following example resulted from a discovery simultaneously made at UCSF and at SmithKline Beecham:

> **Caterina et al. (April 2000) "Impaired Nociception and Pain Sensation in Mice Lacking the Capsaicin Receptor."** ***Science***
> "The capsaicin (vanilloid) receptor VR1 is a cation channel expressed by primary sensory neurons of the "pain" pathway. Heterologously expressed VR1 can be activated by vanilloid compounds, protons, or heat (>43°C), but whether this channel contributes to chemical or thermal sensitivity in vivo is not known. Here, we demonstrate that sensory neurons from mice lacking VR1 are severely deficient in their responses to each of these noxious stimuli. VR1−/− mice showed normal responses to noxious mechanical stimuli but exhibited no vanilloid-evoked pain behavior, were impaired in the detection of painful heat, and showed little thermal hypersensitivity in the setting of inflammation. Thus, VR1 is essential for selective modalities of pain sensation and for tissue injury–induced thermal hyperalgesia."

> **Davis et al. (May 2000) "Vanilloid receptor-1 is essential for inflammatory thermal hyperalgesia."** ***Nature***
> "The vanilloid receptor-1 (VR1) is a ligand-gated, non-selective cation channel expressed predominantly by sensory neurons. VR1 responds to noxious stimuli including capsaicin, the pungent component of chilli peppers, heat and extracellular acidification, and it is able to integrate simultaneous exposure to these stimuli (…). Here we have disrupted the mouse VR1 gene using standard gene targeting techniques. (…) Although the VR1-null mice appeared normal in a wide range of behavioural tests, including responses to acute noxious thermal stimuli, their ability to develop carrageenan-induced thermal hyperalgesia was completely absent. We conclude that VR1 is required for inflammatory sensitization to noxious thermal stimuli but also that alternative mechanisms are sufficient for normal sensation of noxious heat."

These excerpts describe two sets of results obtained by examining the behavior of mice lacking a specific receptor (VR1). Both teams have found that mice in which the VR1 gene had been disrupted exhibit normal reactions to a variety of stimuli but become completely insensitive to one specific stimulus (carrageenan-induced thermal hyperalgesia). One of the team (Caterina et al.) conducted its research within academia and the other team (Davis et al.) in a firm. Both papers were submitted within a month (respectively, January 18th 2000 and December 20th 1999). In short, the (nearly) simultaneous discovery of the capsaicin receptor in two different environments led to the disclosure of the same new knowledge in two distinct papers.

We use simultaneous discoveries as an "experiment" from which it is possible to compare the relative impact of the academic and corporate environments on follow-on invention. Specifically, our empirical strategy exploits three key aspects of the phenomenon associated with the production of paper-twins:

a. since they disclose the same discovery, the knowledge disclosed in each of the paper-twins has intrinsically the same potential for follow-on inventions;
b. since simultaneous discoveries emerge in different environments, the knowledge from each discovery might not actually be turned into new inventions at the same rate;
c. citation and non-citation of each of the twin papers in the patent literature are a noisy but useful measure of the occurrence (or non-occurrence) of follow-on inventions.

3.4. DATA AND METHODS

3.4.1. Sample definition

The data for this study is based on the first automatically and systematically collected dataset of simultaneous discoveries. The full dataset consists in 1,246 papers disclosing 578 discoveries and operationalized as 720 paper twins[9] published between 1970 and 2009. The core of the analysis presented in this paper is, however, based on a subset consisting of 90 scientific

[9] In the data a triplet appears as 3 paper twins, a quadruplet as 6 paper twins

publications disclosing 39 simultaneous discoveries having involved at least one industry-based team and one team based in a public research organization. We disclose the entire dataset of 49 academia-industry paper twins as supplemental material. The method used to build the dataset of paper twins, its theoretical foundations, the algorithm, the dataset, and the robustness analysis, are detailed in a separate paper (Bikard 2012).

The algorithm used to build this dataset is based on the insight that two papers disclosing the same simultaneous discovery are systematically cited together in the follow-on scientific literature, not only in the same papers, but also in the same parenthesis, or adjacently (Cozzens 1989). Figure 1 summarizes the algorithm.

[Insert Figure 1 about here]

Our empirical work relies on the fact that paper-twins are indeed simultaneous discoveries and have therefore inherently the same potential for cumulative invention. Observed variance in the citation rate of two twin papers in the patent literature ought therefore to be due to the different environments in which the research took place rather than on differences in the discovery itself. We test this comparability assumption in several ways. First, we examine the number of months separating the publications of two twin papers. As noted above, the algorithm matches on co-citation rather than publication month. If two alleged paper twins were not really the same, one would expect them to be on average six months apart or more. The 720 paper twins were in fact published on average 1.8 months apart, a lag considerably shorter than the average time between paper submission and publication. In fact, 373 pairs of twins were published the exact same month and 267 of them were published in the same issue of the same journal. Second, we verify the semantic similarity of two twin papers by using the Pubmed related citation algorithm. If the twins were not very closely related, they should not be using the same words and should therefore be ranked far from each other. Pubmed ranks two papers of the same twin right next to each other 42% of the time. The rank difference is inferior to 10 for 90% of the twins[10]. Finally, we collected the opinion of the discoverers themselves. We selected randomly 10 discoverers and asked them to describe the discovery process. Nine of them told us

[10] Rank difference calculated after dropping articles that are published more than a calendar year apart.

about the other twin paper without us asking[11]. After we asked the tenth person why he did not mention the twin paper, he asserted angrily that he deserved all the credit and that his idea had been stolen. Of course, the fact that two teams have published twin papers does not mean that they conducted the exact same experiment or that they obtained the exact same results or that they interpreted them in the exact same way. It also does not mean that both discoveries were independent of one another. However, consistent adjacent co-citation indicates that both papers are very closely related. The fact that the community of experts ruled that credit ought to be shared does indicate (1) that both teams are widely believed to have had the capability to make the discovery and (2) that each team has provided convincing evidence supporting their claim to priority.

Our data is drawn from several sources. Data about each publication comes from ISI Web of Science and Scopus. Details about the corresponding author come from an analysis of the text of the publications. Patent citation data (through May 2011) and information about each citing patent were collected using Google Patents. Table 1 provides a list of variables and definitions.

[Insert Table 1 about here]

3.4.2. Measurement

We tracked our main outcome measures, follow-on inventions, by examining the citations of each of the 90 papers in the patent literature. We used a web crawler that searched for the title of each paper in the patent's body. References in patents are important since they define the scope of the claimed novelty. As such, they are the responsibility of the inventor, the attorney and the examiner. In the US, the applicant has a strong incentive to disclose all prior art that he or she is aware of because failure to do so can lead to patent invalidation, a rule known as the doctrine of "Inequitable Conduct." Citations in the patent literature are an imperfect measure of knowledge diffusion because not all innovations are patented (Cohen, Nelson, and Walsh 2000), not all knowledge flows are cited or citable (Griliches 1990), citations are at time used strategically (Lampe 2010) and a number of them are added by the examiner (Alcácer and Gittelman 2006). Yet they are a readily available, comprehensive and well understood measure of knowledge dissemination and are therefore widely used (Jaffe, Trajtenberg, and Henderson

[11] Some of the discoverers noted that the twin papers emphasize different aspects of the same discovery.

1993; Narin, Hamilton, and Olivastro 1997; Henderson, Jaffe, and Trajtenberg 1998; Gittelman and Kogut 2003; Sorenson and Fleming 2004). In addition, our particular setting presents three characteristics that ought to attenuate some of the concerns associated with this measure. First, we are studying life sciences, an area in which patents are widely used and strategic citation is limited (Lampe 2010). Second, we are studying citations to scientific papers, which tend to be less added by the examiner, less strategically used, and overall a better measure of knowledge diffusion than patent citations to other patents (Cohen, Nelson, and Walsh 2002; Roach and Cohen 2012). Finally, we study published knowledge and prior work has shown that citations in patents are a better indicator of knowledge flow when the latter is more codified (Roach and Cohen 2012). Empirically, our goal is not to estimate whether universities or firms get the "paired patent" (Murray 2002) on the newly discovered knowledge itself. Receiving the paired patent depends mostly on the exact timing of the discovery as well as on the patent application strategy. In contrast, the focus of this paper is on the long-term use of the new knowledge as a springboard for invention.

Table 2 reports summary statistics. Our sample of 39 simultaneous discoveries disclosed in 90 scientific publications has received 533 citations in the patent literature. The distribution of patent citations per twin is highly skewed, which suggests that these 39 discoveries have very different technological potential. One discovery in the data is cited in 41 patents. On the other hand, 17 discoveries are cited in no patents at all. Our main dependent variable CITATION takes a value of 1 if the citation has taken place between the patent and the paper and 0 otherwise. The dataset includes 867 potential citations of which 61% are realized. We also distinguish between different types of assignees. Specifically, we consider separately patents assigned to one of the discoverers and those assigned to third parties; as well as patents assigned to firms and those assigned to universities. While two twin papers are consistently cited together in the scientific literature, the same is not true in the patent literature. In our dataset, the intersection of the forward citations of paper twins in the patent literature is only 21% of the union.

[Insert Table 2 about here]

Our main explanatory variable, ACADEMIA, is a dummy variable equal to one for all papers whose corresponding author was based in a university or public research organization. This measure was chosen because our interviews revealed that as project leader, corresponding authors typically determine whether and how the team will keep building on the new knowledge. For robustness, we have also run our analysis considering that academic (industry) papers are those in which the majority of the authors are from academia (industry). The results remained unchanged. In our dataset, the 39 simultaneous discoveries took place in 41 unique public research institutions and 25 unique firms. The most common public research institutions in the data are Harvard University (4 papers), UT Houston (2 papers), and Stanford University (2 papers) and the most common firms are Genentech (6 papers), GSK (5 papers), and Amgen (4 papers).

Our analysis includes two types of control variables. First, we control for characteristics of the discovery team. US AUTHOR is an indicator variable which equals one if the corresponding author is based in the US and # AUTHORS is the count of the number of authors that are listed on the discovery paper. Second, we consider the characteristic of the patent-to-paper dyad. CITATION LAG is the number of years separating the publication of the paper and the awarding of the patent. GEOG DISTANCE is the geographic distance separating the address of the paper's corresponding author and the address of the first inventor listed on the patent. Finally, the indicator variable SELF PATENT equals one if the patent was assigned to an organization present in the publication's address field.

3.4.3. Methods

At their core, the two perspectives on technology spawning lead to contrasting predictions concerning whether follow-on patents would cite predominantly academic or industry papers. If ease of access is the main driver of cumulative invention, the rate of citations to academic papers should be superior to the rate of citations to their industry twin. Since unobserved characteristics of the inventor or invention (e.g., familiarity with the scientific literature) might be correlated with the origin of the scientific discovery, we use citing patent fixed effects in order to avoid an omitted variable bias. The binary nature of the outcome variable could be modeled using a logistic regression with citing patent fixed effects. However, considering the small number of observations per citing patent, such a model would not be consistent. The well-known incidental parameter problem can be solved by using a conditional likelihood function instead of the usual maximum likelihood. We therefore carry out the estimation using a conditional logit model (Chamberlain 1982). There is, however, a countervailing cost in this approach; it drops all observations in which the patent cites all of the discovery twins. To ensure the robustness of the reported results, all of the regressions were also run using OLS. The results of the analysis are essentially unchanged. Our baseline empirical test for the impact of the academic environment on the extent to which invention *j* has drawn knowledge from paper *i* of twin *k* is:

$$CITATION_{ijk} = f(\varepsilon_{ijk}; \alpha_0 + \alpha_1 ACADEMIA_i + \alpha_2 X_{ij} + \gamma_{jk})$$

where γ_{jk} is a fixed effect for patent *j* citing discovery (paper twin) *k*, $\alpha_2 X_{ij}$ is a vector of control variables and $ACADEMIA_i$ is our main explanatory variable. Robust standard errors are clustered at the level of the citing patent.

In addition to this baseline test, our empirical setting offers the opportunity to test the access and control views of technology spawning in a more nuanced way. First, the importance of control should be particularly salient for follow-on inventions awarded to the discoverers themselves. Admittedly, inventors do not face any access cost to the new knowledge that they discovered. As a measure of discoverer invention, we can count the number of citations in the patent literature that (1) originate from a discoverer and (2) that cite one of the discovery papers. We can use paper-twins fixed-effects to examine the extent to which follow-on invention varies across discovery teams while keeping the discovery constant. Empirically, measuring follow-on invention using patent citations implies that we must account for its form as count data skewed to the right, calling for the use of a count model such as a fixed-effect Poisson with quasi-maximum likelihood (i.e., "robust") estimates (J. Hausman, Hall, and Griliches 1984). We cluster our robust standard errors at the level of the twin. Our test for the impact of the academic environment on invention by the discovery team of paper *i* of twin *k* is therefore:

$$\#\,SELF\ PATENTS_i = f(\varepsilon_{i,k}; \alpha_0 + \alpha_1 ACADEMIA_i + \alpha_2 X_i + \gamma_k)$$

where γ_k is a paper-twin fixed effect, $\alpha_2 X_i$ is a vector of control variables and $ACADEMIA_i$ is our main explanatory variable. Second, the important of ease of access should be particularly salient for follow-on inventors that did not take part in the discovery. These inventors will have a strong incentive to draw their knowledge from the source where access is the least costly—i.e. academia rather than industry. We can test this proposition using our baseline empirical test but restricting our dataset to patents that were awarded to teams that did not take part in the simultaneous discovery.

Finally, the validity of our empirical test rests on the argument that patent citations to papers are not entirely driven by citation norms or strategies. Above, we discussed how the patent citation literature informs our expectations in this regard. This concern should be attenuated by the fact that we are considering patent citations to papers (not to patents) and that our sample is primarily composed of life sciences discoveries. We also attempt to address this concern empirically in three ways. First, we examine the interaction between our main effect and variables such as geographic distance—which ought to be associated with knowledge dissemination but not with citation norm or strategy. Second, we distinguish between citations by academic and corporate patents. Third, we conducted 17 interviews with scientists in order to inquire about the process of technology spawning as well as their citation decision.

3.5. RESULTS

3.5.1. Academic vs. industry science

The significance of our empirical approach depends on the contention that scientific knowledge produced in firms tends to be more directly relevant to the development of new inventions than knowledge produced in universities. We explore the validity of this claim by comparing our subset of 39 simultaneous discoveries in which at least one team is based in industry and another in academia ("matched sample") to our entire dataset of 578 simultaneous discoveries including 1,246 papers of which 51 were authored by a firm and 1,195 were authored by an academic institution ("unmatched sample"). Descriptive statistics for these two samples are presented in Table 3 and Figure 2. The two graphs on top of Figure 2 show the different rates of yearly patent citations for the "unmatched sample." Clearly, the average academic paper receives far fewer patent citations than the average industry publication. The two bottom graphs show the same results for the "matched sample." As apparent from the graph, the difference in citation rate is much smaller. Table 3 similarly compares the unmatched and the matched samples and presents the same result numerically. Interestingly, while firm papers have overall larger teams than academic papers (12.7 versus 7.2 authors in the unmatched sample) this difference seems entirely explained by the different type of science that is conducted in both types of organizations. In the matched sample, industry teams are actually smaller than those from academia (13.2 versus 14.5 authors). Thus, absent a close control for technological potential, comparisons of the research output of academic and industry scientists are not informative of the relative impact of the discovery environment.

[Insert Figure 2 &Table 3 about here]

3.5.2. Academic vs. industry environment

Table 4 presents the main result of our analysis. It considers the entire population of patents that build on one of our simultaneous discoveries and predicts realized citation as a function of whether the paper is from academia or from industry. Paper-twins can be used to observe the non-citation since all patents citing at least one discovery paper could potentially have cited its twins too. The negative impact of the academic environment on patent citation appears substantive (20-30%), statistically significant, and robust to the inclusion of a number of control variables, including characteristics of the paper such as number of authors, US-based, and characteristics of the patent-paper dyad such as time lag, geographic distance or whether the discoverer and the inventor are the same person. This result is consistent with the idea of "Ivory Tower," that the corporate research environment is more propitious to technology spawning than academic research labs. Our data therefore suggest that in our setting, access to the knowledge produced (i.e. the open academic environment) seems less important than controlling the knowledge (i.e. the corporate lab) as a driver of science-based invention.

[Insert Table 4 about here]

3.5.3. Driver of science-based invention: control over the output

In order to get further insights into the importance of an organizational context favoring control rather than openness as a driver of technology spawning, we focus on invention by the discoverers. Admittedly, these scientists can access the knowledge that they created at no cost. Figure 3 presents descriptive statistics and shows that industry papers are associated with a higher number of discoverer patents (z=1.64 and p=0.10 in two-sample Wilcoxon Mann-Whitney test). Table 5 presents the results of the regression analysis. The first two columns use as dependent variables the count of discoverer patents and the two right-hand-side columns present the same regressions but use as dependent variable the count of follow-on patents assigned to the organization of discovery (rather than the individual discoverer). The number of observations is very small since we observe within-discovery variation in the propensity to self-patent for only 13 discoveries involving 32 teams. The results are again in line with the idea that control is a crucial determinant of technology spawning. We find that firm discoverers produce on average over three times more patents than academic discoverers based on the same discovery. Interestingly, the coefficient is even stronger for the organization of discovery than it is for the individual discoverers. This result indicates that follow-on invention in firms does not necessarily involve the discoverer. The same is apparently less true in universities.

[Insert Figure 3 and Table 5 about here]

3.5.4. Driver of science-based invention: ease of access

The importance of ease of access should be particularly striking in the case of third party inventors. Indeed, organizations that did not make the discovery are more likely to draw knowledge where access is the least costly. The descriptive statistics presented in Figure 4 shows that, surprisingly, the share of realized citations is higher for industry papers than for academic ones (z=1.87 and p=0.19 in McNemar chi-square test for matched pairs). Table 6 presents the results of the baseline conditional logistic, excluding every patent awarded to one of the discoverers. Confirming the descriptive results, third party inventors seem to draw their knowledge more from firms than from universities. This result was not expected since prior work has argued that knowledge would flow more easily outside of universities than outside of firms. The negative impact of the academic environment on follow-on invention by non-discoverers appears modest (10-20%) but statistically significant and robust to the inclusion of a number of control variables.

[Insert Figure 4 and Table 6 about here]

3.5.5. Robustness and further nuances

As discussed above, one could be worried that our results on third party citations might be driven not by the flow knowledge but instead by some norm or some strategic decision to cite corporate rather than academic discoveries when both are available. While we cannot entirely disprove this possibility empirically, our results seem more consistent with the knowledge flow explanation.

First, unlike what a pure norm or strategic citation argument would predict, we find that our main effect is not stable over time, location, or size of the discovery team. Table 7 presents an analysis of the variation in our main effect as a function of the number of authors, the time after discovery, whether the inventor and the discoverer were based in the same country, and whether they were located geographically close to each other. In line with the explanation that the difference in cumulative invention is driven by the denser connection of firm scientists in the inventor community, we find that the negative impact of the academic environment increases with the size of the discovery team. This same negative effect also seems to be particularly salient in the years immediately following the discovery and to become weaker overtime. Similarly, the negative effect appears weak in the instance in which the discoveries and the inventions were made in different countries and is stronger when both happened in the same country. Finally, the last column of the table shows that the negative effect of the academic environment decreases the further one is from the place of discovery. Predicted values from this regression are plotted in Figure 5. This statistically significant interaction effect is particularly telling since citation norms and strategies admittedly are not dependent on whether the inventor is geographically close to the discovery team.

