COLLABORATION AND CREATIVITY:

EFFECTS OF TIE STRENGTH

A Dissertation
Presented to
The Academic Faculty

by

Jian Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Public Policy

Georgia Institute of Technology
December 2013

COLLABORATION AND CREATIVITY:

EFFECTS OF TIE STRENGTH

Approved by:

Dr. Diana Hicks, Advisor
School of Public Policy
Georgia Institute of Technology

Dr. Julia Melkers
School of Public Policy
Georgia Institute of Technology

Dr. John Walsh
School of Public Policy
Georgia Institute of Technology

Dr. Juan Rogers
School of Public Policy
Georgia Institute of Technology

Dr. Yajun Mei
School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Sybille Hinze
Institute for Research Information and Quality Assurance (iFQ)

Date Approved: June 25, 2013

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee members, Diana Hicks, Julia Melkers, John Walsh, Juan Rogers, Yajun Mei, and Sybille Hinze, for their advice and support for my dissertation work. I thank the NETWISE team for its collective effort in the development of the NETWISE I data and iFQ for giving me the access to Web of Science data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation studies the relationship between collaboration networks and scientific creativity. It finds significant knowledge spillover from new collaborations to repeated collaborations, and proposes a network approach to understand scientific creativity at the egocentric network level beyond the boundary of teams. To understand the network effect (specifically, effects of tie strength) on creativity, it integrates literature on small groups and social networks and adopts a creative-process model. An inverted U-shaped relationship between tie strength and creativity is observed, because of the mixed impacts of tie strength at different stages of the creative process. Furthermore, it explores the effect of tie configurations and finds that the skewness of tie strength distribution moderates the effect of tie strength. In addition, it also tests two competing explanations for the association between strong tie and low creativity: creativity-decline hypothesis versus cost-reduction hypothesis. Finally, there is no evidence that collaboration networks would raise the visibility of previously published papers, but there is a significant prestige effect in gaining citations.

CHAPTER 1

INTRODUCTION

It is creativity that drives the advancement in science and the progress in many other aspects of human lives and society, and collaboration plays an important role in scientific creativity. This dissertation contributes to the understanding of scientific creativity and the effects of collaboration networks on creativity.

Creativity has become an important topic in the field of psychology since Guilford's 1949 presidential address to the American Psychological Association, and initial creativity studies focused primarily on personal traits as determinants of creativity (Barron & Harrington, 1981; Guilford, 1950; Mednick, 1962; Simonton, 1999), such as sensitivity to problems, ideational fluency, flexibility of set, ideational novelty, synthesizing ability, analyzing ability, reorganizing or redefining ability, span of ideational structure, and evaluating ability. In addition to personal traits, creativity also depends on a number of social and environmental factors, such as group composition and organizational culture. Psychologists have also devoted to investigating group structure, group process, and their effects on group creativity (Hackman & Morris, 1975; King & Anderson, 1990; Levine & Moreland, 2004; Paulus & Nijstad, 2003). Furthermore, they have integrated personality-, cognitive-, and social- psychology explanations of creativity and proposed comprehensive frameworks for individual-, group-, and organizational-creativity (Amabile, 1983; Ford, 1996; Woodman, Sawyer, & Griffin, 1993).

While psychologists increasingly acknowledge the social underpinnings of creativity, sociologists of science emphasize the production of science as a social process, studying the organization and institution of modern sciences at the macro-level (Merton, 1973; Whitley, 2000) and laboratory settings at the micro-level (Latour & Woolgar, 1986). Following this tradition emphasizing institutions and organizations, some scholars have investigated organizational characteristics that facilitate highly creative or breakthrough research (Heinze, Shapira, Rogers, & Senker, 2009; Hollingsworth, 2004, 2009), such as complementary diversity, autonomy, and flexible funding.

Since scientific creativity comes from a social process, to understand scientific creativity, it is important to understand the social process of science production and particularly the transition in science production from individual-based to collaborative models. Scientific knowledge is increasingly created from collections of collaborators instead of solo researchers (de Solla Price, 1986; Wuchty, Jones, & Uzzi, 2007). Therefore, scientific creativity partly depends on the structure and process of collaboration. Furthermore, although a novel idea can often be traced to a flash of intuition from an individual or a brainstorm within a group, the intra-personal and intra-group thought-process is deeply embedded in broader social networks. Therefore, it is important to search for network explanations for scientific creativity. For example, Simonton (1984) argued that the understanding of creativity demands that the creative individual be placed within a network of interpersonal relationships.

Social networks have proven powerful in explaining a variety of phenomena, such as dropouts of high school students (Coleman, 1988), job-related rewards for individuals (Burt, 1992; Granovetter, 1973), commitment and satisfaction of employees (Krackhardt & Porter, 1985), survival of firms (Uzzi, 1996, 1997), and knowledge transfer within firms (Hansen, 1999; Reagans & McEvily, 2003). More recently, scholars have started exploring the effect of social networks on creativity (Fleming, Mingo, & Chen, 2007; McFadyen & Cannella, 2004; McFadyen, Semadeni, & Cannella, 2009; Perry-Smith & Shalley, 2003; Singh & Fleming, 2010; Uzzi & Spiro, 2005).

However, there are still many unanswered questions concerning the relationship between collaboration networks and creativity. First, it is unclear how a specific network structure leads to certain creativity outcomes. To bridge this gap between network structure and its effects, I integrate literature on small groups and social networks. Based on group-process literature, I highlight three steps in the collaboration creative process: idea generation, idea convergence, and idea implementation. The success of these stages requires different conditions, such as cognitive diversity, cognitive capital, and relational capital, and these conditions depends on certain network structures, specifically, the strength of ties in the network. This process model serves as a micro-foundation to explain the effect of network structures on scientific creativity (**Figure 1**).

Anther unanswered question pertains to the existence of many competing network theories. There have been long-standing debates between the weak tie theory (Granovetter, 1973, 1983) and the strong tie theory (Krackhardt, 1992; Uzzi, 1996, 1997), and between the structural hole theory (Burt, 1992, 2005) and the network closure theory (Coleman, 1988, 1990). These theories provide competing predictions about which type of networks are more advantageous to performance or other outcomes of interest, and both sides are supported by a number of empirical studies. A process-perspective helps to integrate these competing theories for a more coherent and comprehensive network theory. For example, tie strength has different effects at different stages of the creative process. Therefore, we may observe evidence favoring weak ties in some cases and evidence favoring strong ties in other cases, depending on which stage we are studying.

Given the prevailing research interest in studying teams, there is a fundamental question confronting my study of network effects: Is it legitimate to study creativity at the egocentric network level? After all, the dominating model of science production is team-based, and the prevailing norm is to study teams (Wuchty et al., 2007). I argue that the egocentric-network-level analysis is needed because of (1) the fuzzy boundaries of team responsible for a scientific output, (2) the fluidness of teams, and (3) interactions and knowledge spillovers between teams. Therefore, a considerable amount of creativity comes from activities outside of the team, and therefore it’s important to search for sources of creativity in dynamics egocentric collaboration network beyond the boundary of closed teams.

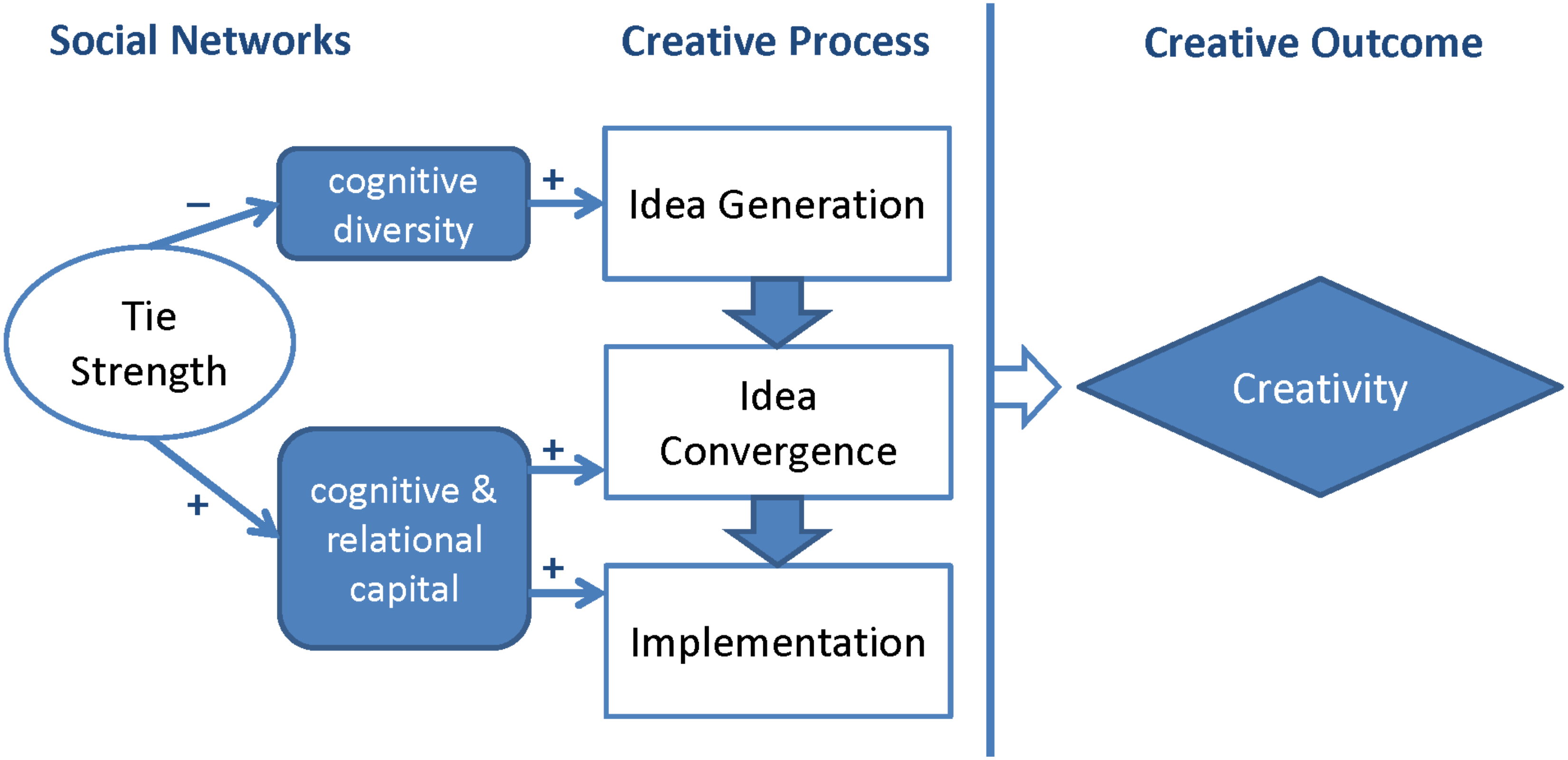


Figure 1. Theoretical Framework

After justifying the egocentric network approach, I investigate the effect of network-level average tie strength on creativity. Specifically, I hypothesize an inverted U-shaped relationship between tie strength and creativity, because of the mixed impacts of tie strength at different stages of the creative process. Furthermore, I explore the effect of network-level tie configuration, instead of taking a simple dichotomy between weak and strong ties. I hypothesize that a more skewed network is more creative than a less skewed one, when the network average tie strength is very high, and that tie strength skewness moderates the effect of network average tie strength.

I argue that strong-tie-collaborations are less creative because of the path-dependency and low cognitive diversity. However, the transaction cost theory provides an alternative explanation. Because the costs for collaborating with strong-tie-collaborators are low, it is "profitable" for scientists to do both trivial and experimental research with them. Therefore, we may observe that products of strong-tie-collaborations have lower average creativeness but also more likely to be highly creative. These two competing hypotheses are tested, and the creativity-decline but not the costs-reduction hypothesis is supported.

My analyses rely primarily on bibliometric data, which have some limitations. Therefore I have a replicate study based on survey data, conducting a factor analysis to construct a survey-based measure of tie strength, and testing if it concurs with the bibliometric measure of tie strength. In addition, I also validate some of my findings using survey data with richer contextual information.

Finally, I move to a different context to study the effect of collaboration networks on knowledge diffusion: Will current collaboration networks help to raise the visibility of a previously published paper? The focus here is no longer about knowledge creation. I test this collaborator-marketing effect and decompose three related effects: collaborator-marketing, prestige, and intellectual relevance. While there is no evidence of collaborator-marketing effect, there are significant prestige effects driving citations.

This dissertation is organized as follows. The "THEORY AND HYPOTHESES" section firstly reviews literature about scientific creativity, research collaboration, and creative process, to provide a theatrical foundation and framework for this study. Subsequently I propose hypotheses concerning knowledge spillover, tie strength effects, creativity vs. costs issues, and knowledge diffusion.

The "METHOD" section presents basic data information and discusses limitations of and corresponding treatments for using citation counts as proxies for creativity and coauthorship for collaboration. Mixed-effect models for hierarchical or panel data are frequently used, so this section also briefly introduces linear-, generalized-linear-, and quantile- mixed-effect models. Because different datasets are used for testing different hypotheses, details about model specification, sample restriction, and model estimation are not presented in this section but reported separately in the "RESULTS" section.

The "RESULTS" section is divided into five sub-sections, four of them are devoted to testing the four sets of hypotheses, and "A Replicate Study" reports the study using survey data.

Finally, the "CONCLUSION" section presents the intellectual structure (a roadmap) of this dissertation, summarizes major findings, and then discusses theoretical contributions and policy implications.

CHAPTER 2

THEORY AND HYPOTHESES

2.1. Scientific Creativity

Following the definition of creativity proposed by Amabile (1983), this dissertation highlights two criteria of the creative product: novelty and usefulness. Correspondingly, at the level of individual scientist, creativity refers to an individuals' tendency to produce novel and useful research outputs. Psychologists have proposed diverse definitions of creativity in terms of creative process, creative person, and creative product, but there is an emerging consensus that the product definitions are the most useful for creativity research and that novelty and appropriateness/value are the two most important criteria for creative products (Amabile, 1983; Ford, 1996; Woodman et al., 1993). In addition, Amabile (1982) pointed out that a major obstacle to creativity studies pertains to translating the conceptual definition of creativity into an operational one in order to allow empirical assessment of creativity. Creativity is not an intrinsic quality that can be measured by some ultimately objective criteria, so the assessment of creativity is ultimately subjective. Ford (1996) further defined creativity as "a domain-specific, subjective judgment of the novelty and value of an outcome of a particular action."

This perspective is also implicitly shared by various measures or proxies for creativity used in empirical studies: Nobel laureates as an indicator of eminent scientist (Zuckerman, 1967), prestigious prizes to identify path-breaking discovers in biomedical research (Hollingsworth, 2004), surveying experts to nominate highly creative accomplishments (Heinze et al., 2009), financial success and critics' reviews for the Broadway musicals (Guimera, Uzzi, Spiro, & Amaral, 2005), citation counts for patents (Fleming, 2001; Fleming et al., 2007; Singh & Fleming, 2010), journal impact factor for collaboration teams (Guimera et al., 2005), and publications and citations to measure creativity of scientists (Simonton, 1999, 2004). In spite of their difference in many important aspects, these studies share one thing in common, that is, creativity is subjectively assessed by experts, consumers, users, or peers.

These two ideas, (1) the emphasis of creativity on novelty and usefulness and (2) creativity is a collective judgment made by a target group, are particularly relevant to the world of science. First, the importance of novelty/originality is self-evident. In the world of science, "the emphasis on the value of originality has a self-evident rationale, for it is originality that does much to advance science" (Merton, 1957). Furthermore, the scientific reward system is constructed as such that "recognition and esteem accrue to those who have … made genuinely original contributions to the common stock of knowledge" (Merton, 1957). Second, the criterion of usefulness and appropriateness is also important for gaining scientific rewards. The "institutional commitment to novelty in the modern sciences is counterbalanced by their other major distinctive feature – the collective appropriation of task outcomes to produce new knowledge" (Whitley, 2000). Therefore, research is valued only if it can influence, direct, and is essential for the work of colleagues in the field. Third, the merit of a scientist or a scientific output is judged by colleagues, as the modern sciences are "reputational work organizations" (Whitley, 2000), and peer recognition serves as the foundation for the institution of science (Merton, 1957).

From Merton's perspective, citation serves as an elementary building block of the scientific reward system, and therefore can be viewed as a good proxy for scientific creativity. For a paper, the acceptance for publishing indicates the acknowledgement of its original contributions to science from peers in the field. Being cited further indicates the peer-recognition of its value and its impact on the scientific community (De Bellis, 2009; Merton, 1973). In other words, citations indicate the impact/recognition of creativity and therefore can be used as an indirect measure of creativity. Furthermore, since creativity is not an inherent quality that can be measure objectively, and the evaluation of creativity ultimately relies on the collective judgment from others (i.e., users, reviewers, and peers), peer recognition embodied in citations can be used as a proxy for creativity. Empirically, Garfield (1973) found that the majority of Nobel laureates were amongst the top 0.1% most cited authors. Cole and Cole (1967) studied 120 physicists and found that the number of citations was more significant than the number of publications in eliciting recognition through the receipt of awards, appointment to prestigious academic departments, and being widely known in the scientific community. Furthermore, Newman and Cooper (1993) found that papers, which explored new paradigms or carried a paradigm into more unknown territory, received more citations than others that refined or extended existing theories. Therefore, at the operationalization level, this paper uses the number of citation as a proxy for creativity. This approach has many problems (Martin & Irvine, 1983). These problems and the treatments undertaken to address them will be further discussed in the method section.

2.2. Research Collaboration

De Solla Price (1986) has shown a noticeable increase of scientific collaboration since the beginning of the 20th century. This phenomenon has drawn a lot of attention from the academia (Barnett, Ault, & Kaserman, 1988; Cronin, 2005; de Solla Price & Beaver, 1966; Katz & Martin, 1997; McDowell & Melvin, 1983; Wuchty et al., 2007).

Underlying this remarkable increase in research collaboration is the change in the environment of scientific research. First, collaboration is driven by the intellectual need to accomplish a project in an environment in which research is increasingly specialized and interdisciplinary (Beaver & Rosen, 1978; Katz & Martin, 1997; Laband & Tollison, 2000). A similar argument is the "burden of knowledge" thesis (Jones, 2009): Individuals face an increasing educational burden because of knowledge accumulated by previous generations. Because of the accumulation of previous knowledge and limited individual capacity, individuals are forced to concentrate in narrower fields and rely on teamwork for scientific practice. Second, De Solla Price (1986) emphasized the importance of economic factors driving collaboration: One important motivation to collaborate is to squeeze “full papers out of people who only have fractional papers in them at that particular time.” In addition, *big science* requires expensive facilities, large personnel, and massive funding, which in turn drives collaboration (Hwang, 2008; Katz & Martin, 1997; Luukkonen, Persson, & Sivertsen, 1992).

In addition to the macro-level intellectual and economic factors, a specific collaboration can be driven by a variety of micro-level motivations. Beaver and Rosen (1978) highlighted eighteen motives: access to special equipment and facilities, access to special skills, access to unique materials, access to visibility, efficiency in use of time, efficiency of use of labor, to gain experience, to train researchers, to sponsor a protégé, to increase productivity, to multiply proficiencies, to avoid competition, to surmount intellectual isolation, need for additional confirmation of evaluation of a problem, need for stimulation of cross-fertilization, spatial propinquity, and accident or serendipity. The authors provided a conceptual analysis but no empirical data to assess these motives. A survey done by Melin (2000) revealed that the major reason for collaboration was that the coauthor had special competence, followed by the reason that the coauthor had special data or equipment, and then social reasons such as previous friendship, collaborative, or mentor-student relationships.

Melin (2000) also observed a goal-oriented attitude towards collaboration, that is, people collaborate in order to gain something, such as methods, equipment or special competence, and otherwise they don’t collaborate. Correspondingly, he found that the primary benefits from collaboration were increased-knowledge and higher scientific quality, that is, each collaborator contributes with his/her special knowledge and brings in different perspectives for the invested problem, so that a scientist gains knowledge and quality from the collaboration. This benefit in terms of knowledge exchange and cross-fertilization of ideas has been intensively discussed (Beaver & Rosen, 1978; Nilles, 1975; Pierce, 1999). Other benefits of collaboration include: intellectual companionship, networks expansion, and more future collaboration opportunities (Katz & Martin, 1997). Empirically, collaboration has been confirmed to be beneficial to winning the Nobel Prize (Zuckerman, 1967), productivity in terms of number of publications (de Solla Price & Beaver, 1966; Landry, Traore, & Godin, 1996; Lee & Bozeman, 2005), and research impact as measured by received citations (Guimera et al., 2005; Katz & Hicks, 1997; Smart & Bayer, 1986). On the other hand, collaboration also has costs such as monetary costs for travelling and communication, time, increased administration, and institutional and culture barriers (Cummings & Kiesler, 2007; Walsh & Maloney, 2002).

In the literature, coauthorships are frequently used as proxies for collaboration. Similarly, this dissertation studies collaborations which lead to coauthored papers but leaves out collaborations which are not embodied in coauthorships. Limitations of this approach are further discussed in the method section.

2.3. Creative Process

Collaboration plays a critical role in scientific creativity. By pooling together different expertise and perspectives, collaboration contributes to cross-fertilization of ideas and enables combining different pieces of knowledge to make something novel and useful (de Solla Price, 1986; Katz & Martin, 1997; Melin, 2000). Furthermore, collaboration allows members to build off of others' ideas to create new knowledge which is not originally possessed by any collaborator individually, so that the amount of knowledge in a collaborative team is greater than a simple summation of knowledge possessed by each individual. However, collaborative teams are often found to perform below their potential. For example, brainstorming in interactive groups generates fewer ideas than does brainstorming by individuals working alone (Levine & Moreland, 2004;

Skilton & Dooley, 2010). This problem is called “process loss” and has many causes, such as opportunistic behavior, failure to share information, and lack of coordination (Hackman & Morris, 1975; Levine & Moreland, 2004).

Therefore, in order to facilitate creativity in collaborations, it is important to understand the creative process in collaboration. This dissertation brings in insights from literature on small groups for a better understanding of the social processes underlying collaboration. Adapted from previous literature on creative process at the group level (Hackman & Morris, 1975; Levine & Moreland, 2004; Skilton & Dooley, 2010), this dissertation highlights three steps of the creative process: idea generation, idea convergence, and idea implementation. Scientific creativity requires divergent thinking to generate a set of novel ideas, convergent thinking to select the best alternative, and coordinated action to implement the selected ideas. This creative process framework derived from studies of small group is also applicable to describe creative process at the dyadic level between two collaborators.

Idea Generation. Scholars have long considered divergent thinking to be an important cognitive skill for creativity at the individual level (Guilford, 1950; McCrae, 1987). The importance of divergent thinking also applies at the group level (Levine & Moreland; Skilton & Dooley, 2010; Woodman et al., 1993). At the beginning of a collaborative project, the problems and goals are ambiguously defined, and the solutions are unclear, so the creative process starts with generation of a variety of ideas about the problem and potential solutions (Drazin, Glynn, & Kazanjian, 1999; Ford, 1996; Woodman et al., 1993). The search for problem definition and solution can go in various directions, and a broader search increases the possibility of discovering better (i.e., more novel and useful) solutions. In this process, cognitive diversity is very important: collaborators are expected to contribute their diverse prior experiences, perspectives, methodologies, and expertise, actively interact with the domain of the study and collaborators, in order to generate diverse ideas.

Idea Convergence. After a variety of ideas have been generated, collaborators need to reconcile their differences and form a consensus on which idea is the best and should be implemented. This process involves not only a cognitive convergent-thinking process but also non-cognitive aspects, such as negotiations and compromises to resolve interest conflicts and align commitments (Latour & Woolgar, 1986; Skilton & Dooley, 2010). Furthermore, the idea convergence process is heavily affected by the personal relationships between collaborators. First, some knowledge similarity and overlap is needed to enable collaborators to understand each other, evaluate different ideas contributed by different individuals, integrate different knowledge sources for idea refinement and improvement, and eventually converge on the best idea (Cohen & Levinthal, 1990; Nahapiet & Ghoshal, 1998; Star & Griesemer, 1989). Second, mutual trust, affection, and obligation embodied in strong personal relations can help reconcile conflicts, smooth negotiation, and align commitments (Krackhardt, 1992; Lin & Ensel, 1989; Uzzi, 1996).

Idea Implementation. After idea convergence, collaborators have to implement the chosen idea, carry out the project, and translate it into successful publications. Generating ideas and implementing ideas are two distinct processes, and they may respond differently to certain individual-, group- and social- level factors, that is, a situation optimal for idea generation might not be so desirable for idea implementation (Mumford & Gustafson, 1988; Obstfeld, 2005; West, 2002). One important aspect of idea implementation is selling the research outcome to colleagues in the field by convincing them of the validity of the results and of their value for future research, since the merit of the scientific work is judged by peer recognitions. Publishing is an important component of the scientific communication system, and communicating/writing affects peer's appreciations (Beyer, Chanove, & Fox, 1995; Laband & Piette, 1994). On the other hand, one of the benefits from collaboration is higher degree of technical competence and the opportunity for cross-checking and pre-submission “internal refereeing” (Gordon, 1980; Presser, 1980).

This paper depicts these three steps in a linear fashion with one step following another. However, the collaboration project may consist of several iterations of the generation-convergence-implementation processes. For example, new problems emerge in the implementation process and require collaborators to generate new ideas to address those new problems. In addition, some complete process may be nested in a bigger process, for example, a team developing a new technology may have several technical performance goals, and choose to tackle these sub-problems either sequentially or in parallel. Furthermore, there might be feedbacks between different stages, for example, when idea convergence fails, collaborators have to go back to idea generation to search for new alternatives. However, iterations, nested-structures, or feedbacks in the creative process do not cause problems in this dissertation, which explores network effects at the holistic-process-level instead of single steps. In other words, this dissertation studies the final success of the whole creative process, which requires success at each stage.

2.4. A Network Approach to Creativity

Before discussing effects of egocentric collaboration networks on creativity, it is important to compare this network approach with the team approach, since the prevailing norm in creativity studies is to investigate team structures and dynamics. In this dissertation, a team/group refers to a group of researchers working specifically on one project or a series of projects, while an egocentric collaboration network of a focal scientist consists of all his collaborators, while these collaborators may be grouped in multiple overlapping teams.

Group process and its impact on group performance have been extensively studied on corporate R&D teams. Some of them especially related to group creativity have been reviewed in the last section. These studies should shed light on the understanding of scientific creativity, especially because the science production has shifted from an individual- to a team- based model (Wuchty et al., 2007). However, previous wisdom might not be directly transplantable to the field of scientific collaboration, which is a very different environment compared with corporate teams. As Whitley (2000) pointed out, one important characteristic distinguishes the modern sciences from other systems of work organization is its autonomy and self-governance. This special setting of modern sciences may require different lens for understanding the organization of teams in scientific collaboration.

First, corporate teams have clear organizational boundaries and mandates, while teams in scientific collaboration are fluid, making the "team" a less useful social construct to study the organization of science. Throughout a scientist's career life, different teams emerge and dissolve, and even at a specific stage of a research project, the boundary of a collaborative team is ambiguous. Because of the autonomy in team organization, people constantly come and go. For example, teams may acquire new members when new expertise is needed, and teams may break apart as the common interests between teammates disappear. To some extent, a collaborative team is co-evolving with the work. This fluidness also exists in corporate teams, and there are studies of fluid and project-based teams (Huckman & Staats, 2011). However, these fluid corporate teams are still designed by managers and embedded in a larger organization, and there is still a team as an unambiguous entity to be evaluated. On the contrary, in the world of scientific collaboration, there is often no unambiguous entity of a team to be credited for the scientific work, otherwise, there wouldn't be longstanding discussions on authorship rules or difficulties confronting scientists when deciding whom should be credited as authors and whom should only be rewarded with positions in the acknowledgements. The difficulty of identifying an unambiguous working team or mapping the working team to the group of authors makes the "team" a less useful social construct for studying research collaboration. For example, *A* and *B* are authors of a paper, and this paper is motivated by their previous collaboration with *C* and benefitted tremendously from comments provided by *D*. In addition, *E* participated in some early stages of the research but was not involved in the writing of the paper, so *E* was not listed as an author. In this case, should we restrict our research within the team of *A* and *B* while ignoring *C*, *D*, and *E*?[1]

Second, corporate teams are relatively independent entities isolated from the external environment, and the bulk of the work-related transactions take place within the team. On the contrary, scientists are often involved in multiple collaborative teams, and these teams are interdependent. Even in studies of relatively independent corporate teams, some scholars adopted an "external" approach to study team behaviors directed outward, toward other parts of the organization, as well as the effects of external activity on group performance (Ancona, 1990; Ancona & Caldwell, 1988, 1992). In these studies, the focus of external activities is about managing external dependence and obtaining critical resources (Pfeffer & Salancik, 1978), for example, obtain technical information, map resources, support, and trends in organizations, to influence those individuals with key resources, and to synchronize work flow (Ancona & Caldwell, 1988). Others further acknowledged external learning as an important component of organizational learning in addition to internal learning (Bresman, 2010; Edmondson, 2002; Wong, 2004), for example, groups learn from others with similar experiences about key aspects of its task or process (i.e., *vicarious learning*) and learn from external sources about key aspects of its context (i.e., *contextual learning*) (Bresman, 2010). However, in this literature, external and internal learning are still distinct, for example, Wong (2004)

[1] The approach took in this dissertation will address this issue partially, but not completely. I will study the whole egocentric coauthorship network, C, D, E not listed on the paper between A and B are likely to show up in other papers authored by A. However, it is also possible that their collaboration relationship will not show up in the coauthorship data. The issue of using coauthorship as proxy for collaborations will be further discussed in the method section.

argued that internal learning focuses on exploiting existing knowledge and enhances efficiency, while external learning focuses on exploring new knowledge and promotes innovativeness. However, the boundary between external and internal learning in scientific collaboration might not be so clear, because scientists participate in multiple collaborative teams simultaneously and there are strong knowledge spillovers across these teams. For example, a scientist may take a novel idea from one team and implement it in another team, use methods developed in one team to help problem-solving in another, and start new lines of research with a new team based on knowledge learned from an old one. In addition, the blurred boundary between external and internal learning also reflects the ambiguity of team boundaries. For example, *A* may publish two closely related papers, one with *B* and *C* and the other with *B* and *D*. It's very unrealistic to assume independence and no knowledge spillovers between these two projects. Furthermore, for the team consisting of *A*, *B* and *C*, is the internal learning between these three people fundamentally different from the external learning between *A*, *B* and *D*? To investigate scientific creativity, should we study these two papers separately and bound our study within two groups respectively, or take the network approach to study all these collaborators at the same time?

Empirical evidence of the fluidness of teams is provided in **Table 1**. Among all identified triplets (i.e., teams with three members), only 9% of them collaborated repeatedly, in other words, about 91% of these triplet-teams only happened once. Furthermore, 63% of these triplets have collaborated in some slightly different teams (for example, only two of them collaborated, or two of them collaborated with someone else).

Table 1. Fluidness of Teams

1	2	3	4	5	6	7	8
Group type	Total number of groups	Number of groups which have more than one paper (repeated groups)	Number of groups which have paper(s) with someone outside of the group	Number of groups which have paper(s) by the ego, a subset of the group, and maybe also someone outside of the group	Number of repeated groups which have paper(s) with someone outside of the group	Number of repeated groups which have paper(s) by the ego, a subset of the group, and maybe also someone outside of the group	Maximum number of papers of the group
doublet	1169	228 (19.50%)	517 (44.23%)		121 (53.07%)		16
triplet	1460	133 (9.11%)	256 (17.53%)	917 (62.81%)	37 (27.82%)	97 (72.93%)	6
quartet	1404	57 (4.06%)	139 (9.90%)	991 (70.58%)	13 (22.81%)	39 (68.42%)	4
quintet	891	31 (3.48%)	56 (6.29%)	684 (76.77%)	5 (16.13%)	23 (74.19%)	3
sextet	601	13 (2.16%)	30 (4.99%)	462 (76.87%)	3 (23.08%)	12 (92.31%)	3

Numbers in the brackets are percentages

Data and sample information is reported in the method section.

- Non-Physics egos, 2005-2007 papers
- Identify doublets from 2-authored papers, triplets from 3-authored papers, …

Column explanations

2. How many doublets/triplets/… are there?
3. How many doublets/triplets/… have more than one paper?
4. Out of all groups, how many groups have paper(s) authored by the whole group and also someone outside of the group?
5. Out of all groups, how many groups have paper(s) authored by the ego, a subset of the group (at least one member is not there), and possibly also someone else? This is actually a conservative estimation, because I require the ego to be there, while it is possible that other members of the group have papers without the ego.
6. Out of repeated groups, ….?
7. Out of repeated groups, …?
8. Among all doublets/triplets/…, there is one group has the largest number of papers, what is this number?

Because of the fluidness of teams and the interdependence between them, the organization of scientific collaboration may be described by the *Garbage Can Model* (Cohen, March, & Olsen, 1972) illustratively but not rigorously. There are streams of problems, expertise, and collaborators in the network. Problems are searching for relevant expertise, expertise searching for problems, and collaborators searching for common research interests (i.e., problems) and complementary expertise. A collaborative team emerges when these three streams converge. However, the emergence of a team is not the end of a chapter. Instead, the team still interacts with these three streams and co-evolves with them. Therefore a network approach is needed to understand scientific creativity, searching for sources of creativity in dynamics networks beyond the boundaries of closed teams. Different networks may have different problem, expertise, and collaborator streams, and therefore lead to different final creative outcomes. In addition, this network approach does not deny the relevance of team-level dynamics: The structure of the temporary team still affects the creative process and final outcomes. The main intension of this network approach is to take a broader perspective to understand scientific creativity as a result of collaboration networks and to take into account the interdependence between teams. In other words, the main argument is that an egocentric network has impacts on creativity cross all the collaborative teams of this ego.

At the core of this network argument is the premise of knowledge spillover across collaborative teams. Some studies have highlighted that team members bring in lessons learned from previous group experience to new group situations (Ancona, 1990; Nonaka, 1994; Reagans, Argote, & Brooks, 2005). Creativity models at the group level also acknowledge the importance of individuals' previous experiences and other resources that they bring with them (Amabile, 1983; Ford, 1996; Woodman et al., 1993). In his discussion on the ontological dimension of knowledge, Nonaka (1994) not only acknowledged the importance of social interaction for knowledge creation, but also argued that "[a]t a fundamental level, knowledge is created by individuals. An organization cannot create knowledge without individuals." (p. 17). One common theme of these different streams of literature is that individuals are the fundamental creator and container of knowledge. Therefore, a scientist may serve as the media to transfer knowledge from one collaborative team to another.