[Insert Figure 5 and Table 7 about here]

Second, unlike what a strategic citation argument might predict, our effect holds for both academic and corporate inventors. Strategic citation is admittedly less likely among academic scientists since they are likely to be less concerned about getting sued than industry scientists. In addition, university inventors might be more familiar with the work of other university scientists than firms' researchers. Table 8 splits the sample between third party inventors from academia and those from firms. Interestingly, we find that the negative impact of the academic environment is just as strong—although of lower statistical significance—for inventors from universities and firms. We do not find that knowledge circulates better within the "ivory tower".

[Insert Table 8 about here]

Finally, our interviewees argued against the existence of such a citation norm or strategy. We conducted 17 interviews with discoverers and inventors in order to inquire about the process of technology spawning as well as about their citation decision. All the inventors affirmed that they cited in their patents all the papers that they were aware of and that they regarded as relevant. Inventors typically justified the non-citation by mentioning "lack of awareness" of the twin paper and "lack of time." Other, less common explanations focused small perceived differences between twin publications such as paper clarity, difference in procedure used or different interpretation of the results.

How can we explain the fact that follow-on inventors are more likely to draw on knowledge from firms even though accessing that knowledge is likely to be more costly? Two themes emerged from our interviews. First, inventors watch firms a lot more than universities. While access to a given piece of new knowledge might be less costly if it emerges in academia, inventors systematically invest in monitoring industry science and might therefore ignore or disregard potentially promising knowledge that emerges in universities. In the words of an inventor in a large West Coast biotechnology company: "We monitor our competitors all the time because our bread and butter, our paycheck, depends on how well or how poorly they do." Second, the interviews with inventors also uncovered considerable skepticism toward academic science in the inventor community. One inventor in a prominent pharmaceutical firm declared: "It's a much higher bar [for industry], higher standards, because every error, or every piece of fraud along the way, the end game is going to fail. (...) Therefore, I have more faith in what industry puts out there as a publication." Strikingly, the argument made is based on institutional logic, and runs parallel to the perspective highlighting knowledge accessibility. However, the conclusion is diametrically opposed. Because academics are interested in publications more than in technology development, their results are less likely to be reproducible—i.e. trustworthy. Clearly, these qualitative insights are no demonstration. However, they do provide provoking hypotheses about the mechanism that might underlie the "Ivory Tower" effect that emerged from our quantitative analysis.

3.6. DISCUSSION AND CONCLUSIONS

This paper uses a novel empirical strategy to explore the circumstances under which scientific knowledge might remain underutilized as compared to its technological potential. We contrast two perspectives of the translation of scientific discoveries into new technologies, a process we term "technology spawning." One view emphasizes knowledge accessibility as a driver of technology spawning and the other contends that control over the new knowledge is paramount for attracting private investment. The two views lead to conflicting predictions concerning whether firms or universities constitute the more fertile discovery environment for the emergence of new science-based technologies. The empirical approach presented here exploits the existence of simultaneous discoveries and their instantiation as paper-twins. Because simultaneous discoveries can emerge on both sides of the academia-industry boundary, it is possible to examine the same piece of new knowledge in two different institutional settings. This paper uses the first systematically and automatically generated dataset of simultaneous discoveries, including 578 instances. The core of the study focuses on 39 such discoveries that involved at least one team from academia and one team from industry.

Our results contradict the often-held view that universities spawn more new technologies based on the knowledge that they produce because of their openness. In our data, a given piece of new scientific knowledge seems to lead to more inventions if it emerges in a firm than if it emerges in a university. Academic papers are 20-30% less likely to be cited in follow-on patents than their twins from industry. This apparent "Ivory Tower" effect has two components. On the one hand, the rate of follow-on invention by the discoverer is over three times higher in industry than in academia. On the other, inventors that did not make the discovery are 10-20% more likely to cite the firm paper in their patents than its academic twin. In addition, this result is unlikely to be driven by citation strategy or norms. If inventors monitor and trust knowledge produced by firms more than if it were produced by universities, then academic knowledge may remain underutilized even though it can be accessed easily.

Traditional studies of the commercialization of science have primarily focused on the role of the law, especially intellectual property rights (Williams 2012). The role played by organizations has received far less attention (Murray and O'Mahony 2007). We contrast the impact of a type of organization that provides wide access—universities—and another type of organization that provides more control—firms. We find that organizational structures that provide more control lead to significantly more technology spawning. Ease of access to promising new knowledge is not enough. Our results indicate that the uncertainty concerning the technological potential of new knowledge can be such that inventors use the identity of the knowledge producer as a signaling device. In our setting, signals of value outweigh ease of access as a driver of technology spawning.

These findings should be interpreted carefully. First, we are only measuring the treatment effect for the treated. One could question the generalizability of our results. For instance, since firms are overall less likely to publish their discoveries (Moon 2011), our sample of paper-twins might be selecting relatively open firms. Similarly, the type of knowledge at risk of emerging at the same time in a university and in a firm might be specific. Second, since both papers are published in the same types of journals, citation of one paper might not be independent from citation of its twin. It is therefore difficult to interpret whether non-citation means that the invention would not have occurred absent the twin paper or if it means that it would have occurred but perhaps later, or at a higher cost. On the other hand, considering that the large majority of the discoveries in the dataset are from the life sciences—a field in which university-industry collaboration is particularly intense—the dataset might constitute a lower boundary of the propensity of the academic environment to trap technologically relevant knowledge within its walls.

One should also note that the evidence presented here, that research remains "trapped" inside the Ivory Tower, captures only one aspect of the impact of the academic environment on knowledge dissemination. This paper starts when the discovery process stops, and therefore does not explore the antecedents of knowledge creation, including the ability to stand on other scientists' shoulders (Furman and Stern 2011). Without a detailed accounting of the size of other (positive) effects of the academic environment on knowledge dissemination, it is impossible to calculate the optimal innovation policy towards scientific research.

The academic research environment can be portrayed as an Ivory Tower. On the one hand, research conducted there tends to be more basic and less directly relevant to science-based invention. On the other hand, even the relevant research done there is less likely to be turned into inventions than it would have, had it been conducted in the private sector. By focusing on simultaneous discoveries, it is possible for the first time to observe the non-citation of papers in the patent literature. The list of potential drivers and obstacles to technology spawning is long. Considering the growing desire to see publicly funded scientific research contribute to the economy through its translation into new technologies, the use of simultaneous discovery as "knowledge twins" presents tremendous opportunities for future research.

3.7. REFERENCES

Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein. 2008. "Academic Freedom, Private-Sector Focus, and the Process of Innovation." *The RAND Journal of Economics* 39 (3) (October 1): 617–635.

Alcácer, Juan, and Michelle Gittelman. 2006. "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations." *Review of Economics and Statistics* 88 (4): 774–779.

Argyres, Nicholas S., and Julia Porter Liebeskind. 1998. "Privatizing the Intellectual Commons: Universities and the Commercialization of Biotechnology." *Journal of Economic Behavior & Organization* 35 (4) (May 1): 427–454.

Arora, Ashish, and Alfonso Gambardella. 1994. "Evaluating Technological Information and Utilizing It : Scientific Knowledge, Technological Capability, and External Linkages in Biotechnology." *Journal of Economic Behavior & Organization* 24 (1) (June): 91–114.

Arrow, K. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *R.R. Nelson (ed.), The Rate and Direction of Inventive Activity*, 609–625. Princeton, NJ: Princeton University Press.

Bikard, Michaël. 2012. "Simultaneous Discoveries as a Research Tool: Method and Promise." *MIT Sloan Working Paper.*

Chamberlain, G. 1982. *Analysis of Covariance with Qualitative Data*. National Bureau of Economic Research Cambridge, Mass., USA.

Cockburn, Iain M., and Rebecca M. Henderson. 1998. "Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery." *The Journal of Industrial Economics* 46 (2) (June): 157–182.

Cohen, Wesley M., and Daniel A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation." *Administrative Science Quarterly* 35 (1) (March): 128–152.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh. 2000. *Protecting Their Intellectual Assets: Appropriability Conditions and Why US Manufacturing Firms Patent (or Not).* National Bureau of Economic Research.

———. 2002. "Links and Impacts: The Influence of Public Research on Industrial R&D." *Management Science* 48 (1): 1–23.

Colyvas, Jeannette, Michael Crow, Annetine Gelijns, Roberto Mazzoleni, Richard R. Nelson, Nathan Rosenberg, and Bhaven N. Sampat. 2002. "How Do University Inventions Get into Practice?" *Management Science* 48 (1) (January): 61–72.

Cozzens, Susan E. 1989. *Social Control and Multiple Discovery in Science: The Opiate Receptor Case*. State Univ of New York Press.

Dasgupta, Partha, and Paul A. David. 1994. “Toward a New Economics of Science.” *Research Policy* 23 (5) (September): 487–521.

Fleming, Lee, and Olav Sorenson. 2004. “Science as a Map in Technological Search.” *Strategic Management Journal* 25: 909–928.

Furman, Jeffrey, M. K Kyle, I. M Cockburn, and R. Henderson. 2005. “Public & Private Spillovers, Location and the Productivity of Pharmaceutical Research.” *Annals of Economics and Statistics / Annales d'Économie Et De Statistique* (79/80) (July 1): 165–188.

Furman, Jeffrey, and Scott Stern. 2011. “Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research.” *American Economic Review* 101 (5): 1933–1963.

Gambardella, Alfonso. 1992. “Competitive Advantages from In-house Scientific Research: The US Pharmaceutical Industry in the 1980s.” *Research Policy* 21 (5) (October): 391–407.

Gittelman, Michelle, and Bruce Kogut. 2003. “Does Good Science Lead to Valuable Knowledge? Biotechnology Firms and the Evolutionary Logic of Citation Patterns.” *Management Science* 49 (4) (April 1): 366–382.

Griliches, Zvi. 1990. “Patent Statistics as Economic Indicators: A Survey.” *Journal of Economic Literature* 28 (4) (December): 1661–1707.

Hausman, J., B.H. Hall, and Z. Griliches. 1984. “Econometric Models for Count Data with an Application to the Patents-R & D Relationship.” *Econometrica: Journal of the Econometric Society*: 909–938.

Hausman, N. 2010. “Effects of University Innovation on Local Economic Growth and Entrepreneurship.” National Bureau of Economic Research.

Heller, Michael A., and Rebecca S. Eisenberg. 1998. “Can Patents Deter Innovation? The Anticommons in Biomedical Research.” *Science* 280 (5364) (May 1): 698–701.

Henderson, Rebecca M., Adam B. Jaffe, and Manuel Trajtenberg. 1998. “Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965-1988.” *Review of Economics and Statistics* 80 (1) (February): 119–127.

Von Hippel, Eric A. 1994. “‘Sticky Information’ and the Locus of Problem Solving: Implications for Innovation.” *Management Science* 40 (4) (April): 429–439.

Jaffe, Adam B. 1989. “Real Effects of Academic Research.” *The American Economic Review* 79 (5) (December): 957–970.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca M. Henderson. 1993. “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations.” *The Quarterly Journal of Economics* 108 (3) (August): 577–598.

Jensen, Richard, and Marie Thursby. 2001. "Proofs and Prototypes for Sale: The Licensing of University Inventions." *The American Economic Review* 91 (1) (March): 240–259.

Klevorick, Alvin K., Richard C. Levin, Richard R. Nelson, and Sidney G. Winter. 1995. "On the Sources and Significance of Interindustry Differences in Technological Opportunities." *Research Policy* 24 (2) (March): 185–205.

Lacetera, Nicola. 2009. "Different Missions and Commitment Power in R&D Organizations: Theory and Evidence on Industry-University Alliances." *Organization Science* 20 (3) (June): 565 –582.

Lach, Saul, and Mark Schankerman. 2008. "Incentives and Invention in Universities." *The RAND Journal of Economics* 39 (2) (July 1): 403–433.

Lampe, Ryan. 2010. "Strategic Citation." *Review of Economics and Statistics* 94 (1) (November 3): 320–333.

Mansfield, Edwin. 1995. "Academic Research Underlying Industrial Innovations: Sources, Characteristics, and Financing." *The Review of Economics and Statistics* 77 (1) (February): 55–65.

Merton, Robert K. 1961. "Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science." *Proceedings of the American Philosophical Society* 105 (5) (October 13): 470–486.

———. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton University Press.

Moon, Seongwuk. 2011. "How Does the Management of Research Impact the Disclosure of Knowledge? Evidence from Scientific Publications and Patenting Behavior." *Economics of Innovation and New Technology* 20 (1): 1–32.

Murray, Fiona. 2002. "Innovation as Co-evolution of Scientific and Technological Networks: Exploring Tissue Engineering." *Research Policy* 31 (8-9): 1389–1403.

———. 2010. "The Oncomouse That Roared: Hybrid Exchange Strategies as a Source of Distinction at the Boundary of Overlapping Institutions1." *The American Journal of Sociology* 116 (September): 341–388.

Murray, Fiona, and S. O'Mahony. 2007. "Exploring the Foundations of Cumulative Innovation: Implications for Organization Science." *Organization Science* 18 (6): 1006–1021.

Murray, Fiona, and Scott Stern. 2007. "Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge?: An Empirical Test of the Anti-commons Hypothesis." *Journal of Economic Behavior & Organization* 63 (4) (August): 648–687.

Narin, Francis, Kimberly S. Hamilton, and Dominic Olivastro. 1997. “The Increasing Linkage Between U.S. Technology and Public Science.” *Research Policy* 26 (3) (October): 317–330.

Nelson, Richard R. 1982. “The Role of Knowledge in R&D Efficiency.” *The Quarterly Journal of Economics* 97 (3). The Quarterly Journal of Economics: 453–70.

Nelson, Richard R. 1959. “The Simple Economics of Basic Scientific Research.” *Journal of Political Economy* 67 (3) (June 1): 297–306.

———. 1986. “Institutions Supporting Technical Advance in Industry.” *The American Economic Review* 76 (2) (May 1): 186–189.

Owen-Smith, Jason, and Walter W. Powell. 2004. “Knowledge Networks as Channels and Conduits: The Effects of Spillovers in the Boston Biotechnology Community.” *Organization Science* 15 (1): 5–21.

Polanyi, Michael. 1966. “The Logic of Tacit Inference.” *Philosophy* 41 (155) (January): 1–18.

Roach, Michael, and Wesley M. Cohen. 2012. “Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research.” *Forthcoming, Management Science.*

Rosell, Carlos, and Ajay Agrawal. 2009. “Have University Knowledge Flows Narrowed?: Evidence from Patent Data.” *Research Policy* 38 (1) (February): 1–13.

Rosenberg, Nathan. 1990. “Why Do Firms Do Basic Research (with Their Own Money)?” *Research Policy* 19 (2) (April): 165–174.

———. 1994. *Exploring the Black Box: Technology, Economics, and History.* Cambridge Univ Pr.

Rosenberg, Nathan, and Richard R. Nelson. 1994. “American Universities and Technical Advance in Industry.” *Research Policy* 23 (3) (May): 323–348.

Sampat, Bhaven N., David C. Mowery, and Arvids A. Ziedonis. 2003. “Changes in University Patent Quality After the Bayh-Dole Act: a Re-examination.” *International Journal of Industrial Organization* 21 (9) (November): 1371–1390.

Sauermann, H., and P. E Stephan. 2012. *Conflicting Logics? A Multidimensional View of Industrial and Academic Science.* National Bureau of Economic Research.

Singh, Jasjit, and Ajay Agrawal. 2011. “Recruiting for Ideas: How Firms Exploit the Prior Inventions of New Hires.” *Management Science* 57 (1) (January 1): 129–150.

Sorenson, Olav, and Lee Fleming. 2004. “Science and the Diffusion of Knowledge.” *Research Policy* 33 (10) (December): 1615–1634.

Stern, Scott. 2004. “Do Scientists Pay to Be Scientists?” *Management Science* 50 (6) (June 1): 835–853.

Stokes, D. E. 1997. *Pasteur's Quadrant: Basic Science and Technological Innovation.* Brookings Institution Press.

Williams, H. L. 2012. *Intellectual Property Rights and Innovation: Evidence from the Human Genome.* National Bureau of Economic Research.

Youngson, Alexander John. 1979. *The Scientific Revolution in Victorian Medicine.* New York, NY: Holmes & Meier Publishers.

Zucker, Lynne G., Michael R. Darby, and Jeff S. Armstrong. 2002. “Commercializing Knowledge: University Science, Knowledge Capture, and Firm Performance in Biotechnology.” *Management Science* 48 (1) (January 1): 138–153.

Zucker, Lynne G., Michael R. Darby, and Marilynn B. Brewer. 1998. “Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises.” *The American Economic Review* 88 (1) (March): 290–306.

3.8. TABLES & FIGURES

TABLE 1. VARIABLES AND DEFINITIONS

Variable	Definition	Source
Publication characteristics		
PAPER-TWIN$_k$	Dummy variable for each pair of paper twins	Matching algorithm
ACADEMIA$_i$	Dummy variable equal to 1 if the corresponding author of article i is in a university or a government organization; 0 otherwise	Paper itself
US AUTHOR$_i$	Dummy variable equal to 1 if the corresponding author of article i is in the US; 0 otherwise	Paper itself
PUBLICATION YEAR$_i$	Year in which article i is published	WoS
# AUTHORS$_i$	Count of the number of authors of article i	WoS
SELF PATENTS$_i$	# of patents citing article i awarded to discoverer and issued before May 2011	Google Patent
Patent characteristics		
UNIVERSITY ASSIGNEE$_j$	Percentage of assignees that are universities or a government organizations	USPTO
US INVENTOR$_j$	Dummy variable equal to 1 if the corresponding author of article i is in the US; 0 otherwise	USPTO
APPLICATION YEAR$_j$	Year of patent application to USPTO	USPTO
Citation characteristics		
CITATION$_{ijk}$	Dummy variable equal to 1 if article I of paper twin k is cited in patent j; 0 otherwise	Google Patent
CITATION LAG$_{ij}$	APPLICATION YEAR$_j$ - PUBLICATION YEAR$_i$	USPTO; ISI Web of Science (WoS)
GEOG DISTANCE$_{ij}$	Distance, in miles, between the cities of the address of publication i's corresponding author and patent j's first inventor (Law-of-Cosines-based calculation)	Harvard IQSS patent database; itouchmap.com
SAME COUNTRY$_{ij}$	Dummy variable equal to 1 if the corresponding author of article i is located in the same country as the first inventor of patent j; 0 otherwise	USPTO; WoS
SELF PATENT$_{ij}$	Dummy variable equal to 1 if patent j was assigned to an organization present in publication i's address field	USPTO; WoS

TABLE 2. MEANS AND STANDARD DEVIATIONS

Variable	N	Mean	Std. Dev.	Min	Max
Twin characteristics (N=39)					
# PAPERS PER TWIN	39	2.31	0.52	2	4
% TWINS CITED IN PATENTS	39	0.54	0.51	0	1
# CITING PATENTS PER TWIN	39	9.05	13.28	0	41
Publication characteristics (N=90)					
ACADEMIA$_i$	90	0.54	0.50	0	1
US AUTHOR$_i$	90	0.64	0.48	0	1
# AUTHORS$_i$	90	13.92	24.06	2	216
PUBLICATION YEAR$_i$	90	2000.44	3.65	1994	2008
# SELF PATENTS	90	1.01	3.26	0	26
Patent characteristics (N=533)					
% ACADEMIC ASSIGNEE	523	0.24	0.42	0	1
% US INVENTOR	533	0.80	0.40	0	1
APPLICATION YEAR	533	2003.04	3.26	1995	2009
Patent-Paper dyad characteristics (N=867)					
CITATION	867	0.61	0.49	0	1
CITATION (academic papers only)	456	0.52	0.50	0	1
CITATION (industry papers only)	411	0.72	0.45	0	1
TIME LAG (YEAR)*	867	4.01	3.08	-2	15
SAME COUNTRY	867	0.67	0.47	0	1
GEOGRAPHIC DISTANCE (MILES)	867	2062	1915	0	9728