Hypothesis 1: there are significant knowledge spillovers across one scientist's different collaborative projects.

2.5. Tie Strength and Creativity

After justifying the network approach, the next question is: What kind of network is more creative, specifically how is the network-level tie strength related to creativity? Latour and Woolgar (1986) have revealed that the "thought process" (emergence of ideas) in science is not only about scientific logic but also subject to sociological determinants. The research question of this dissertation is therefore about how tie strength affects the creative process. I will firstly discuss the relationship between tie strength and creativity at the dyadic level and then translate the discussion to the egocentric network level, so let's think about the problem at the dyadic level right now.

Granovetter (1973) defined the strength of tie as "a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie." Weak ties are more likely to provide non-redundant information (Burt, 1992; Granovetter, 1973; Uzzi, 1996). People bonded by strong ties are more likely to be similar to each other and connected with similar others. Therefore, information obtained from such networks tends to be redundant. In contrast, weak ties are more likely to bridge *structural holes* between communities that are otherwise unconnected and provide access to information and resources beyond those available in one's own social circles.

Because of the access to non-redundant information, weak-tie-collaborations are more likely to generate novel ideas. Many scholars have suggested that the source of creativity is making unusual but fruitful recombination of ideas (Mednick, 1962; Nelson & Winter, 1982; Schumpeter, 1939). For example, Nelson and Winter (1982) suggested that "the creation of any sort of novelty in art, science, or practical life – consists to a substantial extent of a recombination of conceptual and physical materials that were previously in existence." Many stories of science also concur on this recombination theory[2]. Furthermore, this recombination process is not only intellectual but also social, for example, one scientist discovered the link between the selenium content of water and the effectiveness of his assay, and the story behind this discovery was that one of his students was mandated to take a course in an unrelated field and happened to learn something about selenium (Latour & Woolgar, 1986). Therefore, exposure to diverse knowledge and perspectives could increase the chance of making remote associations between different ideas.

However, the association between weak ties and creativity is not so straightforward because many other factors may affect the creative process. For a successful idea convergence, generated ideas firstly need to be shared. The weak tie argument assumes that people connected by a weak tie are willing to share information (Gabbay, 1997), which may be true in the case of sharing non-sensitive and easy-to-share information. However, intense communication and trust are needed for sharing complex and sensitive knowledge, and willingness or obligations are needed for collaborators to take the costs and share (Hansen, 1999; Reagans & McEvily, 2003; Uzzi, 1997).

[2] Please refer to Latour, B., & Woolgar, S. 1986. ***Laboratory life : the construction of scientific facts***. Princeton, NJ: Princeton University Press. and Simonton, D. K. 2004. ***Creativity in science : chance, logic, genius, and Zeitgeist***. Cambridge, UK ; New York: Cambridge University Press. for many interesting stories and quotes.

Therefore, the lack of mutual trust, obligation, and norm in the collaborative tie may impede information exchange between collaborators (Krackhardt, 1992; Lin & Ensel, 1989; Podolny & Baron, 1997; Uzzi, 1996). Furthermore, collaborators bonded by weak ties may find significant communication and epistemological problems because of the lack of a common knowledge base to integrate different ideas and perspectives (Ahuja, 2000; Cohen & Levinthal, 1990; Nahapiet & Ghoshal, 1998; Star & Griesemer, 1989), and these problems hamper the idea convergence and implementation process.

As the strength of tie enhances, so does the *cognitive capital* (i.e., shared knowledge and understanding) and *relational capital* (i.e., trust, norm, and obligation), and as a result, the collaboration has a more effective creative process. Empirically, many studies have shown the advantage of strong ties for knowledge transfer (Hansen, 1999; Reagans & McEvily, 2003) and knowledge creation (McFadyen & Cannella, 2004; McFadyen et al., 2009; Tortoriello & Krackhardt, 2010; Walsh & Maloney, 2002).

However, this effect may turn negative when the strength is too strong. First, cognitions of the collaborators become very similar, and therefore the ability to generate novel ideas is diminished (McFadyen & Cannella, 2004; Uzzi, 1997). Second, shared collaboration experience gives birth to shared cognitive structures/routines that governs behavior of the collaborators (Granovetter, 1985). Skilton and Dooley (2010) argued that an enduring and shared mental model would emerge from repeated collaboration, and the metal model would shape not only the way that individuals explain, predict, and describe events, but also the way that the team differentiates roles among members. Furthermore, the mental model is inert and constrains subsequent collaboration in terms of how they approach problems and what role they play in the division of labor. As a result, repeated collaboration is less able to generate novel ideas. Furthermore, constrained by the mental model, collaborators may have higher self-censorship of ideas, that is, they are more likely to disclose and share ideas related to prior projects instead of novel ideas not shared by members already (Wittenbaum, 2003). This argument is also consistent with the exploration-exploitation literature that old-timers are more likely to exploit exiting knowledge than to explore new knowledge (Gupta, Smith, & Shalley, 2006; March, 1991; Perretti & Negro, 2006). Empirically, studies have found a negative association between repeated collaboration and creativity in science (Guimera et al., 2005; Porac et al., 2004).

In summary, at the dyadic level, the message is that there is an inverted U-shaped relationship between tie strength and creativity, that is, the effect of tie strength is initially positive and turns negative after a threshold. How to translate this message to the network-level analysis? If we can assume that ties in the same egocentric network are relatively homogeneous, then we can use the network average tie strength to indicate the overall tie strength of the whole network, and then the tie strength effect at the network level is a simple aggregation of effects at the dyadic level. Therefore,

> *Hypothesis 2: there is a quadratic (inverted U-shaped) relationship between network average tie strength and creativity, that is, the effect of network average tie strength is initially positive and turns negative after a threshold.*

This simple aggregation approach is implicitly adopted by many studies in the literature (Abbasi, Altmann, & Hossain, 2011; Gabbay, 1997; Hansen, 1999; McFadyen & Cannella, 2004; McFadyen et al., 2009; Reagans & McEvily, 2003), but it is based on a questionable assumption that network ties are homogeneous. This assumption might be reasonable in many of the above referenced studies, for example, those using survey data to study corporate R&D networks, which are bounded in organizations and have more or less homogeneous ties. In the next section, I will question this assumption in the context of scientific collaboration.

2.6. Tie Strength Skewness and Creativity

Now I will explore the effect of network tie configuration and the discussion is at the egocentric network level. As discussed in the last section, many studies implicitly adopted the network tie homogeneity assumption and used network average tie strength as an indicator for the overall tie strength of the whole network, without acknowledging heterogeneity among network ties (Abbasi et al., 2011; Gabbay, 1997; Hansen, 1999; McFadyen & Cannella, 2004; McFadyen et al., 2009; Reagans & McEvily, 2003). However, Uzzi (1996) found that his interviewees maintained both embedded and arms-length ties, suggesting that the configuration of ties, rather than a simple dichotomy between strong-tie-network and weak-tie-network, should be investigated. Uzzi (1996) used a Herfindahl-type indicator to measure the dominance of strong ties in a network[3]. In addition, some studies define a boundary between strong and weak ties, count them separately, and then investigate their effects separately (Tortoriello & Krackhardt, 2010; Walsh & Maloney, 2002)[4].

This dissertation is interested in the skewness of tie strength distribution. In the world of scientific collaboration, it is normal for a scientist to simultaneously have a small group of colleagues with very intense interaction on the one hand and a number of loose contacts on the other hand. The tie strength distribution of an egocentric network tends to be skewed and very different from a normal distribution, so that using the network average tie strength may hide distinct network configuration characteristics. For

[3] This approach is not adopted here because I am interested in a different aspect of tie configuration, skewness instead of dominance of strong ties.

[4] This approach is, to some extent, shared by some analyses of this dissertation, i.e., classify coauthors into new and repeated ones. In addition, this approach misses a lot of information about the tie strength distribution, so it is not used for studying my main interest concerning tie configuration.

example, from two networks with the same average tie strength, one may have all ties of medium strength, while the other has half strong and half weak ties.

In addition, there is actually anther implicit assumption underlying the simple aggregation approach (i.e., network-level effect is an aggregation of dyadic-level effects), that is, there are no interaction effects between dyads. However, because of knowledge spillover across collaborations (Hypothesis 1), the diversity in dyads may have some positive effects on creativity. Empirically, egocentric collaboration networks have (positively) skewed tie strength distributions, with a long tail on the right side and the bulk of the values lie to the left of the mean. The limited number of strong ties may reflect the limit of carrying-capacity. Scientists have limited amount of time and energy to devote to research, while maintaining strong relations is costly. Therefore, having too many strong collaborative ties is simple infeasible or inefficient (McFadyen & Cannella, 2004; Perry-Smith & Shalley, 2003). On the other hand, a large number of weak ties may reflect scientists' broad search for diverse and complementary knowledge. A large number of weak ties may augment the scientist's knowledge base about the research domain (Perry-Smith & Shalley, 2003; Simonton, 1999) and also enhance his *absorptive capacity* (Reagans and McEvily, 2003).

Specifically, when the network average tie strength is high, a less skewed network will suffer from the lack of weak ties and therefore have low creativity. In contrast, a more skewed network still has a number of weak ties and a "healthy" mixture of strong and weak ties, that is, the benefit of knowledge diversity gained from weak-tie-collaborations can be transferred to strong-tie-collaborations (Hypothesis 1), and therefore the whole egocentric network is still very creative. When the network average tie strength is low, however, the effect of tie strength skewness is unclear. Therefore, I hypothesize that

Hypothesis 3: A more skewed network has higher creativity compared with a less skewed network, when the network average tie strength is high.

Furthermore, ties strength skewness moderates the effect of network average tie strength. Given the heterogeneity between ties in a skew network, the average tie strength is not so accurate to indicate the overall tie strength of the whole network. For example, when the average tie strength is very high, there are still a number of weak ties in a skew network. Therefore, a skew network is less sensitive to the change in network average tie strength. For example, when the network average tie strength is low, a more skewed network already has some very strong ties. Under this circumstance, if we increase the strength of each tie, the network does benefit from the increase in those very weak ties, but not from the increase in those already very strong ones. Therefore, the aggregated positive effect is smaller in a more skewed network than a less skewed one. Applying the same logic, the negative effect caused by further increase in network average tie strength after a threshold is also smaller in a more skewed network than in a less skewed one.

Hypothesis 4: tie strength skewness moderates the effect of network average tie strength, specifically, both the initial positive effect and the later negative effect caused by increase in network average tie strength are smaller in a more skewed network than in a less skewed one.

2.7. Creativity Decline vs. Cost Reduction

My discussion in this section will return to the dyadic level. In section 2.5, I argued that, at the dyadic level, creativity is low when the tie strength is too strong, and I explained that it is because of path-dependency and low cognitive diversity, which is from a psychologist's perspective. However, from an economist's perspective, the transaction cost theory (Williamson, 1981) may provide a different explanation: strong-tie-collaborators have lower observed creativity because they are collaborating on different types of projects, which are "profitable" in strong-tie-collaborations with low transaction costs but not in weak-tie-collaboration with high costs (Catalini, 2012).

Collaborations, particularly long-distance and cross-institutional ones, face very high communication and coordination costs, which may hamper the performance of the collaboration (Cummings & Kiesler, 2007). As discussed in Section 2.5, weak-tie-collaborators face high transaction costs. Without a prior collaborating history or an intersection of their social lives, weak-tie-collaborators have very low *cognitive*- and *relational*- *capital* at their disposal. As a result, they have to bear high transaction costs in coordinating their collective actions and invest in building mutual understandings and trusts. In contrast, strong-tie-collaborations have low transaction costs.

Furthermore, this difference in transaction costs will lead to different choices of collaborative projects. Assume the payoff of a collaborative project to a scientist is $\pi = f(V) - c$, where V is the creativity of the project, $f(V)$ is an increasing function in V, and c is the transaction cost, then a scientist will choose to do the project only if $f(V) \geq c$. As c decreases, projects with lower V become "profitable." Therefore, weak-tie-collaborations will be more selective, while strong-tie-collaborations will include more trivial projects and correspondingly have lower observed average creativity.

On the other hand, lower transaction costs also makes more experimental projects profitable (Catalini, 2012). Research can be modeled as a costly process of trials and errors, in which a creative idea has to pass through many gates to reach the final successful creative product (Aghion, Dewatripont, & Stein, 2008; Manso, 2011). Adapted from Aghion and colleagues (2011), I propose the following simple model of research. Assume that an idea has to go through *k* stages to become a final successful product, and at each stage, this idea has a fixed cost *c* to proceed and a fixed probability of p to pass (that is, $1 - p$ probability to fail, and $p \in (0{,}1)$). This idea has value V to the scientist if it successfully passes all k stages and value *0* if fails. In addition, the scientist faces two choices at each stage: pursue the project or give up. Therefore, at the final stage (i.e., the k^{th} stage), the scientist's expected payoff of continuing the project is $E(\pi_k) = pV - c$, and he will choose to continue if the expected payoff is positive. At the *k-1*th stage, the scientist's expected payoff of pursuing the idea is $E(\pi_{k-1}) = p \cdot (pV - c) - c$. Following this iteration process, we can derive that the expected payoff of pursuing the project at the first stage is $E(\pi_1) = p^k \cdot V - c \sum_{i=1}^{k} p^{i-1} = p^k \cdot V - c \cdot \frac{1-p^k}{1-p}$. Therefore, projects will be pursued in the first place only if $c < c^* = \frac{V \cdot (1-p) p^k}{1-p^k}$. Given that V is finite and $p \in (0{,}1)$, as *k* approaches to ∞, the threshold c^* approaches to 0. Therefore, an experimental project involves many stages will only be pursued if *c* is very low, in other words, these types of projects are only profitable to be pursued in strong-tie-collaborations with very low *c*. In this case, we should observe that strong-tie-collaborations are more likely to produce extremely creative products. Therefore, based on the cost-reduction argument,

> *Hypothesis 5a: strong-tie-collaborations have lower average but higher maximum observed-creativity, compared with weak-tie-collaborations.*

On the other hand, based on the creativity-decline argument, we should observe that strong-tie-collaborations have both lower average and maximum observed-creativity.

> *Hypothesis 5b: strong-tie-collaborations have lower average and maximum observed-creativity, compared with weak-tie-collaborations.*

2.8. Collaboration and Knowledge Diffusion

The collaboration networks discussed in previous sections are venues where creative products are produced, and the research theme is about network effects on knowledge creation. From a different perspective, this section studies collaboration networks as channels, through which creative products are diffused, and networks discussed here are not necessarily (and typically not) the network that produced the creative products. In summary, this section studies the network effect (at the egocentric network level) on knowledge diffusion.

As discussed in Section 2.1, citations are social rewards to the originality and appropriateness of the scientific work (Merton, 1957; Whitley, 2000). Furthermore, according to Merton's universalistic view of science, the merit of the scientific work is purely evaluated on the work itself, while who produced this work is irrelevant. In other words, the scientific reward system should be open and fair to all individuals regardless of their social status. In reality, however, who produced the work matters. The preceding reputation of the author provides a certification for the validity of his/her work (Latour & Woolgar, 1986). As a result, the scientific reward system is subject to a *Matthew effect*: "The overloading of the scientific communication system leads scientists to choose their reading matter on the basis of an author's preceding reputation, often further enhancing that reputation" (Merton, 1968).

Furthermore, in these discussions on the institution of science, authors are viewed as disinterested reporters and scientific papers as objective documentations of the research. However, others propose that authors of scientific papers have very clear goals of persuading peers to share his opinion of the value of his work, and correspondingly, the scientific papers are strategic communications serving this purpose. Referencing is one rhetorical device in the toolbox of persuasion (Cozzens, 1989; Gilbert, 1977; Latour & Woolgar, 1986). This perspective provides an additional explanation for the Matthew effect. Citing scientists with higher prestige would provide stronger persuasion power.

Therefore, an author would cite a paper from a renowned scientist instead of another similar paper from an unknown scientist, even though the author himself is equally convinced by these two papers.

After acknowledging scientists' strategic behavior in referencing, it is not difficult to understand a variety of "misbehaviors" in citation practices, for example, scientists may cite their colleagues for repaying extra-scientific debts or simply for courtesy, use citations to bribe potential reviewers in exchange for favorable review results, or use referencing to transfer the responsibility for errors or omissions to the referenced sources (De Bellis, 2009).

Therefore, citations depend on not only the intrinsic quality and scientific appropriateness of the paper, but also many other social factors. This dissertation is interested in the marketing effect of collaboration networks: can collaboration networks contribute to higher visibility and therefore more citations of a scientist? However, testing this collaborator-marketing effect is very tricky, because of the existence of other confounding factors, such as the prestige effect and the intellectual-relevance effect.

Collaborator Marketing. Higher visibility in the scientific community leads to higher probabilities to be cited. Aizenman and Kletzer (2011) found that premature death costs economists a large number of citations and argued that death in the midst of an active career eliminates the opportunity of the scientist to raise awareness of his work and also reduces the incentives of others to cite him for strategic reasons. Therefore, a larger collaboration network may contribute to higher awareness of a scientist's work. A scientist's collaborators are likely to know his previous works and may do marketing for his works. For example, they may refer his works to their graduate students, friends, or others in advising activities, conference discussions, and reviewing processes.