* TIME LAG can be negative if the citation was added after the patent was filed but before it was issued

TABLE 3. MEANS CONDITIONAL ON DISCOVERY ENVIRONMENT

	ALL PAPERS ("Unmatched Sample")		TWIN ACROSS UNIVERSITY-INDUSTRY BOUNDARY (“Matched Sample”)	
	Univ. Paper	Firm Paper	Univ. Paper	Firm Paper
# Publications	1195	51	49	41
# Citing Patents (total)	1924	508	235	283
# Citing Patents (self cites)	139	98	21	70
Patent citation characteristics				
# PATENT CITES RECEIVED	1.61	9.96	4.80	6.90
% CITATION IN PATENTS	58.4	68.4	52.4	71.5
AV. CITATION LAG	4.54	4.62	4.31	3.81
Publication characteristics				
US AUTHOR	0.61	0.64	0.61	0.65
# AUTHORS	7.25	12.73	14.47	13.27
PUBLICATION YEAR	2001.0	1999.7	2000.6	2000.2

TABLE 4. IMPACT OF THE DISCOVERY ENVIRONMENT ON PATENT CITATION

	CONDITIONAL LOGIT (level: citing patent) Dependent Variable = CITATION (dummy)		
	Baseline marginal impact; no control	Marginal impact; w/ discovery controls	Marginal impact; w/ discovery and dyad controls
ACADEMIA	0.518***	0.588***	0.541***
	(0.07)	(0.08)	(0.09)
Discovery team characteristics			
US AUTHOR		1.787**	0.712
		(0.48)	(0.27)
# AUTHORS		0.996	0.998
		(0.00)	(0.00)
Inventors-Discoverers Dyad characteristic			
PATENT BY SELF			1.551
			(0.47)
CITATION LAG			0.982
			(0.05)
SAME COUNTRY			4.833***
			(2.22)
LOG (GEOG DISTANCE)			1.072
			(0.30)
# of observations	643	643	643
Patent-twin dyads FE	257	257	257

Values are odd ratios; Robust standard error in parenthesis are adjusted for 257 clusters (patent-twin dyads)
*** p<0.01, ** p<0.05, * p<0.1

TABLE 5. IMPACT OF THE DISCOVERY ENVIRONMENT ON DISCOVERER INVENTION

	FIXED EFFECT POISSON QML (level: simultaneous discovery)			
	DV = # PAT BY DISCOVERER		DV = # PAT BY DISCOVERY ORG	
	Marginal impact; no control	Marginal impact w/ controls	Marginal impact; no control	Marginal impact w/ controls
ACADEMIA	0.267***	0.268***	0.212***	0.258***
	(0.08)	(0.07)	(0.06)	(0.08)
Discovery team characteristics				
US AUTHOR		4.453		4.956e+07***
		(7.89)		(4.58e+07)
# AUTHORS		1.013		1.182
		(0.01)		(0.12)
# of observations	32	32	32	32
Paper-twin FE	13	13	13	13

Values are incident rate rations; robust standard errors in parentheses are adjusted for 13 clusters (simultaneous discoveries); *** p<0.01, ** p<0.05, * p<0.1

TABLE 6. IMPACT OF THE DISCOVERY ENVIRONMENT ON THIRD PARTY INVENTION

	CONDITIONAL LOGIT (level: citing patent) Dependent Variable = CITATION (dummy); Self-cites excluded		
	Baseline marginal impact; no control	Marginal impact; w/ discovery controls	Marginal impact; w/ discovery and dyad controls
ACADEMIA	0.580***	0.697**	0.576***
	(0.09)	(0.11)	(0.11)
Discovery team characteristics			
US AUTHOR		1.666**	0.665
		(0.43)	(0.27)
# AUTHORS		0.995*	0.998
		(0.00)	(0.00)
Patent-Paper Dyad characteristics			
CITATION LAG			1.136
			(0.39)
SAME COUNTRY			5.328***
			(2.54)
LOG (GEOG DISTANCE)			0.996
			(0.07)
# of observations	483	483	483
Patent-twin dyads FE	206	206	206

Values are odd ratios; Robust standard error in parenthesis are adjusted for 206 clusters (patent-twin dyads)
*** p<0.01, ** p<0.05, * p<0.1

TABLE 7. VARIATION IN ENVIORNMENT EFFECT ON INVENTION BY THIRD PARTIES

	CONDITIONAL LOGIT (level: citing patent) Dependent Variable = CITATION (dummy); Self-cites excluded			
	Team size and academia effect	Citation lag and academia effect	Same country and academia effect	Geographic distance and academia effect
ACADEMIA	0.862	0.467***	0.722	0.103**
	(0.39)	(0.14)	(0.21)	(0.10)
Discovery team characteristics				
US AUTHOR	0.737	0.674	0.722	0.696
	(0.31)	(0.27)	(0.29)	(0.28)
# AUTHORS	1.025	0.998	0.997	0.997
	(0.03)	(0.00)	(0.00)	(0.00)

Patent-Paper Dyad characteristics				
CITATION LAG	1.125	1.072	1.167	1.061
	(0.38)	(0.37)	(0.40)	(0.37)
SAME COUNTRY	5.008***	5.295***	5.989***	4.570***
	(2.41)	(2.47)	(2.87)	(2.21)
LOG (GEOG DISTANCE)	0.995	0.997	0.994	0.823
	(0.07)	(0.07)	(0.07)	(0.12)
Interactions				
ACADEMIA*# AUTHORS	0.976			
	(0.02)			
ACADEMIA*CITATION LAG		1.049		
		(0.05)		
ACADEMIA*SAME COUNTRY			0.665	
			(0.27)	
ACADEMIA*LOG (GEOG DISTANCE)				1.270*
				(0.17)
# of observations	483	483	483	483

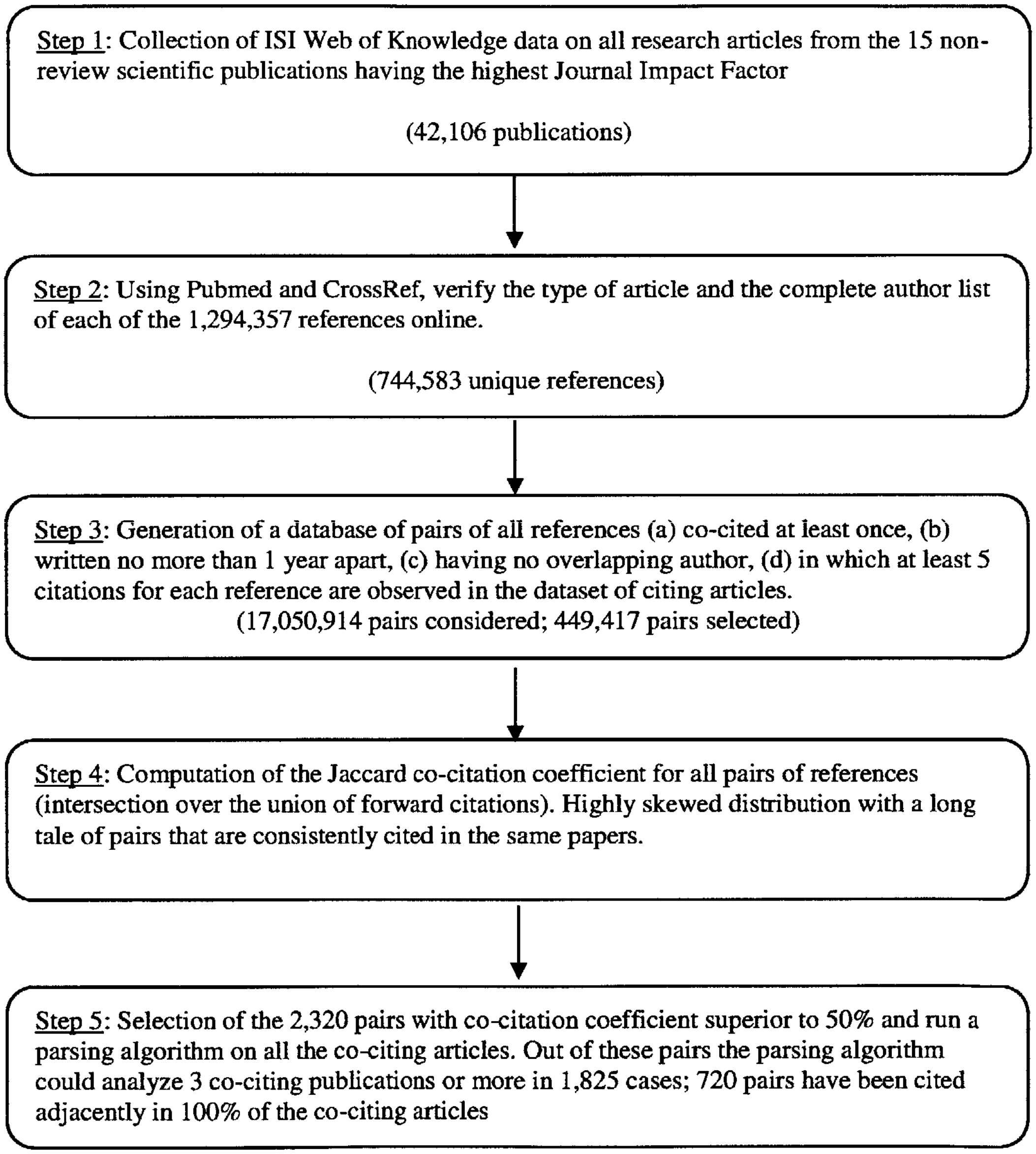
Values are odd ratios; Robust standard error in parenthesis are adjusted for 206 clusters (patent-twin dyads)
*** p<0.01, ** p<0.05, * p<0.1

TABLE 8. ENVIRONMENT EFFECT ON ACADEMIC VS. INDUSTRY THIRD PARTY INVENTION

	CONDITIONAL LOGIT (level: citing patent)			
	Dependent Variable = CITATION (dummy); Self-cites excluded			
	Academia Effect: academic patents only		Academia Effect: corporate patents only	
ACADEMIA	0.556*	0.587	0.664**	0.526***
	(0.18)	(0.21)	(0.13)	(0.13)
Discovery team characteristics				
US AUTHOR	2.152	0.707	1.860*	0.993
	(1.03)	(0.62)	(0.62)	(0.53)
# AUTHORS	1.002	0.998	0.992**	0.998
	(0.01)	(0.01)	(0.00)	(0.00)
Patent-Paper Dyad characteristics				
LOG (DISTANCE)		0.113**		1.704
		(0.12)		(0.72)
SAME COUNTRY		2.571		3.935**
		(2.75)		(2.52)
CITATION LAG		0.877		0.948
		(0.20)		(0.08)
# of observations	117	117	332	332
Patent-twin dyads FE	54	54	137	137

Values are odd ratios; Robust standard error in parenthesis are adjusted for patent-level clusters
*** p<0.01, ** p<0.05, * p<0.1

FIGURE 1. AN AUTOMATED AND SYSTEMATIC METHOD TO GENERATE A LIST OF SIMULTANEOUS DISCOVERIES (Reproduced from Bikard 2012)



Step 1: Collection of ISI Web of Knowledge data on all research articles from the 15 non-review scientific publications having the highest Journal Impact Factor

(42,106 publications)

Step 2: Using Pubmed and CrossRef, verify the type of article and the complete author list of each of the 1,294,357 references online.

(744,583 unique references)

Step 3: Generation of a database of pairs of all references (a) co-cited at least once, (b) written no more than 1 year apart, (c) having no overlapping author, (d) in which at least 5 citations for each reference are observed in the dataset of citing articles.
(17,050,914 pairs considered; 449,417 pairs selected)

Step 4: Computation of the Jaccard co-citation coefficient for all pairs of references (intersection over the union of forward citations). Highly skewed distribution with a long tale of pairs that are consistently cited in the same papers.

Step 5: Selection of the 2,320 pairs with co-citation coefficient superior to 50% and run a parsing algorithm on all the co-citing articles. Out of these pairs the parsing algorithm could analyze 3 co-citing publications or more in 1,825 cases; 720 pairs have been cited adjacently in 100% of the co-citing articles

FIGURE 2. CITATION RATES IN PATENTS: ACADEMIC VS. INDUSTRY PAPERS

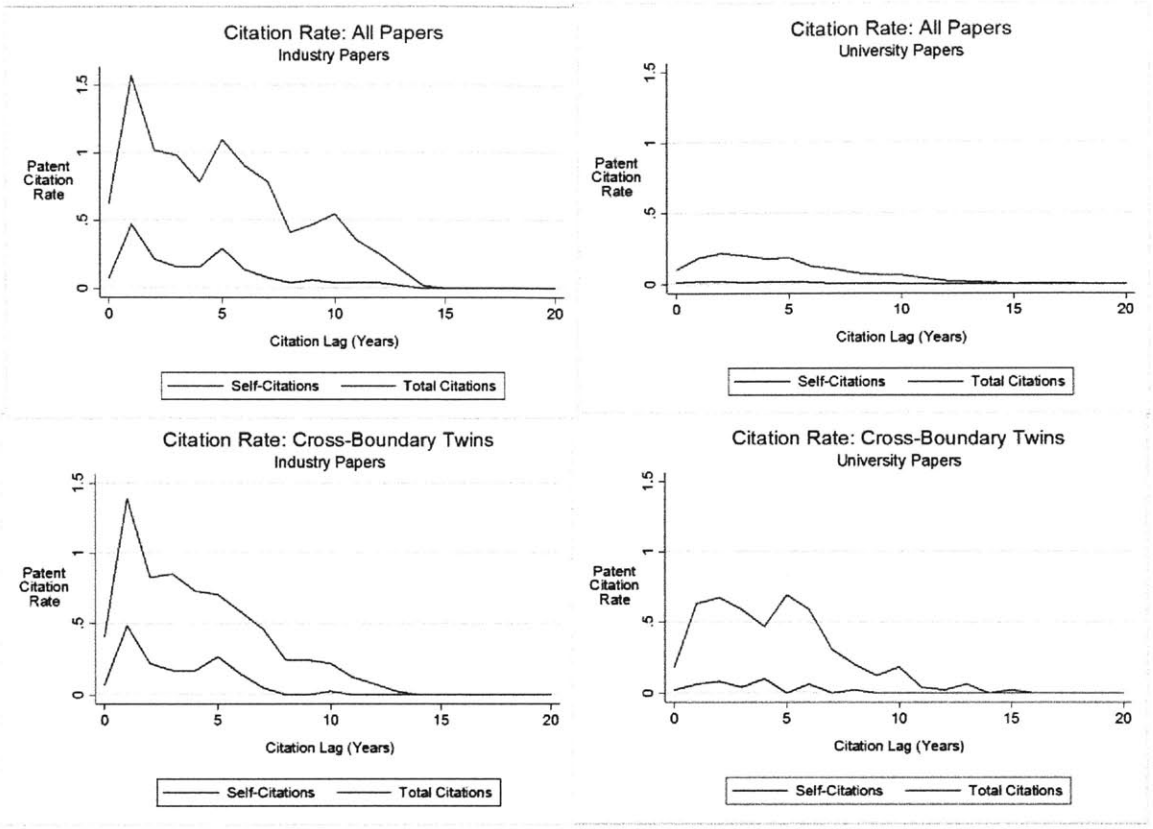


FIGURE 3. DESCRIPTVE STATISTICS: DISCOVERER INVENTION IN FIRMS AND UNIVERSITIES

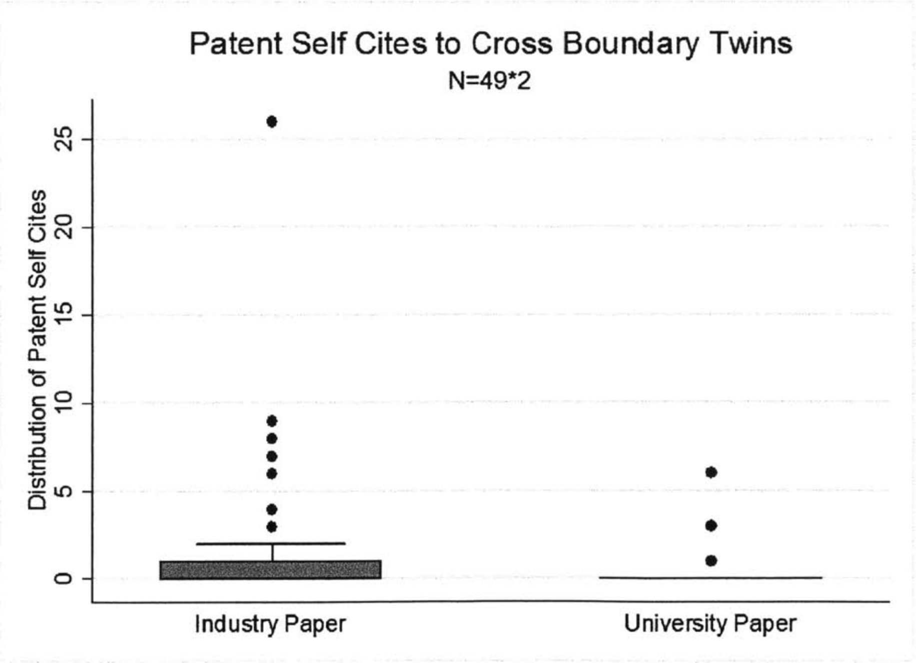


FIGURE 4. DESCRIPTIVE STATISTICS: CITATIONS BY NON-DISCOVERERS (self-cites excluded)

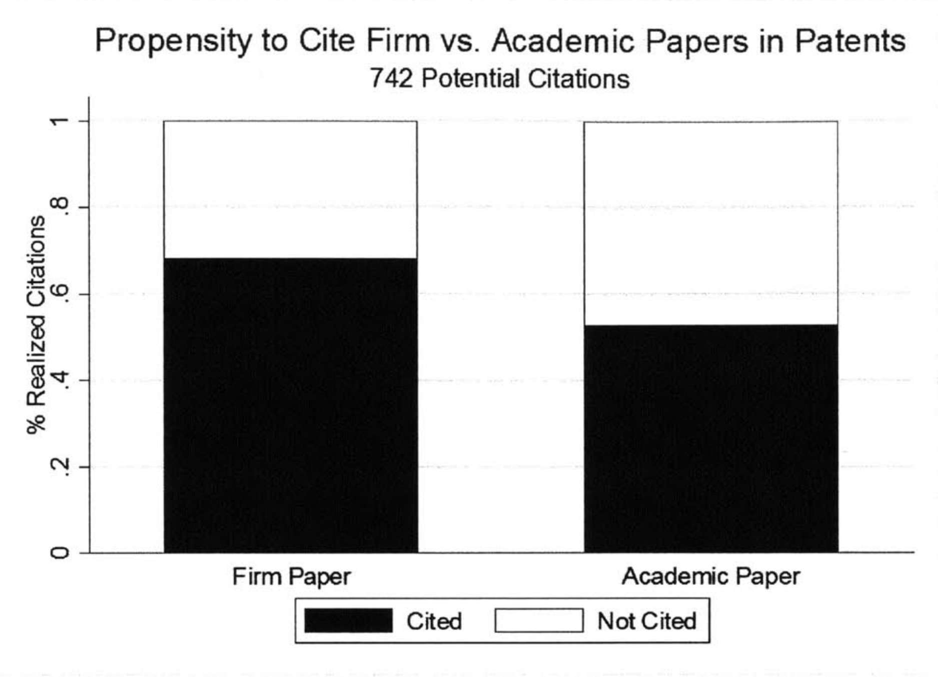
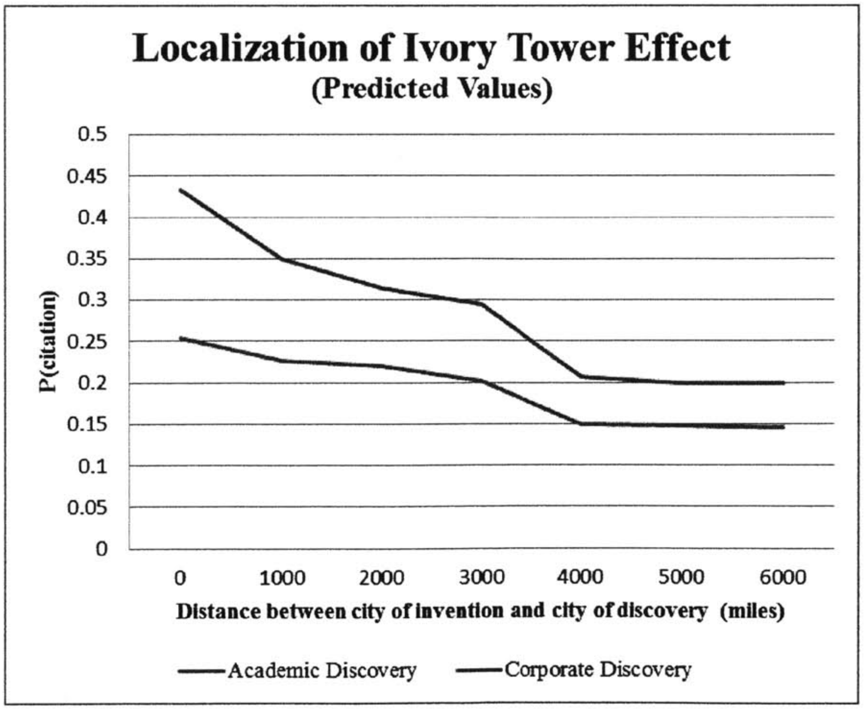


FIGURE 5. IMPACT OF GEOGRAPHIC PROXIMITY ON THE RATE OF PAPER CITATION (Predicted Values)



Chapter Four

Exploring Tradeoffs in the Organization of Scientific Work: Collaboration and Scientific Reward

(with Fiona Murray and Joshua Gans)

4.1. INTRODUCTION

In 2008, the *Journal of Instrumentation* published a paper entitled "The ATLAS Experiment at the CERN Large Hadron Collider" which documented the installation and expected performance of the ATLAS detector that had been installed as a critical component of the Large Hadron Collider to extend the frontiers of particle physics. As the paper states "This detector represents the work of a large collaboration of several thousand physicists, engineers, technicians, and students over a period of fifteen years of dedicated design, development, fabrication, and installation" (p.1). This crisply illustrates the changing nature of scientific work with the need for large numbers of individuals with distinctive expertise to work collaboratively in the solution of a complex scientific problem (Jones 2009). However, while the demands for new more expansive modes of organization push scientists towards larger collaborative groups the reward system for science has not necessarily changed as dramatically: The paper described above has over 1000 authors listed alphabetically, thus, raising the question of whether and how individual authors receive credit for their scholarly contributions. Posed more broadly, how should knowledge workers with high levels of organizational autonomy – such as academic scientists, computer programmers and independent inventors – organize their creative activities? How should they structure their collaborative choice in the light of the potential tradeoff between collaboration on the one hand and credit allocation on the other?

This question is of normative interest as autonomy becomes more prevalent among those engaged in the production of new knowledge, thus, allowing many more individuals to choose the degree to which they work collectively and in collaboration with others in the pursuit of creative outcomes. Not simply a question shaping the daily lives of academic scientists, this is also an issue of managerial import as complex tasks yield only to growing teams leaving open the question of how to allocate credit and other task-based incentives (Holmstrom, 1982; McAfee & McMillan, 1991). The collaboration versus credit question is also of considerable theoretical interest to scholars in light of the increased collective organization of knowledge work inside organizations, in the Academy and in knowledge communities (see for example DiMaggio 2003; Cummings & Kiesler 2007, Adler et al. 2008).

The rise in collective work, in general, and collaborative work, in particular, suggests that collaboration is a highly advantageous organizational choice, particularly for scientists (Wutchy et al. 2007). Empirical evidence repeatedly showing that the creative outputs accomplished by a larger number of people tend to be of higher quality particularly for scientists (Singh & Fleming 2010; Wuchty et al. 2007) but also, for instance, in paintings (Hargadon 2008) and theatre (Uzzi and Spiro 2005). These "facts on the ground" are also greeted with great optimism among scholars who enthusiastically describe the emergence of a "new norm" of collectiveness replacing the age-old tradition of the individual genius (Beaver 2001; Wray 2002; Johansson 2004). Certainly, many studies highlight collaboration's positive aspects: the ability to tap into diverse sources of knowledge (Fleming et al. 2007), the potential to democratize knowledge production (Von Hippel 2005), and its critical role in greater levels of creativity (Hargadon 2003).

Should we, therefore, assume that collaboration is the most effective way to organize knowledge work? Or are there hidden or unmeasured costs associated with the collaborative organization of knowledge production? Scholars in social psychology have provided a more nuanced perspective on the costs of collaboration on creativity (Paulus & Nijstad 2003). Others taking an efficiency perspective (see for example recent analysis by Lee and Bozeman 2005) note that "a trivial but obvious cost [of collaboration], only one person can talk at a time during meetings – assumedly, such communication is instantaneous and almost costless within an individual" (Singh and Fleming (2010, p. 53). A further cost borne by the individual scientist relates to the allocation of credit. Particularly within the scientific community, the central reward system for scientific work is grounded in the provision of credit in reward for novel contributions to the knowledge base (see Dasgupta and David 1994). Traditional modes of credit allocation have been grounded in manuscript authorship and citations to a particular paper – a system that is particularly effective when knowledge work was a largely "solo" activity but is rendered much more complex as knowledge work (and with it authorship) expands to including growing numbers of individuals. Thus, for an individual scientist, the choice of collaboration is made in the shadow of possible tradeoffs in credit allocation as well as other efficiency considerations (Engers et.al., 1999; Gans and Murray, 2013). However, the current analysis of collaboration only in terms of output (i.e. the quality of papers produced) fails to evaluate tradeoffs at the individual level. Thus, the literature ignores whether the benefit to individuals of collectiveness are offset by high potential costs in terms of credit allocation (and other efficiency costs).

In this paper, we take an individual level perspective and evaluate the key tradeoff between the possible benefits of collaboration for the generation of specific outputs –in terms of quantity and quality – and the costs of collaboration to individuals' overall productivity and credit allocation. To do so, we develop a theoretical model that focuses on the decision of an individual scientist in managing their portfolio of research activities building on the model of Becker and Murphy (1992) (that is unrelated to scientific work but highlights production choices). Our model makes three assumptions: that a scientist has a fixed time to allocate to all projects, has discretion in the mode of collaboration and is motivated not only by maximizing quality (citations) but maximizing citations allocated to them. While stylized in nature, these assumptions allow us to derive a set of predictions regarding collaborative behavior and credit allocation tradeoffs. We then test these assumptions by examining the academic publications of 661 faculty-scientists from one institution – the Massachusetts Institute of Technology – over a thirty year period from 1976 to 2006.

Our approach is narrower in scope than the massive data-based efforts analyzing millions of knowledge outputs (Newman 2001; Wuchty et al. 2007) but larger than qualitative small-scale investigations (Melin 2000; Hara et al. 2003). Nonetheless, an individual-level approach (to theoretical modeling and empirics) allows us to consider not only the output of collaboration but the *net* value of collaboration. It presents three crucial advantages over prior studies. First, we can make a realistic examination of the relationship between collaboration and credit at the scientist-year level. Second, we can control for individual's tendency to consistently take part to larger or smaller projects by adding individual-level fixed-effects. Third, we can control for the broader organizational environment by focusing on one institution (adding department-year level fixed-effects).

Our empirical results suggest that collaboration (among MIT researchers) is associated with more highly cited work on a per paper basis, and on an annual basis with more fractional credit – suggesting that credit allocation is not simply divided among the authors of a paper. A given individual in our sample can hope to see their papers receive on average over 60% more citations if they choose to collaborate with a coauthor as opposed to working alone. Up to 4 coauthors, collaboration is also associated with the publication of more papers per author. Using a revealed preference approach, our data also indicates that scientists might be disproportionately rewarded for more collaborative work—i.e. that credit for a given collaborative paper is shared across coauthors in a way that sums up to more than 1. Not all collaborations are equal, however. In line with theories of cross-fertilization of ideas and division of labor, we find that cross-departmental collaborations tend to produce higher quality papers at a lower productivity cost than within-departmental work. Free-riding is also apparent: the quality gain is particularly low and the productivity loss is particularly high when collaborating with a more senior scientist, especially if that scientist is from the same department.

The paper is organized into five sections. In Section 2 we outline the tradeoffs between collaboration as an input into scientific work and credit sharing in the output of collaboration. Section 3 lays out a formal model of this tradeoff from which we derive clear hypotheses. Section 4 describes our setting and method. We detail our results in Section 5. We end with discussion and conclusions.

4.2. COLLABORATION *VERSUS* CREDIT TRADEOFF

Enthusiasm for collaboration is most visible among practitioners: A large number of popular press articles, books, and consulting business reports claim that collaboration provides a superior form of work organization (Hoerr 1989; Dumaine & Gustke 1990; Katzenbach et al. 1993; Orsburn & Moran 2000; Koplowitz et al. 2009). Similarly, in scientific research, the vast majority of policy-makers have embraced the trend toward larger research groups and supports its further development (J. S. Katz & Martin 1997; Landry & Amara 1998; Stokols et al. 2005).

In the US, for instance, the National Institute of Health (NIH) Roadmap for Medical Research lists changing "academic culture to foster collaboration" as one of its four main objectives.[12] Accordingly, it has made available a number of grants to support collective science. For example, the aptly named "Glue Grants" program[13] from the National Institute of General Medical Sciences allocates tens of millions of Dollars to encourage scientists to collectively "tackle complex problems that are of central importance to biomedical science." Overall, the positive perception of collaboration in scientific research was crystallized in the *Science* editorial written by former National Science Board[14] chairman arguing that: "It is clear that knowledge and distributed intelligence holds immense potential, both from a scientific standpoint and as a driver of progress and opportunity for all Americans" (Zare 1997).

Edward Lawler, in an interview for *Fortune*, takes a more nuanced view in line with our theoretical and empirical approach when he noted that "teams are the Ferraris of work design, they're high performance but high maintenance and expensive" (Dumaine 1994, p.2). This highlights the central tension between the positive benefits of collaboration and the possible negative tradeoffs for creative work. In laying out the tradeoffs, we focus on the benefits of collaboration (versus working alone) from a variety of theoretical perspectives and then contrast this with the costs including efficiency considerations but also more centrally the costs in terms of credit allocation.

4.2.1. Collaboration's Benefits

Researchers, like many practitioners, are traditionally optimistic about the impact of collaboration on creative work. At the core of this perspective lies the notion that the division of labor allows individuals endowed with different knowledge, beliefs, skills, and social networks to come together, thus enabling creativity and novelty. Accordingly, groups establish an ideal context for creativity through the recombination of existing ideas (Gilfillan 1935): the variety of ideas and contexts to which group members have been exposed can be easily united during

[12] http://nihroadmap.nih.gov/
[13] http://www.nigms.nih.gov/Initiatives/Collaborative/GlueGrants/
[14] Governing body of NSF

collaborative work, potentially igniting an explosion of novel ideas – a phenomenon popularized as "the Medici Effect" (Johansson 2004). It has been argued that collaborative groups enhance the circulation of knowledge by bringing together members with different information, social networks, and skills (Cummings 2004; Singh 2005; Ding et al. 2010). They do so in part because individuals serve as brokers fostering inspiration across domains (Hargadon & Sutton 1997; Obstfeld 2005; Fleming et al. 2007; Singh & Fleming 2010; Girotra et al. 2010). More specifically, researchers have documented that social interactions can indeed lead to fleeting moments of collective creative insight (Hargadon & Bechky 2006) and that collective work enables members to identify and filter out bad ideas before they fully develop (Singh & Fleming 2010). In addition, groups can be safe arenas for individuals to express original ideas without fearing ridicule (Edmondson 1999). With regards scientific work, by bringing together individuals endowed with different types of knowledge (Porac et al. 2004; Hara et al. 2003), scholars argue that collaboration allows scientists to take advantage of specialization in the deep stock of scientific knowledge while at the same time gaining the benefits of breadth (Jones 2009).

Empirical evidence supports the view that collaboration leads to significant benefits on a variety of output dimensions: The commercial success of creative work such as comic books, Hollywood productions and Broadway musicals, as well as its reception by critics, has been linked to collaboration (Taylor & Greve 2006; Cattani & Ferriani 2008; Uzzi & Spiro 2005). Survey data and field work in firms also highlight the positive performance of groups performing creative work compared to individuals (Obstfeld 2005; Burt 2004; Hargadon & Bechky 2006). As noted in the introduction, more systematic quantitative evidence linking more creative tasks to larger groups is largely based on analyses of both scientific knowledge - patents and papers. Here, the data show that outputs authored by more scientists tend to receive more citations (Adams et al. 2005; Wuchty et al. 2007; Fleming 2007). For instance, Wuchty and colleagues (2007) studying 20 million scientific publications and over 2 millions patents find a clear and increasing advantage of collaborative work in all broad research areas. Specifically, Science and

Engineering papers written by two authors received 1.30 more citations than sole-authored papers in the 1950s and that this ratio increased to 1.74 by the 1990s.[15]

Beyond assessing the *average* effect, collaboration is thought to impact the variance in creative outcomes. The direction of this relationship, however, is complex and current results are contradictory. On the one hand, Taylor and Greve (2006) find that collaboration in comic books increases the variance in good and bad outcomes. On the other, in an analysis of US utility patents, Fleming (2007) finds the opposite – i.e. individual inventors are the source of more failures and more breakthroughs. More recently, a careful study of the creative outcome distribution of over half million patents (as captured by their citations), using quartile regressions shows that collaboration reduces the probability of poor outcomes while increasing the probability of extremely successful ones (Singh and Fleming 2010).

4.2.2. Tradeoffs – coordination and credit

Research (as personal experience) suggests a number of potential *coordination costs* associated with collaboration. These costs consume time and have a variety of origins including conflicting goals and incentives, communication difficulties, the need for translation for or education of collaborators of different backgrounds and the need for processes and routines to distribute work, synchronize, and monitor progress. The issue of synchronization is perhaps most eloquently described by Leslie Perlow in her study of the organization of time at work among Ditto's software engine. Using data from a nine-month field study, Perlow (1999) documents how interactive activities can foster insights and learning. More importantly, she also shows that these same activities have a high cost of individual productivity when they are not synchronized, phenomenon leading to "time famines" for knowledge workers. Coordination costs are documented in academic research for instance by Porac and colleagues who have found that the most heterogeneous collaboration in their study, Eco, had the most issues of communication and synchronization, but yet saw a large increase in productivity after its members had learnt to work

[15] In their data, self-citations account for only 5-10% of the relative citation advantage of collaboration, therefore even accounting for self-citations "the relative citation advantage of teams remains essentially intact" (Wuchty et al. 2007, p.2)

together (Porac et al. 2004). Similarly, Cummings and Kiesler have found that multi-university scientific collaboration impose considerable coordination costs and leads to under-performance absent of a significant coordination effort (Cummings & Kiesler 2007).

Further drivers of coordination costs have been explored in the social psychology literature which outlines several cognitive processes leading to inefficiencies in collaboration (Diehl & Stroebe 1987). First, "production blocking" results from the chaotic interactions of the group, which impede the emergence of a consistent train of thoughts. Second, "evaluation apprehension" stems from the fear that some members might have of the others' judgment of their ideas. Finally, some authors have emphasized "information bias," which stems from a search for consensus within groups (Paulus 2007). It should be noted that a recent lab study by Girotra, Terwiesch and Ulrich (2010) finds that many of these drawbacks of collaboration can be mitigated through hybrid structures, in which individuals first work separately and then work together. Overall then, prior research suggests that the coordination difficulties stemming from collaboration in creative work are generally associated with a loss of individual productivity.

Credit allocation is the second major potential cost to collaboration. This arises because the credit is central to the reward system in knowledge work, particularly for scientific research conducted in the Academy in accordance with the norms of open science (Dasgupta and David 1994). However, while credit can be linked to a particular publication of "piece" of knowledge work, such credit must also be allocated to its producers – the authors. When researchers work alone and publish alone they serve as the sole recipients of credit for the quality of the output. In contrast, collaboration requires a more complex allocation calculus. The central importance of this issue for collaboration in creative work arises because as Merton noted: "[citations] are in truth central to the incentive system and an underlying sense of distributive justice that do much to energize the advancement of knowledge" (Merton 1988, p.621). Nonetheless, citations counts have been criticized for a number of reasons including the fact that – independently of the article's "intrinsic" merits – the amount of citations it is likely to receive will depend on the year of its publication, its field, the journal where it is published, its style, its author, its availability online, etc (Bornmann & Daniel 2008). While some have tried to disentangle quality from popularity (Salganik et al. 2006), such distinctions are problematic in creative work, where –as

Stein's definition suggests[16] – broad acceptation by the audience is often considered the only standard upon which quality can be assessed (Stein 1953).

In science, as in other types of creative work, impact is paramount. "For science to be advanced, it is not enough that fruitful ideas be originated or new experiments developed or new problems formulated or new methods instituted. The innovations must be effectively communicated to others. That, after all, is what we mean by a *contribution*[17] to science – something given to the common fund of knowledge" (Merton 1968, p.59). It is of course possible that research developed via collaboration will have a greater impact because the larger team has a superior ability to communicate, mobilize support for, and bring attention to novel ideas. Collaborations play both a social and a cognitive role in this respect. In its social role, a group provides greater communication channels for the dissemination of novel ideas, thus enabling more visibility because each group member is endowed with a distinct set of relationships that he or she can use to promote the novel idea (e.g. Allen 1978; Tushman & R. Katz 1980; Valderas 2007). Collaboration can also be instrumental in bringing legitimacy to a novel idea. Merton, for instance, noted that famous researchers lend visibility and credibility to a paper and that therefore students sometimes "feel that to have a better known name on the paper will be of help to them." (Merton 1968, p.57) a proposition recently validated in the case of the protocols submitted to the Internet Engineering Task Force (Simcoe & Waguespack 2011). Similarly, in Hollywood, legitimacy can be gained through collaboration with individuals that are central to the network of producers (Cattani & Ferriani 2008).

Nonetheless, while garnering greater attention overall, each individual contributor to the research must consider how this additional impact will shape their own credit allocation– a consideration that has not been heretofore examined. More specifically, researchers must consider the tradeoff between the greater impact overall and the credit allocation they receive and how it is spread among numerous authors. As the Hadron Collider paper illustrates, if individual authors only receive fractional credit allocation consistent with a linear function of the number of authors, collaboration becomes a much less appealing prospect (absent other modes of

[16] Morris Stein famously defined creative work as "a novel work that is accepted as tenable or useful or satisfying by a group in some point in time." (Stein 1953, p.311)

[17] Italicized in the original text

credit for research activities). As an illustration, consider the decision of a talented scientist – should she spend a year engaged in two collaborative projects each with one partner i.e. engaging 50% of her time in each of the two projects or should she work alone? That decision is tied importantly to the amount of credit received for collaborative projects compared with other projects. When a scientist devotes time to a collaborative project, not only must they take into account the balance of quality versus coordination costs but also the possibility that they receive only fractional attribution for the resulting output. Thus, the collaborative projects that are actually observed will likely reflect the highest quality amongst those projects (an outcome that likely biases current results around the returns to collaboration).