Prestige Effect. Preceding reputation enhances subsequent recognitions, either because papers of prestigious authors are more "reliable," more powerful for persuading, or because they are cited for other strategic reasons. Empirically, Crane (1965) found that productive biologists, political scientists, and psychologists at higher-status universities received higher recognitions than equally productive scientists at lower-status universities, and Helmreich, Spence, Beane, Lucker, and Matthews (1980) found that prestige of psychologists' departments were positively correlated to their subsequent citations.

Intellectual Relevance. From the perspective of research evaluation, the *halo effect* (Martin & Irvine, 1983), but not the *Matthew effect*, challenges the validity of citations as a useful measure for research evaluation. In other words, the phenomenon that prestigious authors get more citations is not necessarily biased, and it is biased only when two comparable papers are treated differently because of the difference in the authors' statuses. Performance of scientists is unequal and skewed (Fox, 1983; Stephan & Levin, 1991), and the observed enormous differences between scientists could be normally caused by some slight differences in their talents, education, and factors other than discrimination (Ghiglino, 2012; Simonton, 2004). Therefore, the Matthew effect itself does not challenge the validity of citation counts, while how to translate evaluation results into policies, such as funding allocations, is a different question (Hicks & Katz, 2011). Therefore, the observed Matthew effect need to be decomposed into two parts: the prestige effect because of discrimination, and the intellectual-relevance effect purely because of the scientific appropriateness. For example, some renowned scientists are much more highly cited than others, and it is simply because their work successfully re-directed other colleagues' research and generated new fruitful lines of research.

These three effects (i.e., collaborator marketing, prestige effect, and intellectual relevance) are distinct effects but strongly related to each other. Motivated by the question about network effect on knowledge diffusion, I hypothesize that

Hypothesis 6: A paper gets more citations if the author has more collaborators, after controlling for prestige effect and intellectual relevance.

CHAPTER 3

METHOD

3.1. Data

Survey and bibliometric data for 1,323 American scientists in five disciplines are used in this dissertation. The sample of scientists are from the NETWISE I project (NETWISE, 2007), which is funded by the United States National Science Foundation (NSF). The NETWISE I survey was conducted in 2007 on 3,677 stratified randomly sampled American scientists in six disciplines: biology (BIOL), chemistry (CHEM), computer science (CS), earth and atmospheric sciences (EAS), electrical engineering (EE), and physics (PHYS). The random sample was stratified by sex, rank, and discipline, from the population of academic scientists and engineers in these six disciplines in Carnegie-designated Research I universities (150 universities). The population was constructed by manually retrieving information from the websites of the relevant departments or university directories, and copying the faculty information for assistant, associate, and full professors. Sample weights were calculated using the inverse of the probability of selection and will be employed in the ego-level analyses. Of the 1,774 completed surveys, 176 were removed because of ineligible rank or discipline, resulting in a final total sample size of 1,598. The overall response rate of the survey, calculated using the RR2 method of the American Association for Public Opinion Research (AAPOR) is 45.8%, and the weighted response rate is 43.0%. The responses' distribution of gender, field, and rank are very similar to the survey population.

The NETWISE I survey asked sampled scientists to name up to ten closest collaborators over the past two academic years (five within and five outside of the university), and it is specified that "[c]ollaboration includes proposal generation, working on a research project, writing/presenting an academic paper/book or book chapter, or developing industrial products or patents." Rich information concerning these ties was then collected, such as duration of the relation, frequency of interaction, resource exchange, shared knowledge understandings. The surveyed scientists are referred to as *egos* and their named close collaborators as *alters*. Out of 1,598 egos in six disciplines, 1,435 egos named 7,292 alters, and only 74 (5%) listed the maximum number of 10 alters.

Life-time publication records for these egos were subsequently retrieved from Thomson Reuters Web of Science (WOS) using its online interface. Publication collection firstly required an author name and affiliation match, and then cleaned out false papers of homonymous authors following a name disambiguation algorithm with very high accuracy rate, which is documented in Wang et al. (2012). Coauthor names were standardized, cleaned, and matched to alter names. The bibliometric data were last updated in May 2011, so the publication data in 2011 is incomplete. Because of the complex publishing practice in the field of physics, publication data for physicists were left out, leaving 1,323 scientists in the remaining five disciplines. Only journal articles of these egos were used in this dissertation, while other document types (e.g., reviews, letters, notes, and editorials) were excluded because they might not reflect original research. Out of 1,323 egos, 1,310 egos published 43,996 journal articles in total.

Citation histories of these 43,996 publications were collected from the KB internal database, a bibliometric database developed and maintained by the Competence Center for Bibliometrics for the German Science System (KB). This KB database is built upon the raw data provided by WOS and is updated annually. The version used for this dissertation is updated in April 2012, so I have complete citation counts by years from 1980 to 2011, including both all-citation-counts and non-self-citation-counts (i.e., excluding citations awarded by papers with common author(s)).

3.2. Citation

Citation counts are used as proxies for creativity. A five-year time window is used to count citations, that is, for a paper published in 2007, its citations in year 2007-2011 are counted. Therefore, papers published in different years will have the same time window to accumulate citations. In one case I also study papers published after 2007, and a three-year time window citation counts and journal impact factors are supplemented. For analyses at the paper level, citation counts are used, and for analyses at the level of ego or ego-year (i.e., the unit of analysis is a set of papers), several citations statistics (sum, average, maximum, minimum, and median) are often used. The ratio of cited/uncited papers and the ratio of highly cited papers are also considered but not implemented, because each ego has very limited number of papers, so the denominators of the ratio-type indicators are too small to provide reliable measures.

Several treatments were undertaken to address problems of using citation counts. Martin and Irvine (1983)'s discussion on citation is probably still the best statement at the moment about using citations. One problem is that citation ageing pattern differs across papers: many highly-cited papers take a long time to establish themselves as elite papers, while many papers have very early citation peaks (Glänzel, Schlemmer, & Thijs, 2003; Van Raan, 2004). Therefore, a sufficient citation time window is needed to give reliable citation counts. According to Wang (2013)'s calculation on the whole WOS database, the spearman correlations between five-year time window citation counts and 31-year citation counts are: 0.81, 0.91, 0.85, 0.89, and 0.79 in fields of biology, biomedical research, chemistry, earth and space, and engineering, respectively. The correlations are sufficiently high for this study. Another problem pertaining to citation ageing is that some classical papers become so internalized in the scientific practice and accepted as the norm that people constantly mention them but do not need to provide references to them (Latour & Woolgar, 1986; Merton, 1983). This issue does not cause problems in this study because only recent publications are studied.

Another issue is about self-citations. Some productive scientists may actively cite themselves while self-citations do not reflect the recognition from the community (Aksnes, 2003; Glänzel, Debackere, Thijs, & Schubert, 2006). However, there is still no consensus in the bibliometric community regarding this issue. In addition, identifying the actual individuals from database-indexed names is a challenging task, and the process of excluding self-citations may introduce too many errors. Therefore, self-citations are not excluded in this dissertation. In Section 4.2, non-self-citation counts are also tried and lead to the same conclusion.

In addition, citations are incomparable between fields because of the field differences in size and referencing norms (Moed, Burger, Frankfort, & van Raan, 1985). Furthermore, the most important problem might be the *halo effect* (i.e., a paper of prestigious authors or institutions tends to be evaluated more highly and get more citations than another papers identical in every aspect except for being written by an unknown author). Several procedures are implemented to address these two issues: (1) control for field and ego demographic variables correlated to prestige, (2) include a lagged dependent variable, and (3) add ego or paper level fixed/random effects. The latter two procedures not only control for field or ego prestige effects, but also address the endogeneity problem. These three procedures are used together or separately depending on the data and model.

3.3. Coauthorship

Coauthorship data are used to construct collaboration networks. Based on the belief that collaboration usually results in a published paper and is therefore reflected in coauthorships (Beaver & Rosen, b; Gordon, 1980), coauthorship data have been widely used to study research collaboration. Furthermore, there are also several practical advantages of coauthorship data: invariant and verifiable, inexpensive, large amount of data are available, and un-intrusive and non-reactive (Katz & Martin, 1997).

However, the validity of coauthorship as a proxy for collaboration relies on two assumptions: all coauthors actually participated in the collaboration, and all collaborations result in coauthored publications (Glänzel & Schubert, 2005; Laudel, 2002). The first assumption is challenged by the phenomenon of honorary authorship (Katz & Martin, 1997), which is a very serious problem particularly in the biomedicine field (Biagioli, 1999). The second assumption is also attacked by many researchers (Edge, 1979; Katz & Martin, 1997; Van Raan, 1998), for example, collaborators may choose to publish collaborative work separately, and valuable suggestions and comments are not reflected in coauthorships.

There are also several empirical studies about the relationship between collaboration and coauthorship. In a small scale study in a university, Melin and Persson (1996) found that only 5% of the authors have experienced situations in which collaboration did not result in coauthored papers. Moreover, the reason for excluding some authors was usually because they made only minor contributions. However, based upon 101 semi-structured interviews with research group leaders and at least one group member, Laudel (2002) identified six types of collaborations with different patterns of rewards. She found that about half of the collaborations were not visible in coauthorships and that about one third of collaborations were rewarded only by acknowledgements.

Another challenge to coauthorship data is that they do not indicate what kind of contributions are made by each author, and therefore cannot reflect the complex human interaction process underlying collaboration (Bordons & Gomez, 2000).

Nevertheless, coauthorship data have been widely used, and proven powerful to reveal the relationship between network and performance (Abbasi et al., 2011; McFadyen et al., 2009), and the social structure of sciences (Crane, 1969; Guimera et al., 2005; Moody, 2004). Most analyses in this dissertation are based on coauthorship-based collaboration networks, but this dissertation also includes a replicate study based on survey data to validate some findings.

3.4. Mixed Effect Models

Many analyses in this dissertation use hierarchical or panel data, and therefore mixed-effect models are frequently used. This section provides a brief introduction to mixed-effect models, and model specifications in the latter sections will only present equations but not all specification details, such as variances and covariances of random effects and random errors, unless they are specified differently from here.

3.4.1. Linear Mixed Effect Models

Take an analysis using alter-level data as an example. The alter-level data are grouped by egos, so the alter-level equation for alter *j* of ego *i* is

$$y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + \cdots + b_{iq} z_{qij} + \varepsilon_{ij}$$

$$b_{ik} \sim N(0, \psi_k^2), Cov(b_k, b_{k'}) = \psi_{kk'}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2), Cov\left(\varepsilon_{ij}, \varepsilon_{ij'}\right) = 0 \; for \; j \neq j'$$

where

- y_{ij} is the value of the dependent variable for the *j*-th alter of the *i*-th ego.
- β's are the fixed-effect coefficients, which are identical for all egos.
- x's are the fixed-effect independent variables for the *j*-th alter of the *i*-th ego.
- b_i's are the random-effect coefficients for ego *i* and assumed to be multivariately normally distributed. The random effects vary by ego. The b_i's are thought of as random variables, not as parameters, and are similar to the random errors ε_{ij} in this respect.
- z's are the random-effect independent variables for the *j*-th alter of the *i*-th ego.
- ψ_k^2 are the variances and $\psi_{kk'}$ the covariances of the random effects, and they are assumed to be constant across egos.

- ε_{ij} are the random errors for the j-th alter of the *i*-th ego. The errors for ego *i* are assumed to be multivariately normally distributed.
- Observations are assumed to be sampled independently within groups and are assumed to have constant error variance, that is, $Cov\left(\varepsilon_{ij}, \varepsilon_{ij'}\right) = 0\ for\ j \neq j'$.

3.4.2. Generalized Linear Mixed Effect Models

In many cases, the dependent variables are count or dummy variables, which do not follow a normal distribution as assumed by linear models, so generalized linear mixed-effect models (e.g., quasi-Poisson, negative binomial, and logistic) should be used. From linear models to generalized linear models, we make two modifications. First, specify a *link function* (i.e. *ln* for Poison and negative binomial models, and *logit* for logistic models) between the predicted mean μ and the linear predictor η:

$$l(\mu) = \eta = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + \cdots + b_{iq} z_{qij}$$

Second, specify an error distribution function, such as Poisson and negative binomial distributions.

3.4.3. Linear Quantile Mixed Effect Models

Most studies in social sciences are concerned with averages, take a paper-level analysis for example, the estimation is about the effect of collaboration type (new and repeated) on the mean of citation counts. However, the effect might be different at different percentiles of the citation distribution, for example, the average citations are not different between new and repeated collaborations, but new collaborations have a lower value of citation counts at the 10th percentile and a higher value at the 90th percentile of the citation distribution. These distribution differences are of interest in some cases, and quantile regression is a powerful tool for modeling distributions.

Assume that the dependent variable y_{ij} is continuous[5]. I will use linear quantile models with random intercepts, so the conditional quantile function (CQF) is

$$Q_{y_{ij}}(\tau|X, b_i) = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_i$$

$$y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_i + \varepsilon_{ij}$$

$$b_i \sim N(0, \psi^2)$$

$$\varepsilon_{ij} \sim AL(0, \sigma^2, \tau), Cov\left(\varepsilon_{ij}, \varepsilon_{ij'}\right) = 0 \; for \; j \neq j'$$

where

- $Q_{y_{ij}}$ is the quantile of y_{ij} with τ, that is, $Pr\left(y_{ij} \leq Q_{y_{ij}}(\tau)\right) = \tau$. τ is a fixed priori, and $0 < \tau < 1$, (e.g., $\tau = .5$ for the median).
- AL(·) is the Asymmetric Laplace distribution.

R packages, nlme (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2013), MASS (Venables & Ripley, 2002), glmmADMB (Skaug, Fournier, Nielsen, Magnusson, & Bolker, 2013), and lqmm (Geraci, 2013) are used to fit linear, quasi-Poisson, negative binomial, and linear quantitle mixed-effect models, respectively.

[5] This assumption actually does not hold for the number of citations, which is a count variable. For count variables, there are special estimation methods developed for quantile fixed-effect models Machado, J. A. F., & Silva, J. M. C. S. 2005. Quantiles for Counts. ***Journal of the American Statistical Association***, 100(472): 1226-1237., but not for mixed-effect modes. In addition, analysis in this dissertation uses the natural logarithm of the citation counts as the dependent variable, making it more “continuous.” Therefore, I will overlook the violation of this assumption in model estimations.

CHAPTER 4

RESULTS

4.1. The Knowledge Spillover Hypothesis

Hypothesis 1: there are significant knowledge spillovers across one scientist's different collaborative projects.

This section tests the hypothesis on knowledge spillover across collaboration teams. To make the text more concise, at the given time point of a focal collaboration, I define *new-collaborators* as collaborators who have not collaborated before and *repeated-collaborators* as collaborators who have collaborated before, correspondingly, I define *new-collaborations* as collaborations with only new collaborators and *repeated-collaborations* as collaborations with only repeated collaborators.

The knowledge spillover hypothesis is very general and is operationalized as follows: (1a) the number of new-collaborations has a positive effect on creativity of repeated-collaborations, and (1b) the number of repeated-collaborations has a positive effect on creativity of new-collaborations. As discussed in Section 2.5, new-collaborations are more likely to explore new knowledge while repeated-collaborations are more likely to exploit existing knowledge. However, the knowledge spillover hypothesis (1a) suggests that the benefits of knowledge diversity gained from new-collaborations can be transferred to repeated-collaborations. In other words, an ego's repeated-collaborations are more creative if he has more new-collaborations. One the other hand, (1b) suggests that a productive relationship with repeated-collaborators might provide a secured position for an ego to better explore new-collaborations.

4.1.1. Model Specification

To test these two sub-hypotheses, unbalanced panel data of ego's citations by year are used, and the unit of analysis is ego-year. In each year, an ego's coauthors are classified into two types: *new* (not coauthored in the last three years) and *repeated* (coauthored at least once in the last three years). Subsequently, an ego's papers are classified into four types: *solo* (single-authored paper), *new* (coauthored with only new coauthors), *rep* (coauthored with only repeated coauthors), and *mix* (coauthored with both new and repeated coauthors).

To address the endogeneity issue, that is, both the number of new- (or repeated-) collaborations and the citations of repeated- (or new-) collaborations are likely to be correlated to the previously performance of the ego, a lagged citation measure, $CiteALL_{i,t-1}$, citations of all papers in the last year is included in the model.

Citations of the *i*-th ego's repeated- and new-collaboration papers published in year *t* are specified as follows respectively

$$\begin{aligned} Cite_{i,t}^{REP} = {} & \beta_0^{REP} + \beta_1^{REP} \cdot t + \beta_2^{REP} \cdot PubREP_{i,t} + \beta_3^{REP} \cdot PubNEW_{i,t} + \beta_4^{REP} \\ & \cdot CiteALL_{i,t-1} + \beta_5^{REP} \cdot Field_i + \beta_6^{REP} \cdot Rank_i + \beta_7^{REP} \cdot Gender_i + \beta_8^{REP} \\ & \cdot Race_i + b_{1i}^{REP} + b_{2i}^{REP} \cdot t + \varepsilon_{i,t}^{REP} \end{aligned}$$

$$\begin{aligned} Cite_{i,t}^{NEW} = {} & \beta_0^{NEW} + \beta_1^{NEW} \cdot t + \beta_2^{NEW} \cdot PubREP_{i,t} + \beta_3^{NEW} \cdot PubNEW_{i,t} + \beta_4^{NEW} \\ & \cdot CiteALL_{i,t-1} + \beta_5^{NEW} \cdot Field_i + \beta_6^{NEW} \cdot Rank_i + \beta_7^{NEW} \cdot Gender_i \\ & + \beta_8^{NEW} \cdot Race_i + b_{1i}^{NEW} + b_{2i}^{NEW} \cdot t + \varepsilon_{i,t}^{NEW} \end{aligned}$$

where

- $Cite_{i,t}^{REP}$is the value of the dependent variable for the *i*-th ego's repeated-collaboration papers published in year *t*, i.e., average citation counts (*Cite.AVG*) and maximum citation counts (*Cite.MAX*). $Cite_{i,t}^{NEW}$has the same meaning, but for new-collaboration papers.