Thus, credit allocation, as well as the norms associated with credit allocation must be incorporated into current empirical and theoretical perspectives regarding collaborative choices, particularly from an individual perspective. This is challenging because we have relatively little systematic data regarding credit allocation practices. The issue of authorship and credit has received widespread discussion in the scientific press, particularly with regards to "ghost" authors who make only limited contributions to a paper. In a recent release, publisher Elsevier noted that "Naming authors on a scientific paper ensures that the appropriate individuals get credit, and are accountable, for the research."[18] Nonetheless, ours is the first paper we are aware of that incorporates credit allocation as well as coordination into a model of collaboration. It is also the first paper that attempts to use empirical data to derive a possible credit allocation function from observable collaborative choices of scientists over many years of research activity.

4.3. FORMAL MODEL AND HYPOTHESES

Empirical evaluation of the costs of collaboration is centrally an issue of measurement: while many approaches can be taken in observing the quality of research output and the level of collaboration in the form of citations and formal co-authorship respectively, these measures are potentially independent of coordination costs and the credit allocation costs because they are captured at the level of the publication. If, instead, we consider a scientist's collaborative choices

[18] http://ethics.elsevier.com/pdf/ETHICS_AUTH01a.pdf

at the individual year level, their portfolio of choices is more revealing of the tradeoffs in coordination and credit, thus providing a clearer window into collaborative choices well beyond observable quality differences.

To shed light on these tradeoffs and formulate hypotheses we have developed a formal model that explicitly considers the drivers of observable variables by formalizing a variety of the different underlying models that scientists might use to determine their own tradeoffs year on year. In this section, we provide that model and use it to motivate our empirical approach and the inferences that might be drawn from it. The goal of the model is to clearly exposit the benefits to collaboration and the possible credit allocation costs in a situation where scientific rewards from collaboration are clearly and consistently defined.

To this end, we focus on the decision of an individual scientist in managing their portfolio of activities. This requires several assumptions that, while stylized, are consistent with the evidence of scientists' broader choices and preferences. First, we assume that the scientist has discretion over the set of projects worked on and on the structure of collaboration for any given project. In reality of course, collaboration is a mutual decision (and an overture to collaboration could be rejected). For simplicity, we assume here that the focal scientist has full discretion over this choice. We will, however, comment on the implications of that simplifying assumption below.

Second, we assume that, not only is the scientist motivated by maximizing the total number of citations they receive for their portfolio of work, but also on the credit attribution of those citations: specifically they are motivated by the citations that are attributed to them rather than those attributed to other collaborators. For instance, if the scientist completes and publishes a single author paper, they receive attribution consisting of the total amount of citations to that paper. However, when the scientist publishes a co-authored paper, their attribution may not be the full amount of citations to that paper. Instead, their 'share' depends upon a variety of factors. While there is a paucity of empirical evidence on attribution, a number of factors are likely to intermediate including the identity of the collaborator (e.g., relative rank, field) and number of collaborators. It should also be noted that, while we use the expression 'share of attribution' as this is a useful way of conceptualizing attribution, as will be seen below, we do not impose a requirement that the 'shares' of all scientists involved in a project sum to one.

Third, we assume that our scientist has a fixed amount of time to allocate across all projects and all of the activities that constitute those projects. In reality, a scientist could choose the amount of time they devote to research as opposed to other activities and this choice may be impacted by collaboration decisions. However, it is most simple to assume that scientists have a fixed allotment of time available for research and to assume that they are maximizing the effective allocation of that time. As a starting point, we build on the model of Becker and Murphy (1992). Their model concerned the division of labor in product activities and was neither about scientific research nor about collaboration in science. However, some elements of their model are well matched to the environment under consideration here. Where our model differs is in the concept of reward attribution and in the notion that exists a portfolio of projects; Becker and Murphy (1992) consider only one project.

4.3.1. Model Set-Up and Assumptions

Let us begin with the 'production function' for citations from a particular paper. Following Becker and Murphy, we assume that there is a continuum of tasks on the unit interval $s \in [0,1]$ that must be performed in order to produce a paper from a research project. To this end, suppose the number of citations for a paper, *i*, is Q where:

$$Q = \min_{0 \le s \le 1} Q(s) \qquad (1)$$

The Leontief production function captures the notion that each task, *s*, must be performed for output to be non-zero. The key assumption here is not the assumption of strict complements between tasks but their complementarity. Each task can itself be performed at a certain degree of quality, $Q(s)$, where we assume that $Q(s) = E(N)T^{\theta}(s)$ where E_i is the productivity associated with total hours, $T(s)$, devoted to task *s*, $\theta \in (0,1)$ and *N* is the total number of collaborators on the project. We assume that $E(1) = 1$ and *E* is increasing in *N*. This is a simple way of capturing the notion that specialization increases productivity.[19] However, collaboration also requires time,

[19] Becker and Murphy (1992) assume that productivity increases also require an allocation of time but ultimately this reduces to specialization increasing productivity. For notational simplicity, we remove that extraneous layer of endogeneity here.

$t(N)$, to be devoted to coordinating the activities of that team.[20] As outlined above, past studies of collaboration have focused on understanding the net effect of changes in E and t with N. As we demonstrate below, measures of these are complicated by time constraints and the fact that scientists pursue a portfolio of projects with varying levels of collaboration over time.

4.3.2. Equilibrium Collaboration for a Single Paper

To begin, we focus on the allocation of time for a given paper. Suppose that a scientist, n, is assigned a set, S_n, of the tasks of a paper. Then total time devoted by n to the paper is $T_n = t(N) + \int_{s \in S_n} T(s)ds$. We assume here that the opportunity cost of that time is $C_n(T_n)$; a function that will be modeled explicitly below. Given this, n solves the following problem:

$$\max_{\{T_n(s)\}_{s \in S_n}} \alpha(N)Q - C_n(T_n) \qquad (2)$$

where $\alpha(N) \leq 1$ is the fraction of total citations from i attributable to n. We assume that if $N = 1$, then $\alpha = 1$. This fraction is considered to be independent of Q realized.[21]

To derive the chosen allocation of time, as they are symmetric, we assume that time devoted to each task is equal. Thus, $\int_{s \in S_n} T(s)ds = S_n T(s)$ and $Q(s) = E(N)T^{\theta}(s)$ so the optimal $Q(s)$ satisfies the minimum of this or the minimal quality achieved for a task by collaborating scientists.

To complete the model, assume that if there are N collaborators to a project, they split the number of tasks between them equally. This is a natural assumption if scientists are symmetric[22]

[20] Becker and Murphy (1992) did not model coordination costs specifically and assumed that those costs were a function $C(N)$. Here we provide an additional layer of endogeneity consistent with our notion that scientists are allocating time across projects and thus, time spent in coordinating activities as an opportunity cost determined by time not allocated to other projects.

[21] One can imagine that the attribution may come from market assessments as to the relative contribution of collaborators in a scientific team and such attribution may itself depend on the performance of tasks the scientist is known for. Thus, the fraction of total citations attributable to n may be dependent upon the realized quality of a project. We assumed away this complex problem here. Gans and Murray (2013) investigate it in more detail with a formal model.

and results in $Q = E(N)T^{\theta}(s)$. Holding the time allocation choices of other collaborators as given, the scientist chooses T_n to maximize:

$$\max_{T_n} \alpha(N)E(N)\left(N(T_n - t(N))\right)^{\theta} - C(T_n) \qquad (3)$$

Note that if $\min_{m\neq n} T_m < T_n$, it is optimal to lower T_n to that minimum. Thus, there is potentially a continuum of equilibria in this game. The equilibrium with the highest allocation of time, T_n^*, is characterized by the first order condition:

$$\alpha(N)E(N)N\theta\left(N(T_n - t(N))\right)^{\theta-1} = C'(T_n^*) \qquad (4)$$

This equation plays a key role in what follows.[23] Specifically, we focus on the equilibrium with the highest allocation of time.

4.3.3. Equilibrium Collaboration with Multiple Papers

Our purpose here is to measure the impact of collaboration on productivity and, in the process, make inferences about the benefits and costs of collaboration and also the structure of the scientific reward function for research teams. The above analysis shows that collaboration can be beneficial because of the exploitation of specialization and the division of labor but incurs a potential cost in coordination. However, collaboration also impacts time available for a scientist to pursue other papers; in particular, sole-authored projects without collaboration. Here we introduce that option into the model.

What follows is an examination of the impact of introducing collaborators on one paper in the portfolio of papers that a scientist is involved in producing during a given time period. To this end, we assume that the scientist can allocate time to an additional paper. That paper is single authored. Consequently, there are no advantages from the division of labor but the

[22] This is a strong simplifying assumption as it assumes that no regard to differences in the opportunity cost in time are taken into account when allocating tasks between collaborating scientists. However, the qualitative predictions of the model that we focus on for this paper would not be changed if this assumption were relaxed.

[23] It could also be used to analyze other issues such as the optimal team size. These are issues explored in Jiang, Thursby and Thursby (2012).

scientist faces no coordination costs and receives attribution equal to the full value (in terms of quality or citations) of the paper. We otherwise assume that the paper's production function is equivalent to that specified above.

For the single-authored paper in the portfolio, if we assume that the total time allocation a scientist has is 1, then

$$C'(T_n) = \theta\left(1 - T_n\right)^{\theta-1} \qquad (5)$$

Using this we can solve for the optimal time allocation to the collaborative project given N:

$$\begin{aligned} &\alpha(N)E(N)N\theta\left(N(T_n - t(N))\right)^{\theta-1} = \theta\left(1 - T_n\right)^{\theta-1} \\ &\Rightarrow T_n^* = t(N) + \frac{1 - t(N)}{1 + N(\alpha(N)E(N)N)^{1/(\theta-1)}} \qquad (6) \\ &\Rightarrow 1 - T_n^* = \left(1 - t(N)\right)\frac{N(\alpha(N)E(N)N)^{1/(\theta-1)}}{1 + N(\alpha(N)E(N)N)^{1/(\theta-1)}} \end{aligned}$$

Given this, scientists face a choice. They can collaborate on one of the papers with N participants (leaving a second paper single authored) or they can pursue two single authored papers. The choice depends not only on the quality improvement (if any) arising from collaboration but also from the time cost (if any) diverted from single authored papers as well as the level of attribution the scientist expects from the collaborative project.

Our model exposes the central issue with empirical analyses of the impact of collaboration on scientific productivity and quality: the challenges with studies that focus purely on collaboration versus non-collaboration without accounting for time considerations or individual scientist effects. Because individual scientists are constrained in the time they have at any particular moment, collaborative projects impact time allocation and hence, the observed quality of collaborative and single authored projects. In particular, from (6) note that T_n^* is decreasing in $N(\alpha(N)E(N)N)^{1/(\theta-1)}$ and

$$\begin{aligned} \frac{\partial\left(N(\alpha(N)E(N)N)^{1/(\theta-1)}\right)}{\partial N} &= (\alpha(N)E(N)N)^{1/(\theta-1)} \\ &\quad + \tfrac{N}{\theta-1}(\alpha(N)E(N)N)^{(2-\theta)/(\theta-1)}\left(\tfrac{\partial\alpha}{\partial N}E(N)N + \tfrac{\partial E}{\partial N}\alpha(N)N + \alpha(N)E(N)\right) \end{aligned}$$

This expression is positive if $\alpha(N)E(N)N$ does not vary much with N. In this situation, an increase in collaboration may allow the scientist to *reduce* the time allocated to the collaborative paper in favor of the non-collaborative paper; consequently, studies focused on collaboration may understate the productivity of collaboration. Alternatively, in cases when the impact of collaboration on productivity is high (i.e., $E(N)$ varies substantially with N) time will be *drawn towards* the collaborative project away from the single authored paper overstating the pure productivity of collaboration. From an empirical standpoint, it is only by controlling for scientist-year fixed effects that these distortions can be mitigated. A similar issue arises with respect to coordination costs from collaboration: these result in a reduction in 'research time' for both the collaborative and single authored project. Again, to properly identify the portfolio effects of collaboration, year effects are required to exploit variations in portfolio mix over time.

There are three hypotheses that can be tested with this model. The first concerns the average quality of publications:

H1: A scientist has higher quality average publications in years in which they collaborate more.

This is a direct implication of the notion that scientists are decision-makers with regard to the collaboration choice. As collaborative publications involve a fractional allocation of credit, i.e., $\alpha(N) < 1$, a scientist will only collaborate if this results in a higher quality over their portfolio of projects.

Second, collaborative projects involve costs in terms of a reduction in the quantity of papers accredited to scientists:

H2: In years when the scientist collaborates more, fractional publications fall.

In our model, when the scientist single-authors all papers, they have an output of 2 papers while if they collaborates on one of those papers, their fractional output is $1 + \alpha(N)$ or $1 + 1/N$ in the case of simple fractional allocation. Note that the converse could be true: collaboration may 'free up' a scientist's time with the result that this hypothesis will be refuted as more single authored projects or alternative collaborations are pursued. This will indicate that $\alpha(N)E(N)N$ does not vary much with N.

Third, suppose that collaborative opportunities are equally or harder to come by than individual research projects, then the 'rate of return' to collaboration in a particular year should be (weakly) positive:

H3: For a given $\alpha(N)$, the fractional quality of the portfolio attributed to the scientist in years they collaborate more should be no less than the quality of the portfolio they achieve in years they collaborate less.

The intuition here is that a scientist takes their expected attribution from collaboration as given and chooses their portfolio to maximize their attributed quality. Thus, for a posited $\alpha(N)$ if we see a negative return, this is evidence that the posited $\alpha(N)$ is not consistent with observed collaborative behavior.

4.4. EMPIRICAL APPROACH

4.4.1. Data and Setting

Given the challenges associated with empirical analysis, we have chosen to focus on the collaborative choices and publication outcomes of a sample of scientists over a long period of time, thus allowing us to include both individual and year fixed effects in our analysis. This contrasts with approaches that compare outputs at the publication level.

Our setting is a comprehensive dataset of research publication activity at a single university – the Massachusetts Institute of Technology – including the research output over the thirty-one year period 1976-2006 of more than 650 faculty members in 7 departments from the Schools of Science and the School of Engineering. This focus on a single university over time is particularly appealing for several reasons. First, a scientist's choice of whether or not to collaborate is little constrained by formal organizational structure. Second, these choices can be easily traced out from one year to the next by following authorship on publications. Third, "quality" can be analyzed using the (albeit imperfect) metric of citations. Fourth, as noted above our setting offers the opportunity to control for individual effects allowing us to tease out the impact of collaboration (Woodman et al. 1993). Lastly by selecting one institution, we can control for institutional setting. Our choice of MIT is not only one of convenience: it has been shown that prestigious institutions participate more fully in collaborative science (Adams et al. 2005; Jones et al. 2008) and, thus, using MIT allows us to examine the "leading edge" of collaboration.

The core of the study is a sample of publishing faculty members drawn from the annual lists as faculty members at MIT (in the Academic Bulletin). Criteria for inclusion are the following: First, we include faculty from the following seven departments - Electrical Engineering and Computer Science, Chemical Engineering, Material Science and Engineering, Mechanical Engineering, Biology, Chemistry and Physics. These were selected because they include both science and engineering disciplines and are among the most well established parts of the MIT research activity. Second, faculty must be listed for at least a consecutive period of 3 years in order to avoid counting visiting professors, whose participation in research groups of particular size might be systematically biased by their short stay. Third, we chose the period 1976 to 2006 because of ISI data limitations and because 1975 was the year in which the still stable departmental arrangements were established.[24] Fourth, we exclude all the scientists who ever took part in projects with more than 20 authors due to the decoupling of authorship and contribution for specific projects and particular fields (Knorr-Cetina 1999): using ISI subfields, this included scientists in 5 "Big Science" subfields – Astronomy & Astrophysics, Multidisciplinary Physics, Nuclear Physics, Instruments & Instrumentation, Particles & Fields Physics. (Note, however, that our results are robust to the inclusion on these scientists in our data).

We identified 846 individual scientists from our set of Departmental and year criteria. We then excluded 128 (most of whom were already Emeriti Professors in 1976) due to a lack of any publication record for the time period. We further excluded 57 scientists who had taken part to projects that included more than 20 authors. We use our list of 661 publishing faculty as the basis of our analysis. For these people, we collected individual level information including PhD year and topic from UMI Proquest Dissertation Database, as well as departmental affiliation and seniority from MIT course catalogue for the 31 years (Assistant, Associate, Full Professor or

[24] In 1975 the department of Electrical Engineering expanded to become Electrical Engineering and Computer Science and the department of Metallurgy and Materials Sciences merged into Materials Science and Engineering.

Emeritus). We collected all the articles written by our scientists during their time at MIT using ISI Web of Science. Between 1976 and 2006, the 661 scientists stayed at MIT for 5,964 faculty-years and wrote 21,054 publications.

4.4.2. Dependent Variables

Quality: We measure quality (*Q* in our formal model) by observing the average number of citations received by a scientist for all the papers he or she published in a given year. As noted above, while citation giving is a part of normal science, citation counts are an imperfect measure of quality, impact or credit. However, across a sample of publications, citations are a relatively objective and convenient measure of an article's quality and impact and have therefore been widely used in science evaluation (Furman & Stern 2011; Leahey 2007; Wuchty et al. 2007). Practically, we used the number of citations received by 2008 ($Cites$) as a measure of impact at the paper-level. Using this metric, we can calculate the yearly quality of the scientist's publications by observing the average number of citations that they receive. For scientist i in year t, the quality of publication k was measured as the average number of citations $cites_k$ in that year's publications: $Cites_{it} = E(Cites_k)_{it}$.

In order to examine whether variation in marketing capability is driving our results about quality, we check the robustness of all our results using another proxy for work quality: the average Journal Impact Factor (2009 data) of every scientist-year. We could not find a JIF for 16.2% of the publications – the discrepancy coming from low ranking journals, conference proceedings, journals which have disappeared and those who have changed name. Each model using JIF was run twice, a first time considering that the missing JIF was 0, a second time by imputing the missing data using the article number of citations. These methods consistently led to the same results.

Productivity (Quantity): We measure $NPubs_{it}$, the productivity of a scientist's work (*E* in our formal model) by keeping the input constant (a scientist-year) and observing the number of papers published in a given year. It is worth noting that of course publication data is only a proxy for the number of projects that a scientist is involved with; publications are only the disclosed final outcomes of projects and may therefore undercount the total number of projects if some lead to no publications. Nonetheless, our ability to aggregate publications to the individual faculty-year level serves as an important step towards accounting for inputs into collaboration (Girotra et al. 2010,·p.593). Beyond simply capturing *NPubs*, our approach builds on Lee and Bozeman (2005) and also examines the "factional count" of paper published. As the most simple functional form of fractional counting of productivity, we compute *Frac_Pubs* the sum of "papers shares" directly attributed to the scientist. In other words, if the scientist *i* in the year *t* has published *n* papers, each of which includes a number $NAuthors_k$ of authors, their fractional publication count for that year is:

$$Frac\_Pubs_{it} = \sum_{k=1}^{n} \alpha(NAuthors_k) \ where \ n = NPubs_{it}$$

Credit Allocation: As detailed in our formal model, we consider scientists' motivation to collaborate to be dependent on $\alpha(N)$, the share of their research output that is attributed to them (rather than to their collaborators). The specific functional form of this attribution is an empirical question that we examine in this paper. For a given $\alpha(N)$, we compute the number of yearly citations attributed to a scientist's work by summing the citations attributed to the author for every paper of the year. In other words, if the scientist *i* in the year *t* has published *n* papers (*k*), each of which includes $NAuthors_k$ and has received $Cites_k$ by 2008, their attributed citation count for that year is:

$$Att\_Cites_{it} = \sum_{k=1}^{n} Cites_k * \alpha(NAuthors_k) \ where \ n = NPubs_{it}$$

At one extreme, if $\alpha(N) = 1$, then the credit for a collaborative paper is not split and each author claims the entire credit for each coauthored paper and its citations. At the other extreme, if $\alpha(N) = 1/N$, then the scientists split the credit across every author and the sum of shares of all scientists involved in a project sums to one. A third possible form is that $\alpha(N) = 1/\sqrt{N}$ – i.e., that scientists can claim less credit for a coauthored than for a sole-authored paper, but that the sum of the shares of credit attributed to all the scientists in a coauthored paper is superior to one.[25]

4.4.3. Independent Variable – Collaboration

We measured the extent of collaboration during a scientist-year by considering the mean number of coauthors (*N* in the formal model) for all the publications of that year (see Wuchty et al. 2007; Adams et al. 2005; Jones et al. 2008). We obtain the number of coauthors on a project ($NAuthors_k$) by counting the numbers of names in the author field of each of our publications (also referred to as the coauthorship index (Bordons et al. 1996)). While co-authorship remains a practice as a form of currency in the cycles of scientific credit (Latour & Woolgar 1986), it only reflects actual collaboration to the extent that authorship reflects participation. A few studies have noted that this measure is an imperfect one (Subramanyam 1983; J. S. Katz & Martin 1997; Cronin et al. 2004): Distinguished researchers are sometimes added to the authorship list despite the fact that their contribution is relatively minor, a practice known as "guest authorship." Conversely, "ghost authors" are individuals who are not recognized as coauthors despite their significant contribution. Recent work has shown that norms of inclusion vary by discipline and that inclusion is often positively correlated to a scientist's social standing (Biagioli 2003; Häussler & Sauermann 2011). Decoupling contribution and authorship, increases measurement error in our analysis: to avoid some these issues we exclude from our sample scientists who have taken part to any publication with more than twenty coauthors. In our regressions, we also control for disciplinary, temporal, individual and career-related patterns. For scientist *i* in year *t*, collaboration was measured as the average number of authors $NAuthors_k$ in that year publications: $NAuthors_{it} = E(NAuthors_k)_{it}$.

[25] Other synthetic measures of performance have been suggested such as the index *h* which measures for each individual scientist the number of papers with citation superior or equal to *h* (Hirsch 2005). According to this measure, a scientist would have an *h*-index of 20 if he or she has published 20 papers having received more than 20 citations and all the other published papers have received fewer than 20 citations. While this measure is attractive because of its simplicity and ability to synthesize quantity and quality at the level of the individual, this measure is not adapted to measuring within-individual variation in performance over time.

4.4.4. Control Variables

Individual Ability: Individual aptitudes are widely believed to be a predictor of creativity and quality (Amabile et al. 1996; Woodman et al. 1993). Prior research has also shown that better scientists tend to work in larger groups (H. Zuckerman 1972). Controlling for individual-level variance in creative ability is, therefore, crucial if we are to disentangle the impact of collaboration on the quality of scientific work. An important advantage of our setting is that we have a number of observations per individual and can therefore introduce in our models a dummy variable for each of our scientist.

Career Stage: Scientific creativity and propensity to collaborate are widely believed to vary over their career span (H. Zuckerman 1972; Stephan & Levin 2001; Jones 2009). It is, therefore, important to control for such career-level variation. We therefore introduce an indicator variable for each of the scientist's career stage: Assistant Professor, Associate Professor, Professor and Professor Emeritus.

Department-Year: General citations patterns vary from one year to the next and are known to be increasing over time due to the fast expansion of knowledge production (Cawkell 1976). Moreover, this expansion might vary from one discipline to the next. To control for such variation, we included an indicator variable for all department-years in the sample.

Authoring Position: In order to check the robustness of our findings and control for authoring position, we introduce a dummy variable when the focal scientist is the first (last) author. At the level of the faculty-year, our *First Author* variable (*Last Author* variable) is the propensity of the scientist to be first (last) author for all of their year's publications.

4.4.5. Empirical Approach

H1: A scientist has higher quality average publications in years in which they collaborate more: We test hypothesis 1 by assessing the impact of an individual's annual collaborative behavior on the average quality of the scientist's publications. The mean number of co-authors for the year proxies for collaboration and $E(Cites_k)_{it}$ is our measure of quality (we also measure the average journal impact factor as an alternative measure of quality). Because we can control for individual and contextual characteristics, we are building on and bringing further robustness to prior results that collaboration is associated with higher quality output (Adams et al. 2005; Wuchty et al. 2007). We use an OLS regression (Adams & Griliches 1998; Adams et al. 2005) with department-year, individual scientist, and career stage fixed effect. In all our regressions, robust standard errors are clustered at the level of the individual scientist to account for the non-independence of observations from the same author. We estimate:

$$ln(Cites_{it}) = f(\varepsilon_{it};\ NAuthors_{it} + \beta_i + \delta_t + \theta_{ic} + X_{it})$$

where β_i is the fixed effect for each scientist, δ_t is the fixed effect for each department-year, θ_{ic} is the fixed-effect for career stage and X_{it} represents a vector of variables (potentially including a squared term as well as measures of authoring position) which may be associated with output quality.

H2: In years when the scientist collaborates more, fractional publications fall: We test H2 by studying the impact of an individual's yearly collaborative behavior on the quantity of the scientist's publications. We, therefore, examine productivity by studying the relationship between collaboration and the number of papers published that year using the *Frac_Pubs* measure to account for the fractional number of publication (and compared to the total number of publication *NPubs*). Because the number of attributed publications per year is a continuous variable skewed to the right, we used the natural log to alleviate this skewness and used OLS with robust standard errors for our estimation. As earlier, we used scientist, career stage and department-year fixed effects. Specifically we estimate the following equations:

$$ln(Frac\_Pubs_{it}) = f(\varepsilon_{it};\ NAuthors_{it} + \beta_i + \delta_t + \theta_{ic} + X_{it})$$

where, as in the previous equation, β_i is the fixed effect for each scientist, δ_t is the fixed effect for each department-year, θ_{ic} is the fixed effect for career stage and X_{it} represents a vector of variables which may be associated with productivity.

H3: For a given α(N), the fractional quality of the portfolio attributed to the scientist in years they collaborate more should be no less than the quality of the portfolio they achieve in years they collaborate less. As is apparent from our formal model, the benefit of collaboration relative to non-collaboration will depend on α(N) the share of the credit attributed to the scientist for a collaborative paper. If scientists behave rationally and maximize their attributed citations, we should find that the hypothesized positive impact of collaboration on quality and its negative impact on quantity would cancel one another. Assuming that collaborative opportunities are scarcer than non-collaborative ones (since collaborators might be hard to find), we might expect that scientists systematically under-collaborate and, therefore, find that the overall returns to collaboration might appear positive. We test hypothesis 3 by examining attributed citations as a function of collaboration for a given α(N). As earlier, we take the natural log of the fractional number of citations and use an OLS estimator with robust standard errors, as well as scientist, career-stage and department-year fixed effects:

$$ln(Att\_Cites_{it}) = f(\varepsilon_{it};\ NAuthors_{it} + \beta_i + \delta_t + \theta_{ic} + X_{it})$$

According to H3, one would not expect that the overall returns to collaboration be negative. Assuming that scientists optimize the citations that are attributed to them every year, for a given $\alpha(N)$, finding negative returns to collaboration would indicate that in our chosen $\alpha(N)$ probably underestimates the actual credit that scientists are getting for the work that they produce. While testing H3, we will consider three different functions for $\alpha(N)$ as described above.

We deepen our understanding of the mechanisms at work shaping the impact of collaboration on quality and productivity by distinguish the effect of *different* types of co-authorship. Specifically, we compare the impact on the focal scientist of collaborations with a more junior scientist, a more senior scientist and a scientist of the same rank. We can also explore the impact of collaborating across departments and/or with non-PIs. To do so, we limit our analysis the subset of the sample of scientist-years in which every published paper involved only MIT-affiliated authors (2,273 faculty-years and 4,617 publications) allowing us to identify all the MIT PIs, their department and their career stage as well as count the number of non-PIs on each paper.

4.5. RESULTS

4.5.1. Descriptive Statistics

Table 1 and 2 present the main variables of our analysis. Over the 5,964 faculty-year observations we have data on a total of 21,054 publications. This allows us to track the extent to which the researcher collaborated by observing the average number of coauthors for the year. Mean group size ($NAuthors$) is 3.8 authors. Collaboration at the faculty-year level varies between 1 and 20. Scientists did not collaborate at all only during 157 faculty-years (2.6%). In 64% of the faculty-years average group size was between 2 and 4 authors. The entire distribution of group sizes in the data is plotted in Figure 1.

[Figure 1 approximately here]

The key dependent variables in our data are quality (average number of forward citations received by the papers produced in a faculty-year), productivity (quantity of papers attributed to the scientist per year), and the overall credit (citations) attributed to the scientist for a year of work. With regards to quality (*Cites*), the average number of forward citations received (by 2008) by the papers written in a faculty-year is 41.3. Turning to productivity, the MIT researchers published an average of 3.5 papers per year (i.e. *NPubs* is 3.5). However, *Frac_Pubs* (i.e. assuming $\alpha(N) = 1/N$) is only 1.1. Both quality and productivity are highly skewed across faculty-years. We also measure the credit attributed to the scientist of the year of work using the three proposed functional forms:

- If $\alpha(N) = 1$, the mean *Att_Cites* is 165.1 citations
- If $\alpha(N) = 1/\sqrt{N}$, then mean *Att_Cites* is 86.4 citations
- If $\alpha(N) = 1/N$ then mean *Att_Cites* is 48.7 citations.

Table 2 presents the correlation coefficients of our main variables. It highlights that average group size and year are positively correlated (+0.23), i.e. collaboration in our sample has evolved over time toward larger groups. Also note that productivity has also been increasing over time, and that this holds across both *NPubs* and *Frac_Pubs* (0.20 and 0.11 respectively). Consistent with the prior literature, we find a positive correlation between collaboration (*NAuthors*) and quality (*Cites*) (+0.09). Table 2 also shows that the correlation between collaboration on the one hand and yearly productivity on the other is highly dependent whether we consider *NPubs* or *Frac_Pubs*: For $\alpha(N) = 1$, collaboration is positively correlated with productivity (+0.13) i.e. not surprisingly, on average, collaboration is associated with more authored papers but the correlation is negative for *Frac_Pubs* (-0.14). With regards to credit attribution, for $\alpha(N) = 1$ the correlation is positive (+0.11) but for $\alpha(N) = 1/N$ the correlation is almost null (-0.01). Interestingly, the correlation between productivity and quality appears overall positive, i.e. we do not find evidence of any intrinsic quality-quantity trade-off in academic research.

[Table 1 and 2 approximately here]

4.5.2. Econometric Analysis of Hypotheses

Our hypotheses 1, 2 and 3 are tested in Table 3. Model (3-1) confirms Hypothesis 1 and in doing so, adds robustness to the result of prior studies in showing that choosing to collaborate in larger groups leads *on average* to higher quality outputs since we can control for individual and context level idiosyncrasies. The highly significant positive coefficient on group size confirms our hypothesis 1. More specifically the coefficient of 0.099 can be interpreted as an increase by about 10% of the number of citations received per paper for the addition of one collaborator on average for the year.

Our hypothesis 2, that collaboration is associated with a loss in productivity is tested in models (3-2) and (3-3). At the level of the year at work, we find in (3-2) that the choice to collaborate in larger groups is not associated with a higher number of authored publications per year. This result is particularly interesting since we have seen in Table 2 that the two variables are overall positively correlated in the data. On average, however, it seems that at the level of the scientist's yearly choice, the potential productivity gains stemming from specialization and division of labor are counterbalanced by the costs of coordination. This result is even more striking if we consider that, by choosing collaboration, scientists cannot really be "allocated" all the resulting publications but rather might consider their fractional contribution to the stock of published knowledge i.e. $\alpha(N) < 1$. Model (3-3) shows that collaboration is associated on average with lower attributed productivity. In the case of fractional publication counts (3-3), group size is negatively correlated (-0.069).

Finally, models (3-4) to (3-6) explore credit attribution from collaboration. Model (3-4) shows that the quality benefit from collaboration is on average superior to its productivity cost where $\alpha(N) = 1$. This result can be interpreted in two ways. First, if one believes that scientists get all the credit for each of their coauthored publication, then scientists might be systematically under-collaborating. A second, more plausible interpretation of the result is that $\alpha(N) < 1$—i.e., that scientists actually share some of the credit with their collaborators. Models (3-5) and (3-6) propose different credit sharing functions $\alpha(N)$. $\alpha(N) = 1/\sqrt{N}$ is explored in (3-5) and is consistent with scientists in our dataset rationally using collaboration to maximize their yearly attributed impact. Indeed, model (3-5) shows no statistically significant correlation between collaboration and yearly attributed impact overall. Model (3-6) uses a more strict credit sharing function in which $\alpha(N) = 1/N$ and shows a statistically significant negative relationship between collaboration and yearly attributed impact. This result suggests that the credit for a given collaborative paper is not shared across coauthors in a way that sums up to 1. Taken together these results suggest that the credit sharing function of $\alpha(N) = 1/\sqrt{N}$ is most closely associated with rational collaborative behavior all else being equal.

[Table 3 approximately here]

Figure 2 presents the regression estimates when we dichotomize our main independent variable, average collaboration size for a faculty-year. We can note that the upper left graph that the relationship between collaboration and output quality seems to have decreasing returns and to pick at 8 collaborators. The middle row graphs display the relationship between collaboration and productivity for *NPubs* ($\alpha(N) = 1$) and for *Frac_pubs* ($\alpha(N) = 1/N$). Interestingly, for relatively large collaborations (average coauthoring groups of 5 or more for the year), the relationship between collaboration and productivity is negative for *Npubs* and *Frac_pubs*. The difference between publication attribution functions comes from relatively small collaboration levels. If coauthors do not "share" the papers they write but instead account for all their papers equally, then collaboration is positively associated with productivity for collaborations of up to three coauthors on average per year. However, if papers are split across coauthors, then collaboration is associated with negative (fractional) productivity for every value of N. This result shows that scientists produce fewer papers when they collaborate than when they work alone — but individually, each of them will have more lines on their CV as long as $N < 5$ on average for the year.

The bottom graphs also show striking consistencies across $\alpha(N)$ functions. For yearly average collaboration of up to 3 coauthors, collaboration is associated with more attributed citations. The different results that we observed in models (3-4) to (3-6) stem from average yearly group size of 4 or more. If credit does not get split, then these highly collaborative years are associated with more attributed citations. However, if it does get split, these years are associated with similar levels of ($\alpha(N) = 1/\sqrt{N}$) or fewer ($\alpha(N) = 1/N$) attributed citations.

[Figure 2 approximately here]

4.5.3. Robustness Analysis

In Tables 4 and 5, we subject our results to additional robustness tests. We begin by examining the relationship between collaboration and quality. While citations have been broadly used as a measure of publication quality, one could worry that they might be associated with some marketing advantage that larger groups might have. In model (4-1), we use an alternative measure of paper quality: the journal impact factor (JIF) of the publishing outlet. For the few publications in the dataset for which we could not find a JIF, we imputed the latter based on the citations that the articles had received.[26] We find that scientists not only receive more citations on average for the publications on which they collaborate more, they also get published in journals of higher impact factor. Models (4-2) to (4-4) examine whether and how authorship position impacts our findings.[27] It is, for instance, possible that in those years in which our scientists collaborate more they are attributed fewer citations because they do not have the "controlling position" of being a first or last author. One way to control for authoring position and to avoid the use of fractional measures is to assign to a scientist only the publications in

[26] In another model, we also replaced missing JIF with the value 0 since journals for which we could not find information are likely to be of minor importance. The results remained unchanged.

[27] Note that some of this variance is already accounted for by the fact that we include in all of our model career-stage fixed effect (authorship position is closely related to career-stage)

which he or she is the last author – and give him or her the entire credit for the publication and resulting citations. Our results are very consistent with those that we find when using $\alpha(N) = 1/N$.

[Table 4 approximately here]

While fractional measures have been used in bibliometric studies for many years (for an early example see (Price & Beaver 1966)), one could also worry that our results for the case in which $\alpha(N) = 1/N$ might be mechanically driven by the fact that our main independent variable, collaboration, is also in the denominator of our fractional measures of output quantity and overall contribution. This worry, however, is unfounded here because what we are really interested in is precisely whether there are decreasing returns to collaboration. In other words, a negative coefficient in our fractional regressions is evidence of a concave relationship between collaboration and creative output. Although most visible through fractional measures, our result can equally be observed without using any fractional variable as in Models (4-3) and (4-4).

Model (5-1) and (5-2) achieve the same results in yet another manner; they show that collaboration is associated with significant decreasing returns to scale concerning both quantity and yearly contribution (as measured through forward citations). The relationship is concave: *N* scientists working separately during a given time publish more articles and receive more forward citations than *N* scientists working together. The inflection points implied by the coefficients in these two models are respectively 5.4 in (5-1) and 9.6 in (5-2). They are also visible in the left-hand side graphs presented in Figure 2 (where alpha(N)=1). Finally, one might worry that our results primarily hold for publication of average quality. If one believes that only the very best publications really matter, then one might be interested in the relationship between collaboration and taking part to a very high impact publication. Models (5-3) to (5-5) consider only the top 5% of the paper published by department-year and shows that the general patterns observed overall also holds for this subset of publications only.

[Table 5 approximately here]

4.5.4. Different Collaborators

In the subset of 2,273 faculty-years in which MIT PIs only publish with a coauthor that was affiliated with MIT, we explore the impact of different types of research collaborators. Table 6 provides the descriptive statistics for this subsample. For these within-MIT years, PIs on average collaborated with 0.3 other MIT PIs and 1.7 non-PIs per year. About half of the inter-PI collaborations (54%) took place between scientists of the same rank, and collaboration took place both within-department and across-departments at a similar rate (0.11 and 0.15 collaborating PIs per year respectively).

[Table 6 approximately here]

Table 7 illustrates that our main results from Table 3 still hold for our subsample of MIT PIs in years in which they have only chosen to collaborate with other MIT PIs. Specifically, researchers who collaborate in larger groups produce higher quality papers (7-1) and get fewer fractional papers (7-3) although if $\alpha(N) = 1$ i.e., they consider all their publications (*NPub*s) then the number of PIs has no significant impact (see (7-2)).

[Table 7 approximately here]

The relationship between collaboration and quality might be driven by a number of distinct mechanisms including increased time-input for collaborative work, but also potentially cross-fertilization from different perspectives and higher credibility of larger groups. These mechanisms would lead to conflicting predictions regarding whether scientists would profit more from collaborating within their department or with PIs from other departments. If the positive relationship between collaboration and quality is driven by credibility alone, then within-department collaboration would be the most advantageous. If it was driven by cross-fertilization, then cross-department collaboration might particularly lead to higher quality papers. Finally, if it is simply driven by higher input, then both types of collaboration should have a similar positive impact on paper quality. Model (8-1) shows that cross-department collaboration is much more strongly associated with higher quality papers than within-department work. This result suggests that the main mechanism driving the positive relationship between collaboration and quality is cross-fertilization rather than credibility or simply higher input.