- β's are the fixed-effect coefficients and identical for all egos.
- b_i's are the random-effect coefficients for ego *i* and vary by ego. Specifically, b_{1i} is the random intercept for ego i, and b_{2i} is the random trend for ego *i*. They are added to account for ego heterogeneities.
- ε's are random errors.
- Two equations use the same set of independent variables.
- t is the year.
- $PubREP_{i,t}$ is the number of repeated-collaboration papers of ego *i* in year *t*. It is natural logarithm transformed for model estimation.
- $PubNEW_{i,t}$ is the number of new-collaboration papers of ego *i* in year *t*. *ln* transformed.
- $CiteALL_{i,t-1}$ is the average citations of all papers published by ego *i* in year *t-1* if the dependent variable is *Cite.AVG*. It is the maximum citations of all papers published by ego *i* in year *t-1* if the dependent variable is *Cite.MAX*. *ln* transformed.
- $Field_i$ is the research field of ego *i*, $Rank_i$ is the academic rank of ego *i*, $Gender_i$ is the gender of of ego i, and $Race_i$ is the race of ego i. They do not change over time, a set of dummies are included in model fitting.

Incorporated variable are listed and descripted in **Table 2**.

Table 2. Knowledge Spillover: Variable Descriptions

Variables	Descriptions
Dependent Variables	
ln(Cite.AVG.rep+1)	The average number of citations received by paper of repeated collaboration, *ln* transformed.
ln(Cite.MAX.rep+1)	The maximum number of citations received by paper of repeated collaboration, *ln* transformed.
ln(Cite.AVG.new+1)	The average number of citations received by paper of new collaboration, *ln* transformed.
ln(Cite.MAX.new+1)	The maximum number of citations received by paper of new collaboration, *ln* transformed.
Independent Variables	
ln(Pub.rep+1)	Number of repeated collaboration papers, *ln* transformed.
ln(Pub.new+1)	Number of new collaboration papers, *ln* transformed.
ln(Cite.AVG.lag+1)	The average number of citations received by all paper published in the last year, *ln* transformed.
ln(Cite.MAX.lag+1)	The maximum number of citations received by all paper published in the last year, *ln* transformed.
Field	Categorical variable: BIOL, CHEM, EAS, EE, and CS.
Rank	Categorical variable: Assistant Professor, Associate Professor, and Full Professor.
Gender	1 if female 0 if male.
Race	1 if non-Hispanic White and 0 if minority.

The unit of analysis is ego-year.

Table 3. Knowledge Spillover: Descriptive Statistics

		n	mean	sd	median	min	max	1	2	3	4	5	6	7
1	ln(Cite.AVG.rep+1)	3337	2.15	0.95	2.20	0	6.22							
2	ln(Cite.MAX.rep+1)	3337	2.40	1.02	2.48	0	6.22	.91						
3	ln(Cite.AVG.new+1)	4958	2.05	1.00	2.08	0	6.80	.27	.25					
4	ln(Cite.MAX.new+1)	4958	2.30	1.08	2.30	0	6.80	.26	.25	.92				
5	ln(Pub.rep+1)	6588	0.48	0.54	0.69	0	2.64	.00	.32	.06	.08			
6	ln(Pub.new+1)	6588	0.70	0.49	0.69	0	2.56	.06	.04	.03	.32	-.39		
7	ln(Cite.MAX.lag+1)	6588	1.68	1.17	1.90	0	5.81	.34	.35	.24	.24	.24	-.07	
8	ln(Cite.AVG.lag+1)	6588	2.15	1.47	2.48	0	6.80	.30	.36	.23	.24	.34	-.09	.93
9	Field	BIOL 1600, CHEM 1890, CS 690, EAS 1587, EE 821												
10	Rank	Assistant Prof 115, Associate Prof. 1025, Full Prof. 5448												
11	Gender	Female 2649, Male 3939												
12	Race	Non-Hispanic White 5609, Minority 979												

The unit of analysis is ego-year.

4.1.2. Sample Restriction

Unbalanced panel data of ego's citations of different types of papers are used for analysis, and there are some restrictions on the sample. First, only observations (i.e., ego-year, a set of papers authored by ego *i* in year *t*) between 1983 and 2007 are included. Excluding observations before 1983 is because the complete WoS data start in 1980, and I classify coauthors into new- or repeated- coauthors based on coauthorship history in the preceding three years. Excluding observations after 2007 is because these papers do not have a complete period of five years to accumulate citations, while five-year time window citation counts are used as dependent variables. Second, if the ego's first paper is published in year *t*, then this ego's observations before (and not including) year *t+3* are excluded for the same reason of classifying coauthors into new and repeated ones. Third, egos with at least five observations are included.

Because egos may have repeated- but no new- collaboration papers in some years, and vice versa, so the sample sizes for estimating knowledge spillovers on repeated- and new- collaborations are slightly different. I have 3,039 observations of 320 ego for testing Hypothesis 1a and 4,877 observations of 518 egos for 1b.

4.1.3. Model Estimation

After natural logarithm transformation, dependent variables are roughly normally distributed (

Figure 2). Linear mixed-effect models (using the *ln* of citation statistics as dependent variables) are fitted, firstly only with focal explanatory variables, and then also including other control variables. Because the number of citations is a count variable, and there is apparent over-dispersion in the data, quasi-Poisson models (using integer values of citations statistics as dependent variables) are estimated. Adding control variables almost has no impact on the coefficients of focal explanatory variables.

4.1.4. Findings

Descriptive statistics are reported in **Table 3**. Results for testing Hypothesis 1a are reported in **Table 4**, and the coefficients on ln(Pub.new+1) test the knowledge spillover effect of new-collaborations on repeated-collaborations. The coefficients are all positive and significant when the average citations are used as dependent variable (column 1-4), and positive but not significant when the maximum citations are used as the dependent variable (column 5-8). Therefore, Hypothesis 1a is supported, that is, the number of new-collaborations does have significant knowledge spillovers to repeated-collaborations, which raises the average creativity, although not the maximum creativity, of repeated-collaborations.

Results for testing Hypothesis 1b are reported in **Table 5**, and the coefficients on ln(Pub.rep+1) test the hypothesized knowledge spillover effect. The coefficients are all positive but not significant. Therefore, Hypothesis 1b is not supported, that is, there is no evidence that the number of repeated-collaborations have significant knowledge spillovers to new-collaborations. In addition, ln(Pub.rep+1), the only explanatory variable of interest, happens to be the only insignificant variable in these models.

There is another interesting finding from comparing **Table 4** and **Table 5**. Coefficients on ego rank, gender, and race are all insignificant in **Table 4** but are all significant in **Table 5**. It seems that rank, gender, and race differences in citations only appear in new-collaborations, but not in repeated-collaborations.

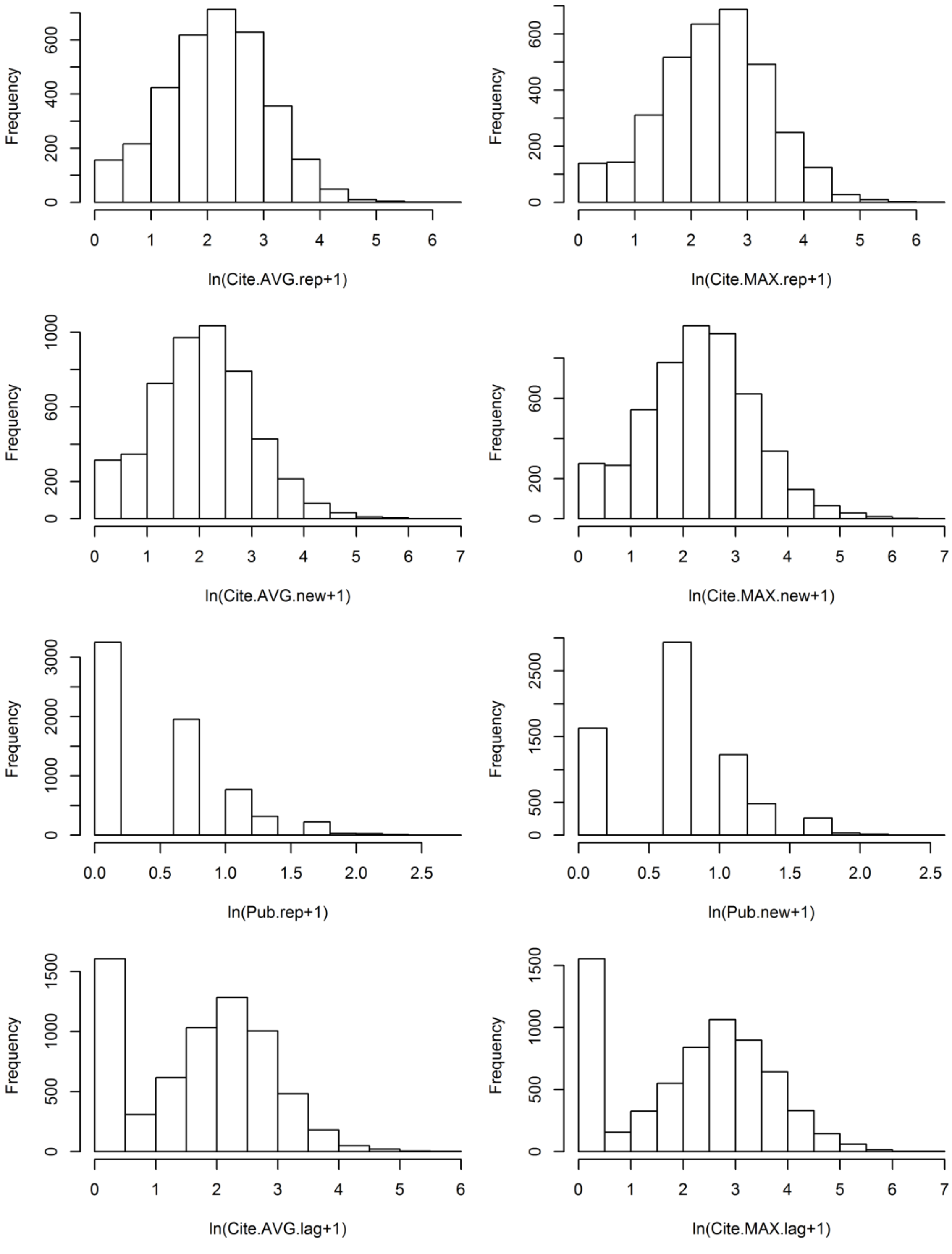


Figure 2. Knowledge Spillover: Histograms

Table 4. Knowledge Spillover: Repeated Collaboration Models

	1		2		3		4	
	ln(Cite.AVG+1)		ln(Cite.AVG+1)		Cite.AVG		Cite.AVG	
	Linear Mixed		Linear Mixed		Quasi-Poisson		Quasi-Poisson	
	Model		Model		Mixed Model		Mixed Model	
ln(Pub.rep+1)	-0.04	(0.05)	-0.02	(0.05)	-0.11 **	(0.05)	-0.10 *	(0.06)
ln(Pub.new+1)	0.05 *	(0.03)	0.05 *	(0.03)	0.07 **	(0.04)	0.07 *	(0.04)
ln(Cite.lag+1)	0.16 ***	(0.02)	0.14 ***	(0.02)	0.17 ***	(0.02)	0.16 ***	(0.02)
Field-CHEM			-0.09	(0.07)			-0.10	(0.09)
Field-CS			-0.80 ***	(0.12)			-0.94 ***	(0.17)
Field-EAS			-0.05	(0.08)			-0.02	(0.12)
Field-EE			-0.62 ***	(0.09)			-0.70 ***	(0.13)
Gender-Female			-0.02	(0.05)			-0.03	(0.07)
Rank-Associate			0.11	(0.26)			0.16	(0.31)
Rank-Full			0.11	(0.25)			0.14	(0.30)
Race-White			-0.07	(0.07)			-0.09	(0.09)
Log-likelihood	-3850		-3820		-4835		-4806	

	5		6		7		8	
	ln(Cite.MAX+1)		ln(Cite.MAX+1)		Cite.MAX		Cite.MAX	
	Linear Mixed		Linear Mixed		Quasi-Poisson		Quasi-Poisson	
	Model		Model		Mixed Model		Mixed Model	
ln(Pub.rep+1)	0.85 ***	(0.05)	0.87 ***	(0.05)	0.86 ***	(0.05)	0.87 ***	(0.05)
ln(Pub.new+1)	0.04	(0.03)	0.04	(0.03)	0.05	(0.03)	0.04	(0.03)
ln(Cite.lag+1)	0.13 ***	(0.01)	0.13 ***	(0.01)	0.14 ***	(0.01)	0.14 ***	(0.01)
Field-CHEM			-0.09	(0.07)			-0.10	(0.08)
Field-CS			-0.75 ***	(0.12)			-0.77 ***	(0.15)
Field-EAS			-0.05	(0.08)			-0.02	(0.11)
Field-EE			-0.61 ***	(0.09)			-0.61 ***	(0.12)
Gender-Female			-0.03	(0.06)			-0.03	(0.06)
Rank-Associate			0.07	(0.26)			0.12	(0.28)
Rank-Full			0.05	(0.25)			0.08	(0.27)
Race-White			-0.08	(0.08)			-0.09	(0.08)
Log-likelihood	-3918		-3892		-5045		-5028	

Number of observations (ego-year): 3039
Number of groups (egos): 320
* p < .10, ** p < .05, *** p < .01
Dependent variables are citation statistics (i.e., *Cite.AVG* and *Cite.MAX*) for repeated collaborations. Linear mixed models (column 1,2,5,6) use *ln* transformed citation statistics as the dependent variable (i.e., *ln(Y+1)*), and quasi-Poisson mixed models (column 3,4,7,8) use the integer value of the citation statistics. For the independent variable *ln(Cite.lag+1)*, *ln(Cite.AVG.lag+1)* is used if the dependent variable is about *Cite.AVG*, and *ln(Cite.MAX.lag+1)* is used if the dependent variable is about *Cite.MAX*. All models include a random intercept and time trend, which are not reported. The reference group for field is BIOL, for Rank is Assistant Professor, for Gender is Male, and for Race is Minority.

Table 5. Knowledge Spillover: New Collaboration Models

	1	2	3	4
	ln(Cite.AVG+1)	ln(Cite.AVG+1)	Cite.AVG	Cite.AVG
	Linear Mixed Model	Linear Mixed Model	Quasi-Poisson Mixed Model	Quasi-Poisson Mixed Model
ln(Pub.rep+1)	0.01 (0.03)	0.01 (0.03)	0.01 (0.04)	0.02 (0.04)
ln(Pub.new+1)	0.01 (0.04)	0.01 (0.04)	-0.05 (0.05)	-0.04 (0.05)
ln(Cite.lag+1)	0.04 *** (0.01)	0.04 *** (0.01)	0.04 *** (0.01)	0.04 *** (0.01)
Field-CHEM		-0.25 *** (0.06)		-0.28 *** (0.08)
Field-CS		-0.99 *** (0.07)		-1.26 *** (0.11)
Field-EAS		-0.29 *** (0.06)		-0.34 *** (0.09)
Field-EE		-0.72 *** (0.08)		-0.86 *** (0.12)
Gender-Female		0.07 * (0.04)		0.11 * (0.06)
Rank-Associate		-0.29 * (0.15)		-0.38 ** (0.18)
Rank-Full		-0.26 * (0.14)		-0.34 * (0.18)
Race-White		-0.15 ** (0.06)		-0.21 *** (0.08)
Log-likelihood	-6501	-6418	-7943	-7868

	5	6	7	8
	ln(Cite.MAX+1)	ln(Cite.MAX+1)	Cite.MAX	Cite.MAX
	Linear Mixed Model	Linear Mixed Model	Quasi-Poisson Mixed Model	Quasi-Poisson Mixed Model
ln(Pub.rep+1)	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)
ln(Pub.new+1)	1.00 *** (0.04)	1.01 *** (0.04)	1.04 *** (0.05)	1.05 *** (0.05)
ln(Cite.lag+1)	0.04 *** (0.01)	0.04 *** (0.01)	0.03 *** (0.01)	0.03 *** (0.01)
Field-CHEM		-0.26 *** (0.06)		-0.27 *** (0.07)
Field-CS		-0.94 *** (0.08)		-1.05 *** (0.10)
Field-EAS		-0.29 *** (0.06)		-0.32 *** (0.08)
Field-EE		-0.70 *** (0.08)		-0.75 *** (0.11)
Gender-Female		0.07 (0.05)		0.09 * (0.05)
Rank-Associate		-0.30 ** (0.15)		-0.36 ** (0.17)
Rank-Full		-0.28 * (0.15)		-0.34 ** (0.16)
Race-White		-0.17 *** (0.06)		-0.20 *** (0.07)
Log-likelihood	-6656	-6582	-8295	-8228

Number of observations: 4877

Number of groups (egos): 518

* p < .10, ** p < .05, *** p < .01

Dependent variables are citation statistics (i.e., *Cite.AVG* and *Cite.MAX*) for new collaborations. Linear mixed models (column 1,2,5,6) use *ln* transformed citation statistics as the dependent variable (i.e., *ln(Y+1)*), and quasi-Poisson mixed models (column 3,4,7,8) use the integer value of the citation statistics. For the independent variable *ln(Cite.lag+1)*, *ln(Cite.AVG.lag+1)* is used if the dependent variable is about *Cite.AVG*, and *ln(Cite.MAX.lag+1)* is used if the dependent variable is about *Cite.MAX*. All models include a random intercept and time trend, which are not reported. The reference group for field is BIOL, for Rank is Assistant Professor, for Gender is Male, and for Race is Minority.