[Table 8 approximately here]

We have also seen that the relationship between collaboration and productivity might be driven by two conflicting mechanisms: coordination cost and division of labor/specialization. The fact that we find an overall negative relationship between collaboration and productivity suggests that the coordination costs outweigh on average the gains from specialization. Should we then conclude that there is no gain from a division of labor in scientific research? Models (8-2) and (8-3) show the contrary. We find that the "productivity cost" to collaboration is lower for collaborations that span departmental boundaries, indicating that division of labor does decrease the cost to collaboration. Overall, then, we find that the apparent trade-off between quality and productivity is not the same for every type of collaboration. Specifically, we find collaboration has more benefits and is less costly when it involves individuals from different departments. This overall difference between within and across department collaboration is also visible when studying the citations attributed to individual scientists.

Models (8-4) to (8-6) show that unless scientists get all the credit for collaborative work, within-department collaboration is associated with fewer attributed yearly citations. In contrast, across department collaboration is associated with significantly more attributed citations if $\alpha(N) = 1$, and no significantly fewer attributed citations if $\alpha(N) = 1/\sqrt{N}$.

[Table 9 approximately here]

Finally, the rank of the collaborator might also influence the collaboration's outcome. On the one hand, a prestigious collaborator might increase both the quality and the visibility of the work output. On the other hand, senior collaborators might also free-ride on the efforts of more junior coauthors. In order to study the influence of these mechanisms in our context, we distinguish in Table 9 between collaborating with a more junior scientist, collaborating with a scientist of the same rank or collaborating with a more senior PI. Our results are more consistent with the free-riding mechanism. Model (9-1) shows that collaborating with a more senior person does not increase quality but does have a cost on productivity, especially if the collaboration is inter-departmental. Models (9-4) to (9-6) show that scientists seem to perform less well when they collaborate with someone who is senior to them. Like collaborations with more senior coauthors, collaborations with more junior ones are not associated with a statistically significant gain in output quality. However, Models (9-2) and (9-3) show that the productivity cost in this case appears considerably lower, leading to a relatively more positive impact of collaboration on attributed citations (Models (9-4) to (9-6)). Interestingly, the positive impact of collaboration on quality and its negative impact on productivity are particularly salient in the case of collaboration with someone of the same rank.

Overall, our analysis of the mechanisms driving our main results provides a richer picture of the micro-foundations of the apparent quality-productivity tradeoff associated with collaboration choices in creative work. Collaboration is particularly associated with higher quality output when it provides more opportunities for cross-fertilization of ideas by bringing together scientists of a similar rank and from different backgrounds. Collaboration also involves a significant productivity loss (at least if $\alpha(N) < 1$). This loss is particularly salient if the coauthor is of a more senior rank. Interestingly, opportunities for division of labor—through cross-departmental collaboration—seem to diminish this productivity loss. In sum, these results indicate that the apparent tradeoff between quality and quantity associated with scientists' collaboration choices might be driven by the decision to allocate one's time in a way that might lead to coordination costs and free-riding but that might also foster cross-fertilization of ideas and a productive division of labor.

4.6. DISCUSSION AND CONCLUSIONS

Considering collaboration at the level of the individual provides insights into the reasons why autonomous creative workers choose to work together or with others with different expertise or with different positions in the status hierarchy. As any researcher knows, the decision to collaborate is endogenous and the focus on creative output (e.g., publications, patents) in prior studies conceals important potential variables that contour collaborative choices. Only through a simultaneous exploration of the benefits and costs of collaboration for individuals can we really seek to understand the phenomenon of collaboration. In this paper, we have taken a step in this direction by developing a theoretical model of collaboration that considers both the potential benefits in terms of productivity, but also the coordination costs and the costs in terms of credit allocation among individuals. Our empirical focus on individual choices over a period of time enables us to hold "talent" constant, thus overcoming (to some extent), the heterogeneous nature of individual knowledge workers. We thus explore these tradeoffs in the organization of scientific work by considering a scientist's decision to allocate her fixed time to more or less collaborative projects.

We find that collaboration is associated with important tradeoffs, including higher quality publications, lower individual productivity and disproportionate credit attribution—i.e. that credit for a given collaborative paper is shared across coauthors in a way that sums up to more than 1. The size of these effects is considerable. A scientist working during a year with one other person on average rather than working alone can hope to receive over 60% more citations per published paper. They will also be able to show more publications on their CV despite the fact that their fractional productivity, in fact, decreased by over 15%--indicating that the two together publish considerably less than they would, had they worked separately. Importantly, scientists' collaboration behavior is consistent with a credit premium of over 33%[28] for collaborating with one person per year on average as opposed to working alone. Taken together, these results suggest that the "net value" of collaboration in creative work might be superior for the credit-seeking worker than it is for the output-focused manager or policy-maker. The benefits of collaborations are particularly high and its costs are particularly low when the collaboration brings together individuals having different skills and perspectives—as in the case of cross-departmental collaborations. On the other hand, the drawbacks of collaboration are particularly salient when scientists collaborate with a person that is senior to them. We find no evidence that junior scientists benefit from collaborating from somebody that is more established in their field.

We regard our results as an important first step in bringing the perspective of the time-conscious and credit-seeking knowledge worker to the debate on collaboration and creativity. Despite the recent surge in interest in collaboration for creativity by organization scholars, the large majority of these studies has focused on the output from the collaboration (typically the quality of the work completed). Our contribution was made possible by a departure from previous studies that have examined how to optimize the quality of a given piece of work. Indeed, we have complemented this approach by switching unit of analysis and focusing on the creative worker's decision to spend their time working alone or in groups of smaller or larger

[28] Percentages were computed using estimates shown in Figure 2. Average credit attribution (yearly citations received) for a collaboration of two when $a(N) = 1/N$ is 18 citations by 2008. It is 24 citations when $a(N) = 1/\sqrt{N}$. In comparison, credit received for the average sole authored year is about 15 citations.

sizes. In so doing, we were able to disentangle a number of tradeoffs associated with collaboration and creativity.

This research is not without its limitations: First, the decision to collaborate in smaller or larger groups is a complex one and is likely to involve other considerations than output quality, individual productivity and credit allocation. For instance, collaboration might have a financial cost or be endeavored to learn rather than to maximize individual credit. Second, we are not able to directly measure the scientists' credit allocation function—which is likely to vary substantially across disciplines (see for instance Maciejovsky et al. 2009). Instead we study that function indirectly by first developing a formal model of scientific collaborative choice and then by testing whether its predictions are consistent with the behavior that we observe in our data. Third, absent an experimental design, we cannot be sure that our empirical results are not at least partially driven by task heterogeneity. This endogeneity could be particularly problematic if tasks that can only accommodate large groups were important in ways that cannot be captured through paper citations or publication journal impact factor. Nonetheless, our setting provides the advantage of presenting the real choices made by creative workers in a number of disciplines over three decades. Our theoretical model generates predictions that are consistent with scientists' behavior and are robust across a variety of specifications. The fact that our findings are obtained after including individual fixed effects is also important– in other words our approach accounts for the variation in the choices made by the same scientist over the course of their career. Analysis of different types of collaborator further illuminates the various mechanisms underlying the tradeoffs that we observe.

Our study is only a first step toward understanding the benefits and drawbacks of collaboration for creative workers. Other correlates are likely to shape collaboration's "net value" beyond what we can observe in our data. On the input side, the amount of financial resources and equipment necessary for a given task are likely to vary with group size (Beaver & Rosen 1978). On the output side, collaboration is often described as a particularly enjoyable organization of work and a paramount driver of circulation of ideas, and learning (J. S. Katz & Martin 1997). Overall, the relationship between these costs and benefits is likely to depend on the group's intensity, structure, and experience (e.g., Porac et al. 2004). These are all important nuances which are likely to impact the net value of collaboration, and which we have not been able to study here. The importance of continuing the investigation of collaboration in the context of creative work should not be understated. As the nature of work is changing, more attention might usefully be brought to the fact that in practice, collaboration is an organization of work that has complex and perhaps distinct implications for creative workers on the one hand and managers and policy-makers on the other.

4.7. REFERENCES

Adams, J.D. et al., 2005. Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999. *Research Policy*, 34(3), pp.259–285.

Adams, J.D. & Griliches, Z., 1998. Research productivity in a system of universities. *Annales d'Economie et de Statistique*, pp.127–162.

Allen, T.J., 1978. *Managing the flow of technology*, MIT press Cambridge, MA.

Amabile, T.M. et al., 1996. Assessing the Work Environment for Creativity. *The Academy of Management Journal*, 39(5), pp.1154–1184.

Ancona, D.G. & Caldwell, D.F., 1992. Bridging the Boundary: External Activity and Performance in Organizational Teams. *Administrative Science Quarterly*, 37(4), pp.634–665.

Beaver, D.B., 2001. Reflections on scientific collaboration (and its study): past, present, and future. *Scientometrics*, 52(3), pp.365–377.

Beaver, D.B. & Rosen, R., 1978. Studies in scientific collaboration. *Scientometrics*, 1(1), pp.65–84.

Becker, G.S. & Murphy, K.M., 1992. The Division of Labor, Coordination Costs, and Knowledge. *The Quarterly Journal of Economics*, 107(4), pp.1137–1160.

Biagioli, M., 2003. Rights or Rewards. In *Scientific Authorship: credit and intellectual property in science*. pp. 253–279.

Bordons, M. et al., 1996. Local, domestic and international scientific collaboration in biomedical research. *Scientometrics*, 37(2), pp.279–295.

Bornmann, L. & Daniel, H.-D., 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), pp.45–80.

Burt, R.S., 2004. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2), pp.349–399.

Cattani, G. & Ferriani, S., 2008. A Core/Periphery Perspective on Individual Creative Performance: Social Networks and Cinematic Achievements in the Hollywood Film Industry. *Organization Science*, 19(6), pp.824–844.

Cawkell, A.E., 1976. Citations, obsolescence, enduring articles, and multiple authorships. *Journal of Documentation*, 32(1).

Cronin, B., Shaw, D. & La Barre, K., 2004. Visible, Less Visible, and Invisible Work: Patterns of Collaboration in 20th Century Chemistry. *Journal of the American Society for Information Science & Technology*, 55(2), pp.160–168.

Cummings, J.N., 2004. Work Groups, Structural Diversity, and Knowledge Sharing in a Global Organization. *Management Science*, 50(3), pp.352–364.

Cummings, J.N. & Kiesler, S., 2007. Coordination costs and project outcomes in multi-university collaborations. *Research Policy*, 36(10), pp.1620–1634.

Dasgupta, P. & David, P.A. 1984. Towards a new economics of science. *Research Policy*, 23, pp.487-521.

Diehl, M. & Stroebe, W., 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology* 53.3 (1987): 497-509.

Ding, W.W. et al., 2010. The Impact of Information Technology on Academic Scientists' Productivity and Collaboration Patterns. *Management Science*, 56(9), pp.1439 –1461.

Dumaine, B., 1994. The Trouble with Teams. *Fortune*.

Dumaine, B. & Gustke, C., 1990. Who Needs a Boss? *Fortune*.

Edmondson, A., 1999. Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly*, 44(2), pp.350–383.

Engers, M., Gans, J.S., Grant, S. & King, S.P., 1999. First Author Conditions. *Journal of Political Economy*, 107 (4), pp.859-883.

Fleming, L., 2007. Breakthroughs and the "Long Tail" of innovation. *MIT Sloan Management Review*, 49(1), p.69.

Fleming, L., Mingo, S. & Chen, D., 2007. Collaborative Brokerage, Generative Creativity, and Creative Success. *Administrative Science Quarterly*, 52(3), pp.443–475.

Furman, J. & Stern, S., 2011. Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review*, 101(5), pp.1933–1963.

Gans, J.S. & Murray, F. 2013. Markets for Scientific Attribution. *mimeo.*, MIT.

Gilfillan, S.C., 1935. *Inventing the ship*, Follett Chicago, IL.

Girotra, K., Terwiesch, C. & Ulrich, K.T., 2010. Idea Generation and the Quality of the Best Idea. *Management Science*, 56(4), pp.591–605.

Hara, N. et al., 2003. An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information Science and Technology*, 54(10).

Hargadon, A., 2008. Creativity That Works. In *Handbook of Organizational Creativity*. Psychology Press.

Hargadon, A., 2003. *How breakthroughs happen: The surprising truth about how companies innovate*, Harvard Business School Press.

Hargadon, A. & Bechky, B., 2006. When Collections of Creatives Become Creative Collectives: A Field Study of Problem Solving at Work. *Organization Science*, 17(4), pp.484–500.

Hargadon, A. & Sutton, R., 1997. Technology Brokering and Innovation in a Product Development Firm. *Administrative Science Quarterly*, 42(4), pp.716–749.

Häussler, C. & Sauermann, H., 2013. Credit Where Credit is Due? The Impact of Project Contribution and Social Factors on Authorship and Inventorship. *Research Policy*, 42(3), pp.688-703.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), pp.16569–16572.

Hoerr, J., 1989. The Payoff from Teamwork: The Gains in Quality are Substantial-So Why Isn't It Spreading Faster. *Business Week (European Edition)*, pp.36–42.

Holmstrom, B. 1982. Moral Hazard in Teams. *Bell Journal of Economics*, 13, pp.324-340.

Jiang, L., Thursby, J. and Thursby, M. 2012. Scientific Disclosure and the Faces of Competition. REER Conference Presentation, Georgia Tech.

Johansson, F., 2004. *The Medici effect: breakthrough insights at the intersection of ideas, concepts, and cultures*, Harvard Business Press.

Jones, B., 2009. The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? *Review of Economic Studies*, 76(1), pp.283–317.

Jones, B., Wuchty, S. & Uzzi, B., 2008. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science*, 322(5905), p.1259.

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

Katzenbach, J.R., Smith, D.K. & Bookspan, M., 1993. *The wisdom of teams*, Harvard Business School Press Boston.

Knorr-Cetina, K., 1999. *Epistemic Cultures: How the Sciences Make Knowledge*, Harvard University Press Cambridge, MA.

Koplowitz, R. et al., 2009. Benchmarking Your Collaboration Strategy. Available at: http://www.forrester.com/rb/Research/benchmarking_collaboration_strategy/q/id/48336/t/2.

Landry, R. & Amara, N., 1998. The impact of transaction costs on the institutional structuration of collaborative academic research. *Research Policy*, 27(9), pp.901–913.

Latour, B. & Woolgar, S., 1986. *Laboratory life: The construction of scientific facts*, Princeton University Press.

Leahey, E., 2007. Not by productivity alone: How visibility and specialization contribute to academic earnings. *American sociological review*, 72(4), pp.533–561.

Lee, S. & Bozeman, B., 2005. The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science*, 35(5), pp.673–702.

Maciejovsky, B., Budescu, D.V. & Ariely, D., 2009. Research Note--The Researcher as a Consumer of Scientific Publications: How Do Name-Ordering Conventions Affect Inferences About Contribution Credits? *Marketing Science*, 28(3), pp.589–598.

McAfee, R.P. & McMillan, J., 1991. Optimal Contracts for Teams. *International Economic Review*, 32, pp.561-577.

Melin, G., 2000. Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, 29(1), pp.31–40.

Merton, R.K., 1968. The Matthew Effect in Science The reward and communication systems of science are considered. *Science*, 159(3810), p.56.

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

Newman, M.E.J., 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2) pp.404-409.

Obstfeld, D., 2005. Social networks, the tertius iungens orientation, and involvement in innovation. *Administrative Science Quarterly*, 50(1), pp.100–130.

Orsburn, J.D. & Moran, L., 2000. *The new self-directed work teams: Mastering the challenge*, McGraw-Hill.

Paulus, P.B., 2007. Fostering creativity in groups and teams. *The handbook of organizational creativity*, pp.159–182.

Paulus, P.B. & Nijstad, B.A., 2003. *Group creativity: Innovation through collaboration*, Oxford University Press.

Porac, J.F. et al., 2004. Human capital heterogeneity, collaborative relationships, and publication patterns in a multidisciplinary scientific alliance: a comparative case study of two scientific teams. *Research Policy*, 33(4), pp.661–678.

Price, D.J. & Beaver, D.B., 1966. Collaboration in an invisible college. *American Psychologist*, 1966, pp.1011–8.

Reagans, R., Argote, L. & Brooks, D., 2005. Individual Experience and Experience Working Together: Predicting Learning Rates from Knowing Who Knows What and Knowing How to Work Together. *Management Science*, 51(6), pp.869–881.

Reagans, R. & Zuckerman, E.W., 2001. Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams. *Organization Science*, 12(4), pp.502–517.

Salganik, M.J., Dodds, P.S. & Watts, D.J., 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762), pp.854–856.

Simcoe, T.S. & Waguespack, D.M., 2011. Status, Quality, and Attention: What's in a (Missing) Name? *Management Science*, 57(2), pp.274 –290.

Singh, J. & Fleming, L., 2010. Lone Inventors as Sources of Breakthroughs: Myth or Reality? *Management Science*, 56(1), pp.41–56.

Stein, M.I., 1953. Creativity and culture. *Journal of Psychology*, 36(2), pp.311–322.

Stokols, D. et al., 2005. In vivo studies of transdisciplinary scientific collaboration Lessons learned and implications for active living Research. *American journal of preventive medicine*, 28(2S2), pp.202–213.

Subramanyam, K., 1983. Bibliometric studies of research collaboration: A review. *Journal of information Science*, 6(1), p.33.

Taylor, A. & Greve, H.R., 2006. Superman or the fantastic four? Knowledge combination and experience in innovative teams. *Academy of Management Journal*, 49(4), p.723.

Tushman, M.L. & Katz, R., 1980. External Communication and Project Performance: An Investigation into the Role of Gatekeepers. *Management Science*, 26(11), pp.1071–1085.

Uzzi, B. & Spiro, J., 2005. Collaboration and Creativity: The Small World Problem. *American Journal of Sociology*, 111(2), pp.447–504.

Valderas, J.M., 2007. Why do team-authored papers get cited more? *Science (New York, NY)*, 317(5844), p.1496.

Von Hippel, E.A., 2003. *Democratizing Innovation.* MIT Press Book

Woodman, R.W., Sawyer, J.E. & Griffin, R.W., 1993. Toward a Theory of Organizational Creativity. *The Academy of Management Review*, 18(2), pp.293–321.

Wray, B.K., 2002. The Epistemic Significance of Collaborative Research. *Philosophy of Science*, 69(1), pp.150–168.

Wuchty, S., Jones, B. & Uzzi, B., 2007. The Increasing Dominance of Teams in Production of Knowledge. *Science*, 316(5827), p.1036.

Zare, R.N., 1997. Editorial: Knowledge and Distributed Intelligence. *Science*, 275(5303), p.1047.