4.2. The Tie Strength Hypotheses

Hypothesis 2: there is a quadratic (inverted U-shaped) relationship between network average tie strength and creativity, that is, the effect of network average tie strength is initially positive and turns negative after a threshold.

Hypothesis 3: A more skewed network has higher creativity compared with a less skewed network, when the network average tie strength is high.

Hypothesis 4: tie strength skewness moderates the effect of network average tie strength, specifically, both the initial positive effect and the later negative effect caused by increase in network average tie strength are smaller in a more skewed network than in a less skewed one.

4.2.1. Model Specification

Cross-sectional data of 576 egos are used for testing the above three hypotheses about effects of network average tie strength and tie strength skewness. Dependent variables are citation statistics (*Cite.SUM*, *Cite.AVG*, and *Cite.MAX*) of an ego's papers published between 2005 and 2007, and tie strength between an ego and his coauthor is measured as the number of their coauthored papers in the same period.

The equation for ego *i* is

$$\begin{aligned} Cite_i = \beta_0 + \beta_1 \cdot TieStrAVG_i + \beta_2 \cdot TieStrAVG_i^2 + \beta_3 \cdot Skew_i + \beta_4 \cdot Skew_i \\ \cdot TieStrAVG_i + \beta_5 \cdot Skew_i \cdot TieStrAVG_i^2 + \beta_6 \cdot NetSize_i + \beta_7 \\ \cdot NetSize_i^2 + \beta_8 \cdot CiteLag_i + \beta_9 \cdot Pub_i + \beta_{10} \cdot CareerAge_i + \beta_{11} \\ \cdot CarrerAge_i^2 + \beta_{12} \cdot Field_i + \beta_{13} \cdot Gender_i + \beta_{14} \cdot Race_i + \varepsilon_i \end{aligned}$$

where

- $Cite_i$ is the value of the dependent variable for the i-th ego, total citations (*Cite.SUM*), average citations (*Cite.AVG*), and maximum citations (*Cite.MAX*).
- $TieStrAVG_i$ is the network average tie strength for ego i. Both $TieStrAVG_i$ and $TieStrAVG_i^2$ are included to test the curvilinear effect of network average tie strength.
- $Skew_i$ is the tie strength skewness for ego i. Interaction terms between $Skew_i$ and $TieStrAVG_i$ & $TieStrAVG_i^2$ are included to test the moderating effect.
- $NetSize_i$ is the network size for ego i, that is, the number of ego i's coauthors.
- $CiteLag_i$ is the lagged dependent variable for ego i. ln transformed.
- Pub_i is the number of papers for ego *i*. *ln* transformed.
- $CareerAge_i$ is the number of years after receiving PhD for ego *i*,$Field_i$ is the research field of ego *i*, $Gender_i$ is the gender of of ego *i*, and $Race_i$ is the race of of ego *i*.
- All variables with squared terms are centered at their mean, to reduce the potential multicollinearity problem.

Further details about these variables are presented in **Table 6**.

Dependent Variables. Citation count of a scientist's papers published between 2005 and 2007 are used as a proxy for creativity. Three different citations counts were used to capture not only the average creativity performance but also the extremely creative instances. These three counts are: the total number of citations of all papers (*Cite.SUM*), the average citation rate (i.e., citations per paper)(*Cite.AVG*), and the maximum citation number among all papers (*Cite.MAX*).

Table 6. Tie Strength: Variable Descriptions

Variables	Descriptions
Dependent Variables	
Cite.SUM	The total number of citation received by an ego's papers published between 2005 and 2007.
Cite.AVG	The average citation rate of an ego's papers published between 2005 and 2007.
Cite.MAX	The maximum citation count of an ego's papers published between 2005 and 2007.
Independent Variables	
Tie Strength AVG	The average number of times that an ego coauthored with all his/her coauthored between 2005 and 2007.
Skewness	The skewness of an ego's tie strength distribution.
Cite.SUM.lag	The total number of citation received by an ego's papers published between 2002 and 2004.
Cite.AVG.lag	The average citation rate of an ego's papers published between 2002 and 2004.
Cite.MAX.lag	The maximum citation count of an ego's papers published between 2002 and 204.
Pubs	Number of papers published by an ego between 2005 and 2007.
Career Age	2007 – Year received the PhD degree.
Gender	1 if female and 0 if male.
Race	1 if non-Hispanic White and 0 if minority.
Field	Categorical variable: BIOL, CHEM, EAS, EE, and CS
Network Size	Number of coauthors an ego has between 2005 and 2007.

The unit of analysis is ego.

Independent Variables. Publication data between 2005 and 2007 are used to construct collaboration networks and independent variables. Many studies of scientists' or inventors' collaboration networks used a three-year time window to construct collaboration networks (Long, 1978; McFadyen, 2004; McFadyen, 2009; Fleming, 2001; Fleming, 2007). Tie Strength at the dyadic level is defined as the number of coauthored papers, and *Tie Strength AVG* at the egocentric network level is the network average tie strength. *Skewness* of the tie strength distribution is calculated using the following formula:

$$Skewness = \frac{\frac{1}{n} \cdot \sum_1^n (x_i - \bar{x})^3}{\left(\frac{1}{n-1} \cdot \sum_1^n (x_i - \bar{x})^2\right)^{3/2}}$$

where *n* is the number of ties in an egocentric network, and the *x*'s are the tie strength measures at the dyadic level. In addition, two other popular skewness formulas[6] are also tried, and all three skewness measures are highly correlated and yield very similar regression results.

Control Variables. Lagged dependent variables, that is, citation counts for papers published between 2002 and 2004, are included to control for unobserved and omitted variables that might be of potential importance to creativity or citation counts, such as author prestige and ability of collaborators. Number of papers published between 2005 and 2007 was also controlled, given that more papers may result in higher total or maximum citation counts. *Network Size* was also controlled, which was measured as the number of coauthors between 2005 and 2007. McFadyen and Cannella (2004) observed

[6] $g_1 = \frac{\frac{1}{n} \cdot \sum_1^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \cdot \sum_1^n (x_i - \bar{x})^2\right)^{3/2}}$ and $G_1 = g_1 \cdot \frac{\sqrt{n(n-1)}}{n-2}$

an inverted U-shaped relationship between network size and journal-impact-factor-weighted number of publication in biomedical sciences, because an increase in network size on the one hand increases cognitive diversity but on the other hand may distract scientists from other more productive activities. Therefore, both *Network Size* and *Squared Network Size* were included in the model.

Age, experience, and rank are important factors of research collaboration and performance (Lee & Bozeman, 2005; Levin & Stephan, 1991; van Rijnsoever & Hessels, 2011; van Rijnsoever, Hessels, & Vandeberg, 2008). The survey data have information about age, career age (number of years after receiving the PhD degree), and rank (full, associate, or assistant professor), but these three variables are highly correlated, so only *Career Age* is included to control for all these three effects, following the same procedure of Lee and Bozeman (2005). A brief summary of all variables are provided in **Table 6**.

4.2.2. Sample Restriction

The initial sample has 1,323 scientists in five disciplines. Scientists with missing survey data are excluded, leaving 1,318 scientists. Several restrictions are further imposed on the sample. First, an active publishing history is needed to construct collaboration networks and count citations, so 928 scientists are included who have at least one publication in each of the following four periods: before 2002, between 2002 and 2004, between 2005 and 2007, and after 2007. Requiring at least one publication before 2002 and after 2007 excludes beginners who have just started to build collaboration networks and drop-outs or retirees. Publications between 2005 and 2007 are needed to construct collaboration networks and count citations. Similarly, publications between 2002 and 2004 are needed for constructing the lagged dependent variables.

Furthermore, papers with a large number of authors may cause problems in this study. Some papers with hundreds of authors are observed in the data, but it is unclear whether authors of this type of papers actually have substantial interpersonal interactions, that is, this type of coauthorship may not be a reliable measure of collaboration. Furthermore, theoretically, the *hyper-authorship* (Cronin, 2001) is beyond the scope of this study. Therefore, 74 scientists who have ever had a paper published between 2005 and 2007 with more than 15 authors are excluded from the sample. Instead of only dropping out one scientist's "problematic" papers while using his remaining papers, the whole observation of this scientist is dropped out, because the former may introduce measurement errors. At the paper level, about 97.5% of the papers have no more than 15 authors. Two different thresholds, 10 (95% of the papers) and 29 (99%) are also tried and would not change the conclusions. In addition, the distribution of author number is very similar between fields, so the same threshold is applied across all the five fields.

Another restriction is about the number of coauthor. A certain number of coauthors are needed to give a reliable measurement of tie strength skewness, so 150 egos with less than five coauthors between 2005 and 2007 are excluded. Removing this restriction would not change the conclusions. Furthermore, some variations in tie strength are needed for calculating skewness, so 128 authors are further excluded. Keeping these authors and using an adjusted skewness measure allowing variance to be 0 would not change the results. In summary, results reported here are based on 576 American scientists.

4.2.3. Model Estimation

The *ln* of the dependent variables are roughly normally distributed (

Figure 3), so the *ln* of these variables are used for OLS models. Furthermore, the dependent variables, total and maximum citations, are non-negative count variables, so the negative binomial models are adopted. The average citations are also non-negative but not always integers, so the integer value of *Cite.AVG* is used for negative binomial regressions. One alternative to deal with count variable is the Poisson model. However, given the evident over-dispersion in the data, the negative binomial model is more appropriate. Likelihood ratio tests also suggested that negative binomial models are significantly better than Poisson models. Furthermore, the dependent variables might be zero-inflated because of using a five-year citation time window and a three-year publication time window. Some scientists observed to have no citations may actually have some citations if I had used a longer time window of observation. To address this problem, zero-inflated negative binomial models are also estimated. However, the Vuong tests are insignificant, indicating that zero-inflated models are not superior to standard ones. Therefore, standard negative binomial fixed-effect models are adopted.

4.2.4. Findings

Descriptive statistics and correlations are reported in **Table 7**. 43% of the sampled scientists are female, 82% are non-Hispanic White in terms of race, and the average career age is about 18.5 years. On average these scientists published about 9 papers between 2005 and 2007, and their 2005-2007 papers received about 141 total citations within five years after publishing. Furthermore, on average, their average citation rate was 14.5, and their mostly highly cited papers got 41 citations.

In addition, Lagged dependent variables and *Pubs* are very highly correlated with each other, as well as with other independent variables, while the correlations between other independent variables are low. Therefore, lagged dependent variables and *Pubs* may cause multicollinearity problems. To address this issue, the following models are also estimated: (1) delete these two variables, (2) delete one control variable at a time (out of all control variables), and (3) delete all control variables. Regression results do not change significantly.

Table 7. Tie Strength: Descriptive Statistics

		mean	sd	min	max	1	2	3	4	5	6	7	8	9	10
1	Cite.SUM	141.07	173.94	1	1673										
2	Cite.SUM.lag	125.71	176.67	0	1847	.61									
3	Cite.AVG	14.44	11.21	.25	80	.67	.37								
4	Cite.AVG.lag	15.11	13.56	0	134.75	.32	.57	.46							
5	Cite.MAX	40.69	46.64	1	516	.79	.39	.79	.29						
6	Cite.MAX.lag	37.19	39.90	0	306	.48	.78	.43	.80	.44					
7	Pubs	9.11	6.35	2	49	.71	.51	.13	.07	.37	.27				
8	Network Size	19.90	14.92	5	129	.64	.53	.24	.16	.42	.29	.74			
9	Tie Strength AVG	1.47	0.35	1.04	3.75	.30	.15	.01	-.04	.12	.06	.46	.08		
10	Skewness	1.75	1.00	-.75	5.04	.28	.21	.14	.11	.23	.13	.34	.51	-.25	
11	Career Age	18.47	9.37	2	46	.05	.07	-.10	-.09	-.03	-.01	.14	.09	.10	.05
12	Gender-Female	0.43	0.50	0	1	-.02	-.03	.04	-.02	.02	-.01	-.08	-.06	.00	-.02
13	Race-White	0.82	0.39	0	1	-.05	-.01	-.02	.01	-.07	-.01	-.03	-.01	-.08	.00
14	Field-BIOL	0.20	0.40	0	1	.01	.07	.10	.19	.00	.11	-.05	.05	-.15	.07
15	Field-CHEM	0.30	0.46	0	1	.18	.17	.07	.06	.04	.08	.21	.17	.24	.02
16	Field-CS	0.11	0.31	0	1	-.13	-.13	-.15	-.11	-.09	-.09	-.10	-.11	-.06	-.06
17	Field-EAS	0.22	0.41	0	1	-.06	-.10	.08	-.05	.04	-.05	-.15	-.07	-.22	.07
18	Field-EE	0.17	0.38	0	1	-.07	-.07	-.15	-.14	-.03	-.09	.05	-.09	.16	-.12

Number of observations (ego): 576. Correlations > 0.082 are significant at p<.05.