Zuckerman, H., 1972. Age, Aging, and Age Structure in Science, reprinted in: Robert K. Merton, 1973. *The Sociology of Science.*

4.8. TABLES & FIGURES

TABLE 1: DESCRIPTIVE STATISTICS AT THE INDIVIDUAL-YEAR AND PUBLICATION LEVELS

Variable	PUBLICATION LEVEL (n=21,054)				INDIVIDUAL-YEAR LEVEL (n=5,964)			
	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
Average Group Size (*Nauthors*)	4.1	2.3	1	20	3.8	1.9	1	20
Average Forward Citations (*Cites*)	46.8	111.9	0	4810	41.3	86.9	0	2595
Productivity -- alpha(N)=1 (*NPubs*)	n.a.	n.a.	n.a.	n.a.	3.5	3.7	1	47
Productivity -- alpha(N)=1/N (*Frac_Pubs*)	n.a.	n.a.	n.a.	n.a.	1.1	1.1	0.1	9.7
Att_Cites -- alpha(N)=1	46.8	111.9	0	4810	165.1	388.8	0	8852
Att_Cites -- alpha(N)=1/sqrt(N)	24.5	67.2	0	4810	86.4	211.7	0	4947
Att_Cites -- alpha(N)=1/N	13.8	51	0	4810	48.7	135.4	0	4819
N_ Highly Cited Publications	0.05	0.22	0	1	0.18	0.52	0	8
Frac_Highly Cited Publications	0.01	0.07	0	1	0.05	0.17	0	2
Average JIF (missing values imputed)	4.86	5.57	0	52.59	4.46	4.62	0.024	29.89
Average JIF (missing values 0)	4.1	5.82	0	52.59	3.57	4.76	0	29.89
Last Authored Paper	0.63	0.48	0	1	0.61	0.39	0	1
First Authored Paper	0.08	0.27	0	1	0.12	0.29	0	1
Year	1995	8.6	1976	2006	1993.2	8.8	1976	2006

TABLE 2: CORRELATION TABLE, MAIN VARIABLES, INDIVIDUAL-YEAR LEVEL (5,964 observations)

	1	2	3	4	5	6	7	8
1. *NAuthors* (mean group size)	1							
2. *Cites* (av. forward cites)	0.0944*	1						
3. *NPubs*	0.1254*	0.0602*	1					
4. *Frac_Pubs* -- alpha(N)=1/N	-0.1414*	0.0415	0.9026*	1				
5. *Att_Cites* -- alpha(N)=1	0.1068*	0.6097*	0.5092*	0.4712*	1			
6. *Att_Cites* -- alpha(N)=1/sqrt(N)	0.0414	0.6210*	0.4742*	0.4786*	0.9711*	1		
7. *Att_Cites* -- alpha(N)=1/N	-0.011	0.6095*	0.4022*	0.4415*	0.8794*	0.9655*	1	
8. Year	0.2301*	-0.1416*	0.1967*	0.1149*	-0.0520*	-0.0742*	-0.0848*	1

Significance level: * p<0.001

TABLE 3: THE EFFECT OF COLLABORATION CHOICE ON QUALITY, QUANTITY & CREDIT

	QUALITY	QUANTITY		CREDIT		
	DV=log(1+ *Cites* by paper)	DV=log(1+*Pubs*)		DV=log(1+*Att_Cites*)		
		NPubs	*Frac_Pubs*	alpha(N)=1	alpha(N)=1/sqrt(N)	alpha(N)=1/N
	(3-1)	(3-2)	(3-3)	(3-4)	(3-5)	(3-6)
Department-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Scientist FE	Yes	Yes	Yes	Yes	Yes	Yes
Career Stage FE	Yes	Yes	Yes	Yes	Yes	Yes
Group Size	0.0990***	-0.00582	-0.0688***	0.0919***	-0.00303	-0.0834***
	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
Constant	3.127***	0.864***	0.653***	3.405***	3.253***	3.079***
	(0.17)	(0.05)	(0.03)	(0.19)	(0.18)	(0.17)
Observations	5964	5964	5964	5964	5964	5964
R-squared	0.24	0.136	0.21	0.157	0.16	0.183

OLS; Robust standard errors in parentheses are clustered at the level of the individual MIT scientist

Significance level: *** p<0.01, ** p<0.05, * p<0.1

TABLE 4: ROBUSTNESS CHECKS (1)

	JOURNAL IMPACT FACTOR	CREDIT GIVEN TO LAST AUTHOR ONLY†		
	QUALITY	QUALITY	QUANTITY	YEARLY CITATIONS‡
	OLS; DV= Av. JIF for the year; missing JIF imputed based on pub cites	OLS; DV= log(1+*Cites*)	OLS; DV=log(1+ *NPubs*-LastAuthor)	OLS; DV= DV=log(1+ *Att_Cites*-LastAuthor)
	(4-1)	(4-2)	(4-3)	(4-4)
Department-Year FE	Yes	Yes	Yes	Yes
Scientist FE	Yes	Yes	Yes	Yes
Career Stage FE	Yes	Yes	Yes	Yes
Group Size	0.387***	0.117***	-0.0846***	-0.199***
	(0.06)	(0.02)	(0.01)	(0.02)
Constant	4.456***	2.890***	0.708***	2.832***
	(0.44)	(0.23)	(0.06)	(0.24)
Observations	5,964	2,265	5,964	5,964
R-squared	0.09	0.28	0.16	0.131

OLS; Robust standard errors in parentheses are clustered at the level of the individual MIT scientist

Significance level: *** p<0.01, ** p<0.05, * p<0.1

† Since (total) yearly collaboration is of interest, the "Group Size" variable was not recalculated to include only last-authored paper

‡ All cites attributed to last author only

TABLE 5: ROBUSTNESS CHECKS (2)

	CONCAVENESS		TOP QUALITY PUBLICATIONS		
	QUANTITY	CREDIT	CREDIT		
	OLS; DV=log(1+ *NPubs*)	OLS; DV=log(1+ *Att_Cites*) N=1	OLS; DV=log(1+*Att_Cites*); alpha(N)=1	OLS; DV=log(1+*Att_Cites*); alpha(N)=1/sqrt(N)	OLS; DV=log(1+*Att_Cites*); alpha(N)=1/N
	(5-1)	(5-2)	(5-3)	(5-4)	(5-5)
Department-Year FE	Yes	Yes	Yes	Yes	Yes
Scientist FE	Yes	Yes	Yes	Yes	Yes
Career Stage FE	Yes	Yes	Yes	Yes	Yes
Group Size	0.0492***	0.250***	0.00457**	-0.000532	-0.00247***
	(0.01)	(0.04)	(0.00)	(0.00)	(0.00)
Group Size Squared	-0.00454***	-0.0130***			
	(0.00)	(0.00)			
Constant	0.782***	3.098***	0.0921***	0.0767***	0.0599***
	(0.05)	(0.21)	(0.03)	(0.02)	(0.01)
Observations	5,964	5,964	5,964	5,964	5,964
R-squared	0.142	0.164	0.041	0.036	0.036

OLS; Robust standard errors in parentheses are clustered at the level of the individual MIT scientist

Significance level: *** p<0.01, ** p<0.05, * p<0.1

TABLE 6: WITHIN MIT COLLABORATION -- DESCRIPTIVE STATISTICS

Variable	PUBLICATION LEVEL (n=4,617 out of 21,054)				INDIVIDUAL-YEAR LEVEL (n=2,273 out of 5,964)			
	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
Overall Group Profile								
# PIs	1.3	0.5	1	4	1.3	0.5	1	4
# non-PIs	1.7	1.3	0	10	1.7	1.2	0	10
Breakdown of Collaborating PIs by Position of the Collaborator								
# junior PIs	0.05	0.2	0	3	0.06	0.2	0	3
# PIs with same position	0.14	0.4	0	3	0.13	0.3	0	3
# senior PIs	0.07	0.3	0	3	0.08	0.3	0	3
Breakdown of Collaborating PIs by Department of the Collaborator								
# PIs from the same department	0.11	0.3	0	3	0.19	0.4	0	3
# PIs from a different department	0.15	0.4	0	3	0.09	0.3	0	2

TABLE 7: THE EFFECT OF PI COLLABORATION CHOICES WITHIN MIT

	QUALITY	QUANTITY		CREDIT		
	DV=log(1+ *Cites* by paper)	DV=log(1+*Pubs*)		DV=log(1+*Att_Cites*)		
		NPubs	*Frac_Pubs*	alpha(N)=1	alpha(N)=1/sqrt(N)	alpha(N)=1/N
	(7-1)	(7-2)	(7-3)	(7-4)	(7-5)	(7-6)
Department-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Scientist FE	Yes	Yes	Yes	Yes	Yes	Yes
Career Stage FE	Yes	Yes	Yes	Yes	Yes	Yes
# PIs	0.196***	-0.0199	-0.297***	0.169**	-0.0945	-0.343***
	(0.07)	(0.02)	(0.02)	(0.08)	(0.08)	(0.07)
# non-PIs	0.119***	0.00134	0.00214	0.121***	0.120***	0.120***
	(0.03)	(0.01)	(0.01)	(0.04)	(0.04)	(0.04)
Constant	1.573***	0.764***	1.042***	1.641***	1.943***	2.232***
	(0.24)	(0.09)	(0.08)	(0.30)	(0.29)	(0.29)
Observations	2273	2273	2273	2273	2273	2273
R-squared	0.245	0.18	0.273	0.221	0.222	0.233

OLS; Robust standard errors in parentheses are clustered at the level of the individual MIT scientist

Significance level: *** p<0.01, ** p<0.05, * p<0.1

TABLE 8: THE EFFECT OF PI COLLABORATION WITHIN AND ACROSS DEPARTMENTS

	QUALITY	QUANTITY		CREDIT		
	DV=log(1+ *Cites* by paper)	DV=log(1+*Pubs*)		DV=log(1+*Att_Cites*)		
		NPubs	*Frac_Pubs*	alpha(N)=1	alpha(N)=1/sqrt(N)	alpha(N)=1/N
	(8-1)	(8-2)	(8-3)	(8-4)	(8-5)	(8-6)
Department-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Scientist FE	Yes	Yes	Yes	Yes	Yes	Yes
Career Stage FE	Yes	Yes	Yes	Yes	Yes	Yes
# PIs from the same department	0.104	-0.0421	-0.323***	0.0488	-0.214**	-0.465***
	(0.10)	(0.03)	(0.03)	(0.11)	(0.10)	(0.10)
# PIs from a different department	0.281***	0.000757	-0.273***	0.282**	0.0175	-0.230**
	(0.10)	(0.03)	(0.02)	(0.11)	(0.11)	(0.10)
# non-PIs	0.116***	0.000517	0.0012	0.116***	0.116***	0.115***
	(0.03)	(0.01)	(0.01)	(0.04)	(0.04)	(0.04)
Constant	1.706***	0.750***	0.749***	1.767***	1.792***	1.819***
	(0.22)	(0.07)	(0.07)	(0.26)	(0.26)	(0.25)
Observations	2273	2273	2273	2273	2273	2273
R-squared	0.246	0.18	0.274	0.222	0.223	0.234

OLS; Robust standard errors in parentheses are clustered at the level of the individual MIT scientist

Significance level: *** p<0.01, ** p<0.05, * p<0.1

TABLE 9: MECHANISM -- THE EFFECT OF PI COLLABORATION WITH DIFFERENT PIs (within MIT only)

	QUALITY	QUANTITY		CREDIT		
	DV=log(1+ *Cites* by paper)	DV=log(1+*Pubs*)		DV=log(1+*Att_Cites*)		
		NPubs	*Frac_Pubs*	alpha(N)=1	alpha(N)=1/sqrt(N)	alpha(N)=1/N
	(9-1)	(9-2)	(9-3)	(9-4)	(9-5)	(9-6)
Department-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Scientist FE	Yes	Yes	Yes	Yes	Yes	Yes
Career Stage FE	Yes	Yes	Yes	Yes	Yes	Yes
Within Department Collaboration						
Junior PI	0.18	-0.02	-0.293***	0.14	-0.12	-0.378*
	(0.18)	(0.04)	(0.05)	(0.21)	(0.20)	(0.20)
Same rank PI	0.14	-0.0726*	-0.325***	0.06	-0.18	-0.399***
	(0.13)	(0.04)	(0.04)	(0.14)	(0.13)	(0.13)
Senior PI	-0.19	0	-0.234***	-0.2	-0.411*	-0.612***
	(0.20)	(0.06)	(0.06)	(0.23)	(0.22)	(0.21)
Across Department Collaboration						
Junior PI	0.704	0.187	-0.0639	0.996**	0.737	0.502
	(0.47)	(0.12)	(0.12)	(0.50)	(0.48)	(0.45)
Same rank PI	0.671***	0.037	-0.252***	0.733**	0.44	0.162
	(0.26)	(0.07)	(0.06)	(0.30)	(0.29)	(0.29)
Senior PI	-0.0428	-0.00984	-0.232***	-0.0776	-0.341	-0.58
	(0.39)	(0.08)	(0.08)	(0.42)	(0.42)	(0.42)
# non-PIs	0.125***	-0.000041	-0.00772	0.125***	0.116***	0.108***
	(0.03)	(0.01)	(0.01)	(0.04)	(0.04)	(0.03)
Constant	1.825***	0.739***	0.721***	1.861***	1.881***	1.903***
	(0.22)	(0.09)	(0.08)	(0.29)	(0.28)	(0.27)
Observations	2273	2273	2273	2273	2273	2273
R-squared	0.248	0.182	0.239	0.225	0.227	0.234

OLS; Robust standard errors in parentheses are clustered at the level of the individual MIT scientist

Significance level: *** p<0.01, ** p<0.05, * p<0.1

FIGURE 1: DISTRIBUTION OF GROUP SIZES BY FACULTY-YEARS (DESCRIPTIVE STATISTICS)

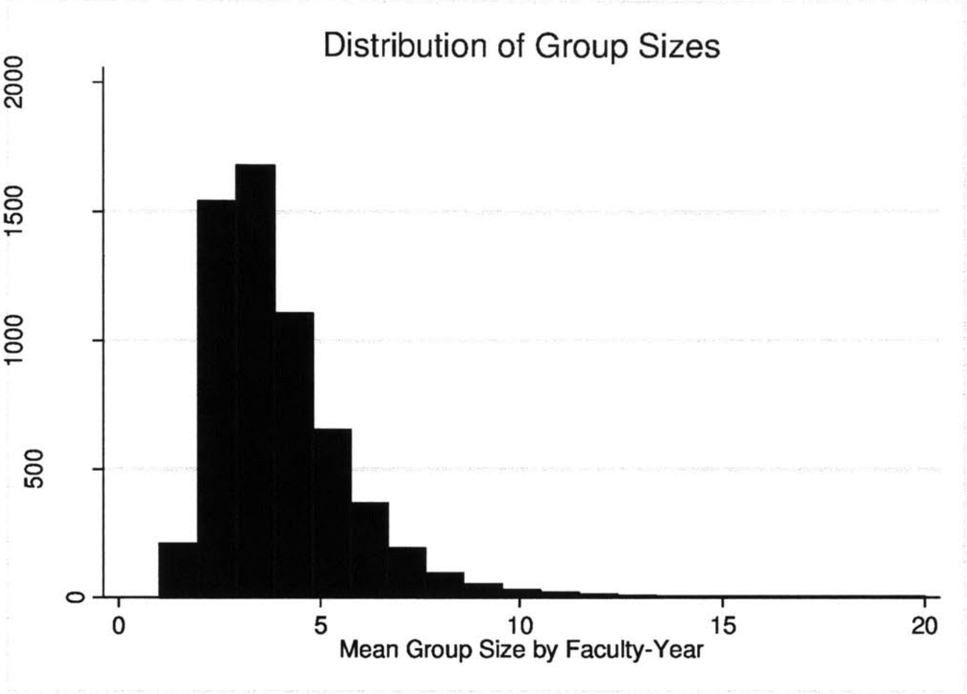
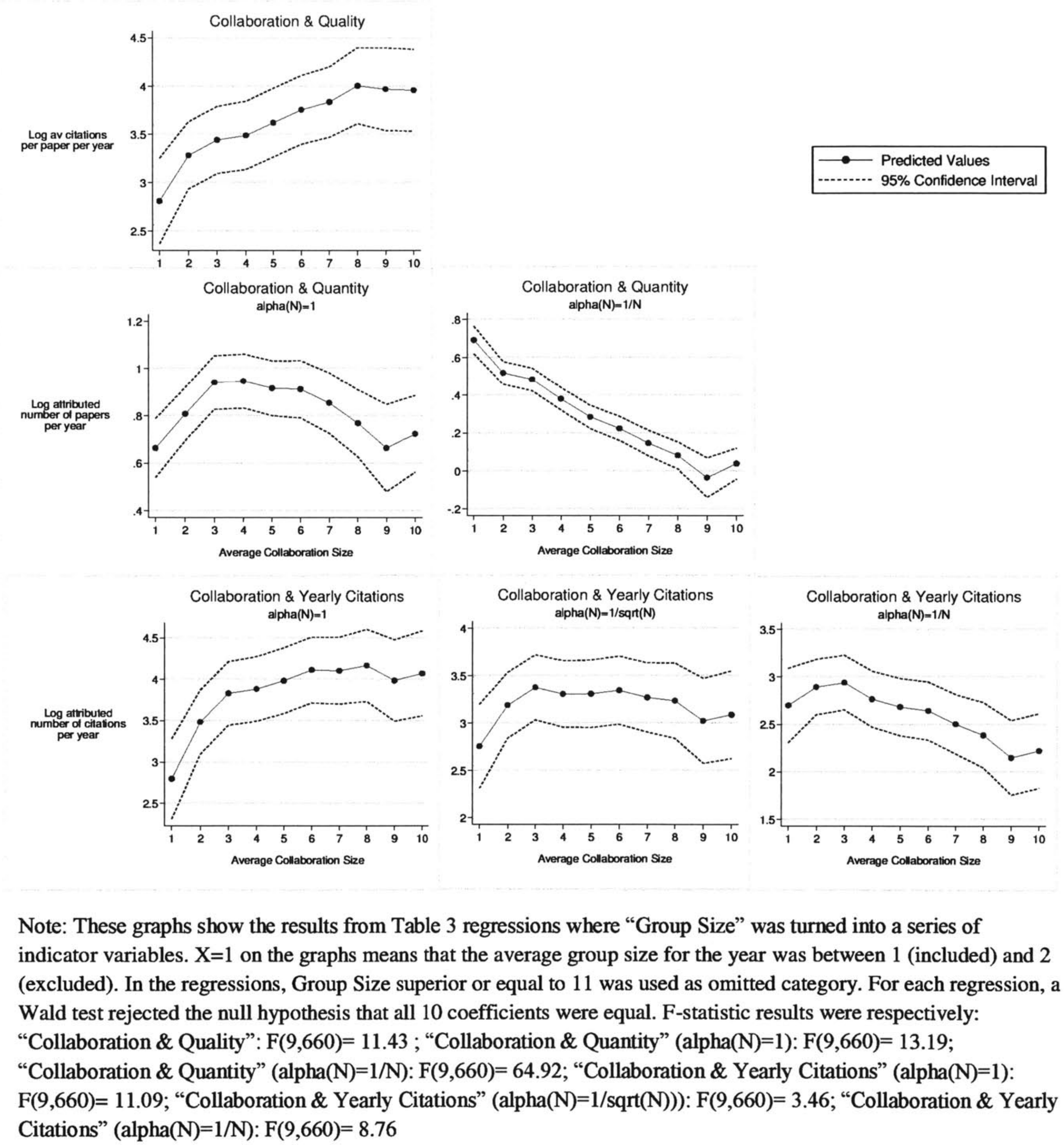


FIGURE 2: RELATIONSHIP BETWEEN COLLABORATION, QUALITY, QUANTITY AND OVERALL YEARLY CITATIONS (ESTIMATES)



Note: These graphs show the results from Table 3 regressions where "Group Size" was turned into a series of indicator variables. X=1 on the graphs means that the average group size for the year was between 1 (included) and 2 (excluded). In the regressions, Group Size superior or equal to 11 was used as omitted category. For each regression, a Wald test rejected the null hypothesis that all 10 coefficients were equal. F-statistic results were respectively: "Collaboration & Quality": F(9,660)= 11.43 ; "Collaboration & Quantity" (alpha(N)=1): F(9,660)= 13.19; "Collaboration & Quantity" (alpha(N)=1/N): F(9,660)= 64.92; "Collaboration & Yearly Citations" (alpha(N)=1): F(9,660)= 11.09; "Collaboration & Yearly Citations" (alpha(N)=1/sqrt(N))): F(9,660)= 3.46; "Collaboration & Yearly Citations" (alpha(N)=1/N): F(9,660)= 8.76