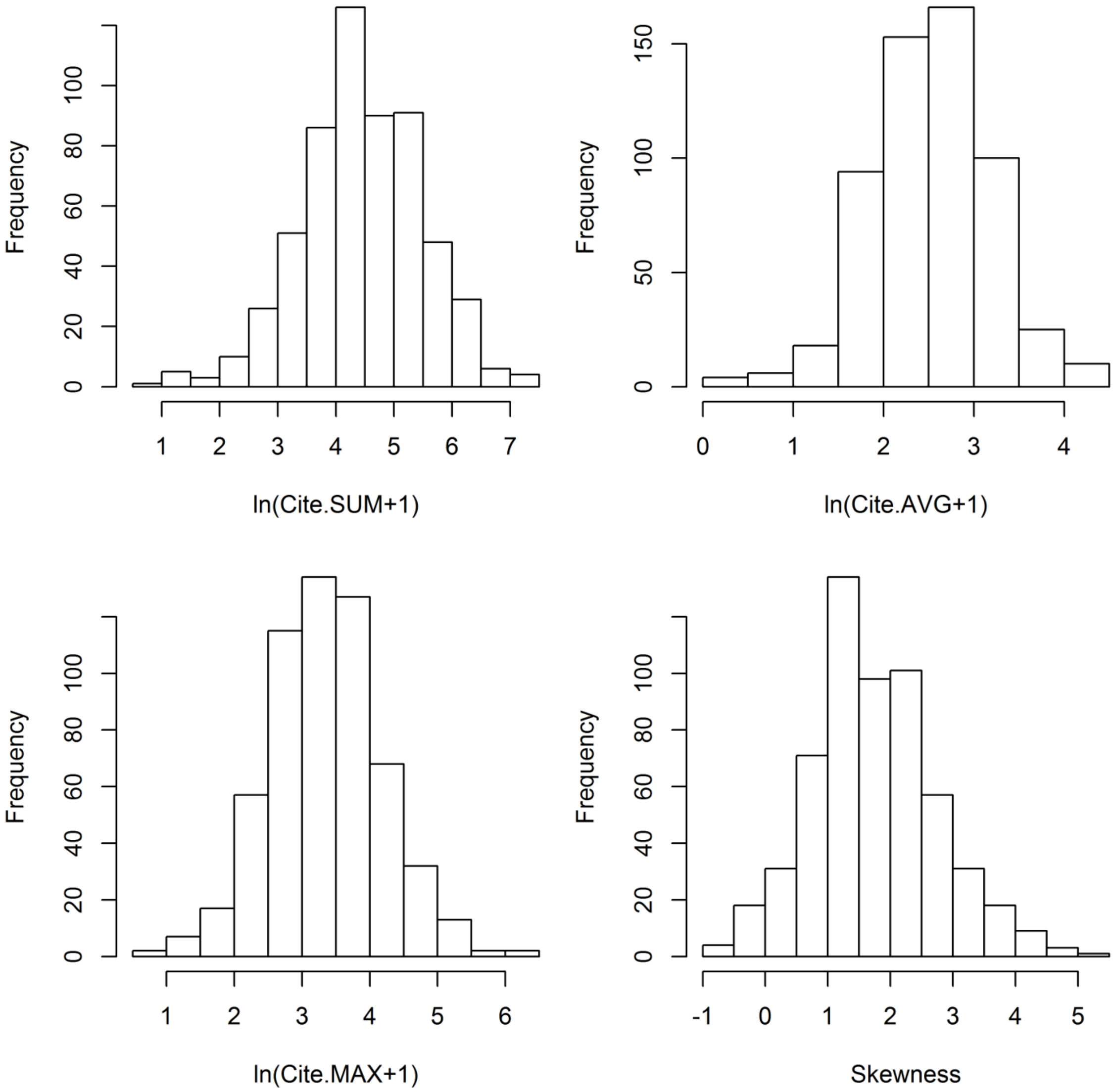


Figure 3. Tie Strength: Histograms

Table 8. Tie Strength: Cite.SUM Models

	1	2	3	4
	Cite.SUM	Cite.SUM	Cite.SUM	Cite.SUM
	NB	NB	NB	OLS
Intercept	2.35 *** (0.07)	2.45 *** (0.07)	2.42 *** (0.08)	1.86 *** (0.20)
ln(Y.lag+1)	0.19 *** (0.01)	0.19 *** (0.01)	0.19 *** (0.01)	0.24 *** (0.03)
ln(Pubs)	0.82 *** (0.03)	0.78 *** (0.03)	0.76 *** (0.03)	0.84 *** (0.08)
Career Age	-0.02 *** (0.00)	-0.02 *** (0.00)	-0.02 *** (0.00)	-0.02 *** (0.00)
Career Age^2	0.00 *** (0.00)	0.00 *** (0.00)	0.00 *** (0.00)	0.00 * (0.00)
Gender-Female	0.06 * (0.03)	0.06 * (0.03)	0.05 (0.03)	0.12 (0.08)
Race-White	-0.09 *** (0.03)	-0.08 *** (0.03)	-0.07 *** (0.03)	-0.05 (0.07)
Field-CHEM	-0.14 *** (0.03)	-0.16 *** (0.03)	-0.16 *** (0.03)	-0.16 ** (0.08)
Field-CS	-0.48 *** (0.04)	-0.48 *** (0.04)	-0.48 *** (0.04)	-0.52 *** (0.11)
Field-EAS	-0.01 (0.03)	0.00 (0.03)	0.00 (0.03)	0.01 (0.09)
Field-EE	-0.36 *** (0.03)	-0.36 *** (0.03)	-0.36 *** (0.03)	-0.30 *** (0.09)
Network Size	0.01 *** (0.00)	0.02 *** (0.00)	0.01 *** (0.00)	0.01 *** (0.00)
Network Size^2	0.00 *** (0.00)	0.00 *** (0.00)	0.00 *** (0.00)	0.00 ** (0.00)
TieStrAVG		0.14 *** (0.05)	0.36 *** (0.10)	0.32 (0.26)
TieStrAVG^2		-0.11 ** (0.05)	-0.37 *** (0.12)	-0.36 (0.33)
Skewness			0.03 * (0.02)	0.02 (0.04)
Skewness * TieStrAVG			-0.07 * (0.04)	-0.07 (0.11)
Skewness * TieStrAVG^2			0.10 * (0.05)	0.09 (0.14)
χ^2	3971 ***	3979 ***	3990 ***	
$\Delta\chi^2$		8.48 **	11.03 **	
log-likelihood	-19965	-19961	-19955	
R^2 adj				0.65

Number of observations (ego): 576

* p < .10, ** p < .05, *** p < .01

The dependent variable is the total number of citations received by papers published between 2005 and 2007, that is, *Cite.SUM*. Column 1-3 are negative binomial models using the original count variable *Cite.SUM*. Column 4 is the OLS model using natural logarithm transformed dependent variable, *ln(Cite.SUM +1)*. *ln(Y.lag+1)* is *ln(Cite.SUM.lag+1)*, where *Cite.SUM.lag* is the total number of citations received by papers published between 2002 and 2004. The reference group for field is BIOL, for Gender is Male, and for Race is Minority.

Table 9. Tie Strength: Cite.AVG Models

	1	2	3	4
	Cite.AVG	Cite.AVG	Cite.AVG	Cite.AVG
	NB	NB	NB	OLS
Intercept	1.92 *** (0.07)	2.03 *** (0.08)	2.01 *** (0.08)	1.52 *** (0.18)
ln(Y.lag+1)	0.37 *** (0.01)	0.37 *** (0.01)	0.37 *** (0.01)	0.42 *** (0.03)
ln(Pubs)	-0.07 *** (0.02)	-0.12 *** (0.03)	-0.13 *** (0.03)	0.00 (0.07)
Career Age	-0.02 *** (0.00)	-0.02 *** (0.00)	-0.02 *** (0.00)	-0.01 *** (0.00)
Career Age^2	0.00 *** (0.00)	0.00 *** (0.00)	0.00 *** (0.00)	0.00 * (0.00)
Gender-Female	0.07 ** (0.03)	0.07 ** (0.03)	0.06 ** (0.03)	0.12 (0.07)
Race-White	-0.13 *** (0.03)	-0.12 *** (0.03)	-0.11 *** (0.03)	-0.06 (0.06)
Field-CHEM	-0.09 *** (0.03)	-0.11 *** (0.03)	-0.11 *** (0.03)	-0.08 (0.06)
Field-CS	-0.46 *** (0.04)	-0.45 *** (0.04)	-0.46 *** (0.04)	-0.37 *** (0.09)
Field-EAS	0.08 ** (0.03)	0.08 ** (0.03)	0.08 ** (0.03)	0.10 (0.08)
Field-EE	-0.30 *** (0.03)	-0.31 *** (0.03)	-0.31 *** (0.03)	-0.20 ** (0.08)
Network Size	0.01 *** (0.00)	0.02 *** (0.00)	0.01 *** (0.00)	0.01 *** (0.00)
Network Size^2	0.00 *** (0.00)	0.00 *** (0.00)	0.00 *** (0.00)	0.00 * (0.00)
TieStrAVG		0.15 *** (0.05)	0.35 *** (0.10)	0.20 (0.22)
TieStrAVG^2		-0.10 ** (0.05)	-0.38 *** (0.13)	-0.31 (0.28)
Skewness			0.02 (0.02)	0.02 (0.03)
Skewness * TieStrAVG			-0.06 (0.04)	-0.01 (0.10)
Skewness * TieStrAVG^2			0.11 ** (0.05)	0.07 (0.12)
χ^2	1536 ***	1545 ***	1556 ***	
$\Delta\chi^2$		9.18 **	10.61 **	
log-likelihood	-12227	-12222	-12217	
R^2 adj				0.38

Number of observations (ego): 576

* p < .10, ** p < .05, *** p < .01

The dependent variable is the average citation rate of papers published between 2005 and 2007, that is, *Cite.AVG*. Column 1-3 are negative binomial models using the integer value of *Cite.AVG*. Column 4 is the OLS model using natural logarithm transformed dependent variable, *ln(Cite.AVG +1)*. *ln(Y.lag+1)* is *ln(Cite.AVG.lag+1)*, where *Cite.AVG.lag* is the average citation rate of papers published between 2002 and 2004. The reference group for field is BIOL, for Gender is Male, and for Race is Minority.

Table 10. Tie Strength: Cite.MAX Models

	1	2	3	4
	Cite.MAX	Cite.MAX	Cite.MAX	Cite.MAX
	NB	NB	NB	OLS
Intercept	1.56 *** (0.11)	2.01 *** (0.12)	1.96 *** (0.12)	1.58 *** (0.30)
ln(Y.lag+1)	0.41 *** (0.02)	0.43 *** (0.01)	0.42 *** (0.01)	0.40 *** (0.04)
ln(Pubs)	0.26 *** (0.04)	0.05 (0.04)	0.05 (0.04)	0.19 * (0.11)
Career Age	-0.01 *** (0.00)	-0.01 *** (0.00)	-0.01 *** (0.00)	-0.01 *** (0.00)
Career Age^2	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Gender-Female	0.11 *** (0.04)	0.10 ** (0.04)	0.09 ** (0.04)	0.06 (0.10)
Race-White	0.01 (0.03)	0.07 ** (0.03)	0.06 * (0.03)	0.11 (0.08)
Field-CHEM	0.07 ** (0.03)	-0.01 (0.03)	-0.02 (0.03)	-0.07 (0.09)
Field-CS	0.27 *** (0.05)	0.30 *** (0.05)	0.28 *** (0.05)	0.05 (0.13)
Field-EAS	0.36 *** (0.04)	0.39 *** (0.04)	0.36 *** (0.04)	0.23 ** (0.10)
Field-EE	0.15 *** (0.04)	0.11 ** (0.04)	0.09 ** (0.04)	-0.02 (0.11)
Network Size	0.02 *** (0.00)	0.02 *** (0.00)	0.02 *** (0.00)	0.01 *** (0.00)
Network Size^2	0.00 *** (0.00)	0.00 *** (0.00)	0.00 *** (0.00)	0.00 * (0.00)
TieStrAVG		0.53 *** (0.05)	1.17 *** (0.11)	0.75 ** (0.29)
TieStrAVG^2		-0.23 *** (0.05)	-0.85 *** (0.14)	-0.72 ** (0.36)
Skewness			0.04 ** (0.02)	0.04 (0.04)
Skewness * TieStrAVG			-0.26 *** (0.05)	-0.14 (0.12)
Skewness * TieStrAVG^2			0.25 *** (0.06)	0.23 (0.15)
χ^2	1524 ***	1617 ***	1671 ***	
$\Delta\chi^2$		9.18 **	10.61 **	
log-likelihood	-11728	-11682	-11655	
R^2 adj				0.42

Number of observations (ego): 396. Only scientists with more than five papers are included. Results are similar if instead using all 576 scientists.

* p < .10, ** p < .05, *** p < .01

The dependent variable is the maximum number of citations among papers published between 2005 and 2007, that is, *Cite.MAX*. Column 1-3 are negative binomial models using the original count variable of *Cite.MAX*. Column 4 is the OLS model using natural logarithm transformed dependent variable, *ln(Cite.MAX +1)*. *ln(Y.lag+1)* is *ln(Cite.MAX.lag+1),* where *Cite.MAX.lag* is the maximum number of citations among papers published between 2002 and 2004. The reference group for field is BIOL, for Gender is Male, and for Race is Minority.

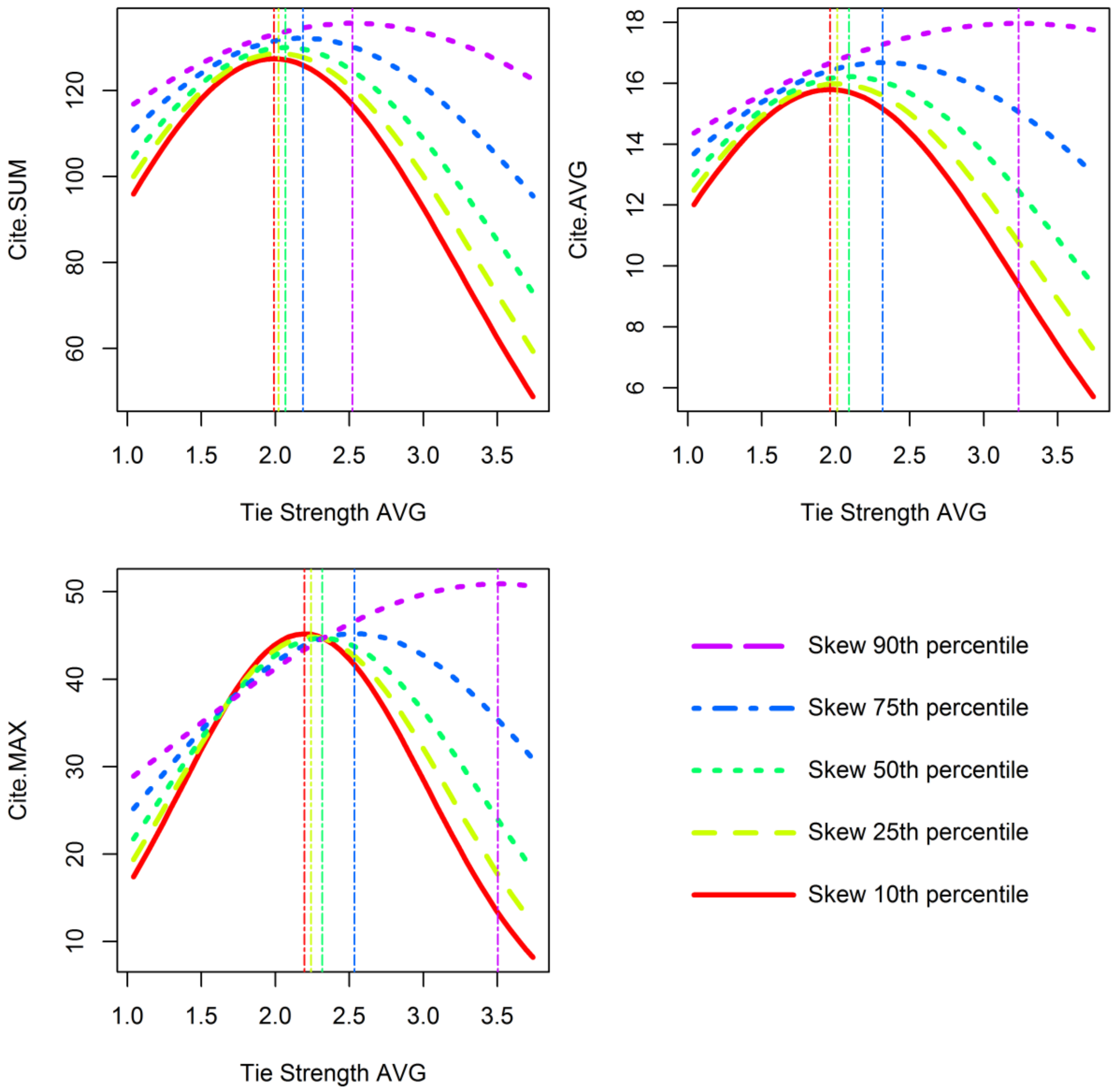


Figure 4. Tie Strength: Effects on Citations

Table 8, **Table 9**, and **Table 10** report regression results for *Cite.SUM*, *Cite.AVG*, and *Cite.MAX* models, respectively. In each table, from column 1 to 3, variables of interest are added sequentially. The fourth column reports OLS regress results as a benchmark. Negative binomial and OLS regression yield very similar results in terms of the direction and size of the coefficients, but coefficients on focal independent variables are almost all significant in negative binomial models but not in OLS models. Furthermore, using three different citation statistics as dependent variables yields similar estimations of network effects.

Given that negative binomial models are more appropriate for analyzing count variables, the following interpretation are based on negative binomial models. Here are hypothesis testing results:

- Hypothesis 2 is supported, that is, there is an inverted U-shaped relationship between network average tie strength and creativity. The coefficients on *Tie Strength AVG* are positive while the coefficients on *Tie Strength AVG^2* are negative, both significant.
- Hypothesis 3 is supported, that is, skewed networks have higher creativity when the network average tie strength is high. The coefficientits on *Skewess* are all positive (significant in **Table 8** and **Table 10** but insignificant in **Table 9**). Because *Tie Strength AVG* is centered, the coefficients on *Skewess* represent its effects when *Tie Strength AVG* is at its mean. Furthermore, effects of *Skewness* are even bigger when the *Tie Strength AVG* is higher, as shown in **Figure 4**, which plots estimated citations at different *Tie Strength AVG* grouped by different level of skewness.
- Hypothesis 4 is supported, that is, more skewed networks are less sensitive to changes in network average tie strength. Interaction effects between *Tie Strength AVG* and *Skewness* are also significant and have the same direction as hypothesized, that is, the positive effect as tie strength increases is smaller for

more skewed networks. The interaction effects between *Tie Strength AVG^2* and *Skewness* are also significantly positive, which confirms the hypotheses that the negative effect of tie strength increases is also smaller for more skewed networks. One exception is that the interaction effect between *Tie Strength AVG* and *Skewness* is insignificant in the *Cite.AVG* model, but in the same direction as hypothesized.

Figure 4 plots the estimated citation counts against *Tie Strength AVG,* in high- and low- skew networks separately. An increase in *Tie Strength AVG* initially has a positive effect on total, average, and maximum citations, but turns into a negative effect after a threshold. However, *Tie Strength AVG* effects are different between high and low skew networks. In high skew networks, citation counts are less sensitive to changes in network average tie strength, that is, the marginal effect is smaller in both the initial positive-effect stage and the later negative-effect stage. Graphically speaking, the slopes (both the initial positive and the later negative sections) of the effect curve are flatter in more skew networks.

4.2.5. Alternative Explanations

For the observed negative effect of network average tie strength after it reaches a certain threshold, I argued that it is because networks dominated by strong ties are unable to generate novel ideas. One alternative explanation is that strong ties may present network constraints impeding creativity. First, strong relationships are binding, imposing obligations to cooperate, which may help performance of the group, but is not optimal for the individual personally. The binding effect reduces individual autonomy (Hansen, 1999; Weick, 1976), preventing him from strategically allocating energy and efforts across different collaborations to maximize his personal gains. Second, network constraint also prevents individuals from altering current network structure to establish new and more creative networks (Gabbay, 1997). To assess this alternative explanation, the Pearson correlation between the number of new collaborators in 2008 and the network average tie strength between 2005 and 2007 is calculated, which is 0.03 and insignificant (*p*=0.31). The Spearman correlation is also 0.03 and insignificant (*p*=0.34). Therefore, there is no evidence that strong ties restricted developing new collaborative relations.

In addition, the performance of scientists is highly unequal and skewed, and the results may be biased by some star scientists who perform extremely well. More specifically, because the bulk of scientists have a medium level of tie strength, so it is more likely to find some extremely creative scientists at this medium level. These extreme cases may significantly drive the estimated citations up and create an artifact the same as observed here. Several checks were performed to address this concern. First, the *ln* transformed dependent variables are roughly normally distributed, so there is so significant evidence of outliners. Second, the *Cook's distance* and the *leverage statistic* of OLS models also suggest no evidence that some observations have significantly higher influence in the regression results than others. Third, if I exclude the top 10% scientists, regression results are not significantly changed. Therefore, it is very unlikely that the observed tie strength effect is actually a data artifact.

4.3. The Creativity vs. Cost Hypotheses

Hypothesis 5a (Cost-Reduction Hypothesis): strong-tie-collaborations have lower average but higher maximum observed-creativity, compared with weak-tie-collaborations.

Hypothesis 5b (Creativity-Decline Hypothesis): strong-tie-collaborations have lower average and maximum observed-creativity, compared with weak-tie-collaborations.

This section tests two competing hypotheses: whether the observed low citation rate of strong-tie-collaborations is because of reduced costs or diminished creative capacity. For empirical testing, strong-tie-collaborations are operationalized as repeated-collaborations (i.e., papers with only repeated collaborators), and weak-tie-collaborations as new-collaborations (i.e., papers with only new collaborators). The defining feature of tie strength is therefore the existence or absence of prior collaborating experiences.

There are two important points concerning this operationalization. First, there is clearly a selection effect making repeated-collaborations more productive or creative than new-collaborations. After the first collaboration, those collaborative relations with an unpleasant experience will be abandoned, in other words, repeated-collaborations have already sorted out the worst ones while new-collaborations still include those. This selection effect counterbalances the creativity-decline effect but reinforces the cost-reduction effect. As a result, I am not very confident that the creativity-decline hypothesis should be rejected if this operationalization rejects it, but I am relatively confident in the confirming result. On the other hand, I am confident in the rejection but not the confirmation of the cost-reduction hypothesis.

Second, according to the creativity-decline hypothesis, creativity of repeated collaborations is likely to decrease gradually, that is, the creativity of second collaboration might be only a little lower than the first collaboration. Therefore, classifying collaborations into new- and repeated- ones is not the best strategy for testing the creativity-decline hypothesis. As a result, I am relatively confident in the confirmation but not the rejection of the creativity-decline hypothesis. On the other hand, the transaction costs are likely to have a dramatic drop after the first collaboration. New collaborators are more likely to have trust issues and behave opportunistically because of information asymmetry, that is, I'd better shirk when I am not sure whether my collaborator would. The situation for the second-time collaboration is much improved because of the selection effect. In addition, it is in the first collaboration that collaborators have communication and epistemological problems and have to invest in teaching each other about their own specialties. Shared mental model develops during the collaboration, and collaborators will have a much easier process at the second time. Therefore, classifying collaborations into new- and repeated- ones captures this disruptive cost-reduction, and I will be confident in the rejecting result of the cost-reduction hypothesis.

4.3.1. Paired Tests

Two different set of analyses are implemented to test these two competing hypotheses. The first strategy is paired *t* and non-parametric tests. Each ego's papers between 2005 and 2007 are classified into four types: *solo* (single authored papers), *new* (with only new collaborators), *repeated* (with only repeated coauthors), and *mix* (with both new and repeated coauthors). Egos may have several or none papers in each category, and there are 443 egos with both types (i.e., *new* and *repeated*) of papers in this period. These 443 egos are used for analysis. Several citation statistics are calculated for the new- and repeated- collaboration papers for each ego: average, minimum, median, and maximum citations. The research question is whether these statistics are different between new- and repeated- collaboration papers. These statistics are skewed, so they took natural logarithm before the t-tests and the Wilcoxon Signed-Rank tests. The Wilcoxon Signed-Rank tests using the original scale of these statistics are also reported.

Testing results are reported in **Table 11**, and conclusions of these three different tests are all the same. New-collaborations have higher average citations, lower minimum citations, no different median citations, and higher maximum citations. The lower minimum citations of new-collaborations reflect the selection effect (repeated-collaborations have sorted out unproductive collaborative relations), and the higher maximum citations of new-collaborations support the creativity-decline hypothesis and reject the cost-reduction hypothesis. The distribution of between-type (new and repeated) differences in these citation statistics are plotted in **Figure 5**.

Table 11. Creativity vs. Cost: Paired Tests

	mean difference (ln)	p-value t-test (ln)	p-value Wilcoxon Signed-Rank Test	p-value Wilcoxon Signed-Rank Test (ln)
Cite.AVG	0.13	0.00	0.02	0.01
Cite.MIN	-0.10	0.09	0.06	0.05
Cite.MED	0.08	0.07	0.43	0.15
Cite.MAX	0.24	0.00	0.00	0.00

Number of paired samples (egos): 443.
All three tests are paired and two-sided.
Citation statistics are the mean (AVG), minimum (MIN), median (MED), and maximum (MAX) citations of each ego’s new (and repeated) collaboration papers published between 2005 and 2007. The difference between new and repeated papers for each observation is the *ln* value of NEW minus the *ln* value of REP, so positive numbers indicate higher value for the NEW group.

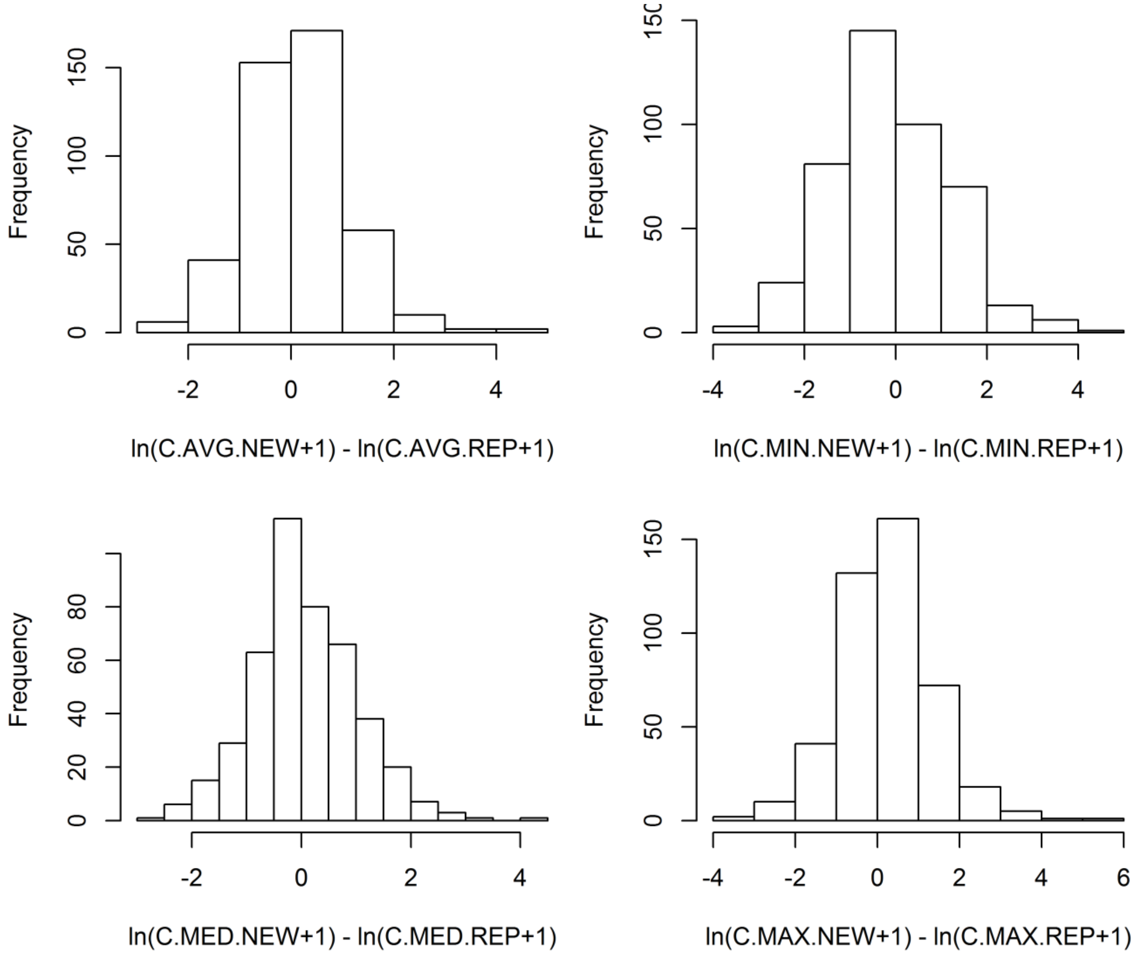


Figure 5. Creativity vs. Cost: Paired Tests Histograms

4.3.2. Quantile Regressions

The second analysis is quantile regressions at the paper level to estimate citation distribution differences between new- and repeated- collaboration papers. There is only one sample restriction at the paper level: papers included for analysis are those published more than three years after the year when the ego’s first paper was published, to allow classifying coauthors and papers. In total, I have 7,408 papers of 1,102 egos for analysis, and these papers are classified into four types: *solo*, *new*, *rep*, and *mix*.

I specify the following equation for citation counts for the j-th paper of the i-th ego

$$Cite_{ij} = \beta_0 + \beta_1 \cdot Type_{ij} + \beta_2 \cdot YearF_{ij} + \beta_3 \cdot Field_i + \beta_4 \cdot Rank_i + \beta_5 \cdot Gender_i$$
$$+ \beta_6 \cdot Race_i + b_{1i} + \varepsilon_i$$

where

- $Cite_{ij}$ is the value of the dependent variable for the *j*-th paper of the *i*-th ego, the number of citations. *ln* transformed.
- $Type_{ij}$ is the type for the *j*-th paper of the *i*-th ego: solo, new, rep, and mix. A set of dummies are used for model fitting.
- $YearF_{ij}$ is the publication year for the *j*-th paper of the *i*-th ego: 2005, 2006, and 2007. It is treated as a factor variable, so a set of dummies are used to control year fixed effects.
- $Field_i$ is the research field of ego *i*, $Rank_i$ is the academic rank of ego *i*, $Gender_i$ is the gender of of ego *i*, and $Race_i$ is the race of of ego *i*.
- b_{1i} is the random intercept to account for ego heterogeneities.

First, models without control variables (*Field*, *Rank*, *Gender*, and *Race*) are fitted, and the field, rank, gender, and race differences are presumably absorbed by the ego random effects. A series of percentiles of the citation distributions are fitted: 5th, 10th, 15th, …, 95th. The estimated 19 percentiles for each type of paper are plotted in **Figure 7**, and estimated models for five of them (10th, 25th, 50th, 75th, and 90th) are reported in **Table 13** column 2. In addition, a least squares dummy variable (LSDV) model is also fitted as a benchmark. It estimates the difference in average citations between different types of papers. In the LSDV model, ego fixed effects are controlled, that is, replace b_{1i} by a set of ego dummies. LSDV model estimations are presented in **Table 13** column 1.

Second, quantile mixed-effect models with these control variables are estimated and reported in **Table 14**, results are not significantly different from models without control variables.

Results (**Table 13**, **Table 14**, and **Figure 7**) suggest that mixed-collaborations have the highest citations at all percentiles, and solo-authored papers lowest at all percentiles. Focusing on differences between new- and repeated- collaborations, **Table 13** column 1 suggests that the mean values of *ln(citations+1)* are not significantly different between new- and repeated- collaborations. However, compared with new-collaborations, repeated-collaborations do have more citations at low percentiles (selection effect) and fewer citations at high percentiles. The difference at high percentiles supports the creativity-decline hypothesis and rejects the cost-reduction hypothesis, but the difference is insignificant. In addition, given that (1) there is a significant selection effect, (2) the selection effect raises the citation counts at high percentiles for repeated-collaborations, (3) citation counts at high percentiles is still lower for repeated-collaborations than new-collaborations, although insignificant, and (4) creativity-decline hypothesis predicts that repeated-collaborations have few citations at high percentiles, while cost-reduction hypothesis predicts the opposite, it should be safe to conclude that the creativity-decline hypothesis is supported, while the cost-deduction hypothesis is rejected.

In conclusion, results of paired tests and quantile regressions concur, and the creativity-decline hypothesis (Hypothesis 5b) is supported while the cost-reduction hypothesis (Hypothesis 5a) is rejected.

Table 12. Creativity vs. Cost: Descriptive Statistics

		mean	sd	median	min	max	1
1	*ln*(Cite+1)	2.29	1.07	2.30	0	6.83	
2	Type-solo	0.02	0.15	0	0	1	-0.07
3	Type-new	0.33	0.47	0	0	1	-0.08
4	Type-rep	0.22	0.41	0	0	1	-0.03
5	Type-mix	0.43	0.50	0	0	1	0.12
6	Type	solo (singled-authored paper) 171, new (coauthored papers and all coauthors are new) 2433, rep (coauthored papers and all coauthored are repeated)1647, mix (coauthored papers and coauthors include both new and repeated ones) 3227					
7	Field	BIOL 1536, CHEM 2279, CS 786, EAS 1572, EE 1305					
8	Rank	Assistant Prof 1246, Associate Prof 1944, Full Prof 4288					
9	Gender	Female 3049, Male 4429					
10	Race	Non-Hispanic White 5952, Minorities 1526					

Number of observations (papers): 7478

ln(Cite+1) is the ln transformed citations of each paper, *Type-solo*, *Type-new*, *Type-rep*, and *Type-mix* are dummies.

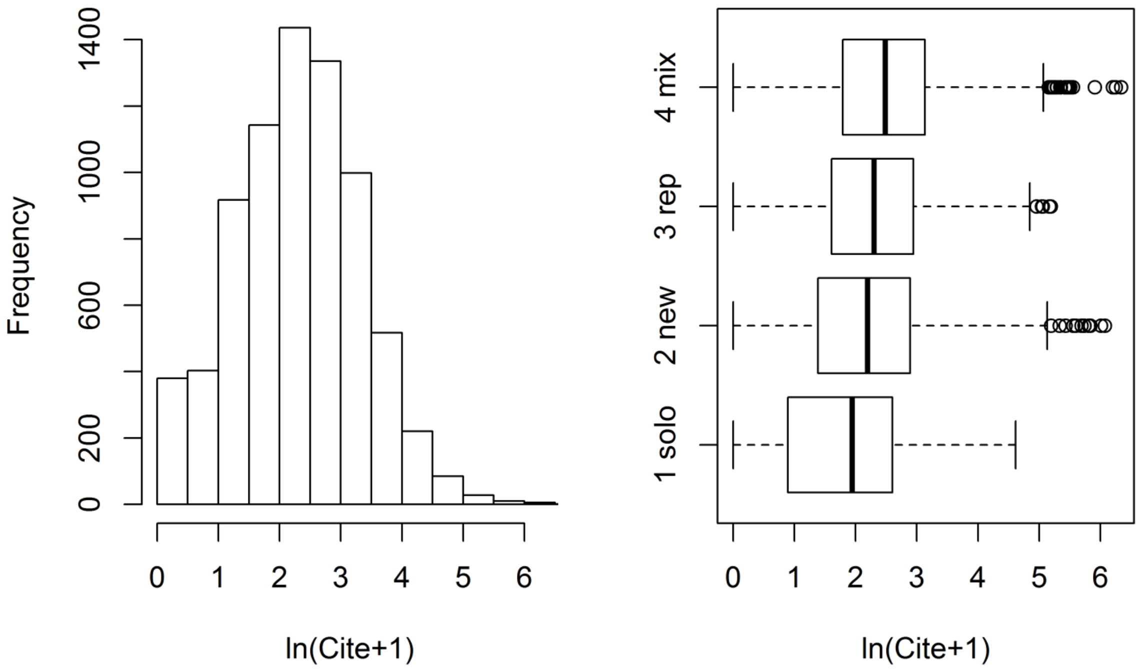


Figure 6. Creativity vs. Cost: Exploratory Plots

Table 13. Creativity vs. Cost: Models A

	1	2				
	ln(Cite+1)	ln(Cite+1)				
	LSDV	Linear Quantile Mixed Model				
		10%	25%	50%	75%	90%
Type-rep	0.02 (0.03)	0.40 ** (0.19)	0.12 (0.08)	0.03 (0.04)	0.00 (0.03)	-0.06 (0.05)
Type-solo	-0.39 *** (0.09)	0.00 (0.08)	-0.39 (0.24)	-0.43 *** (0.15)	-0.26 (0.16)	-0.16 (0.14)
Type-mix	0.12 *** (0.03)	0.41 *** (0.14)	0.30 *** (0.07)	0.20 *** (0.03)	0.12 *** (0.04)	0.13 *** (0.05)
Ego dummies	Yes	No	No	No	No	No
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
R2 adj	0.27					
likelihood ratio		0 ***	118 ***	84 ***	65 ***	69 ***

Number of observations (papers): 7408

Number of groups (egos): 1102

* p < .10, ** p < .05, *** p < .01

The dependent variables for the least squares dummy variable (LSDV) model and linear quantile mixed modes are all *ln* transformed. The intercept of LSDV model is fixed and not reported, the intercepts of quantile mixed models are random and not reported. The reference type group is *Type-new*.

Table 14. Creativity vs. Cost: Models B

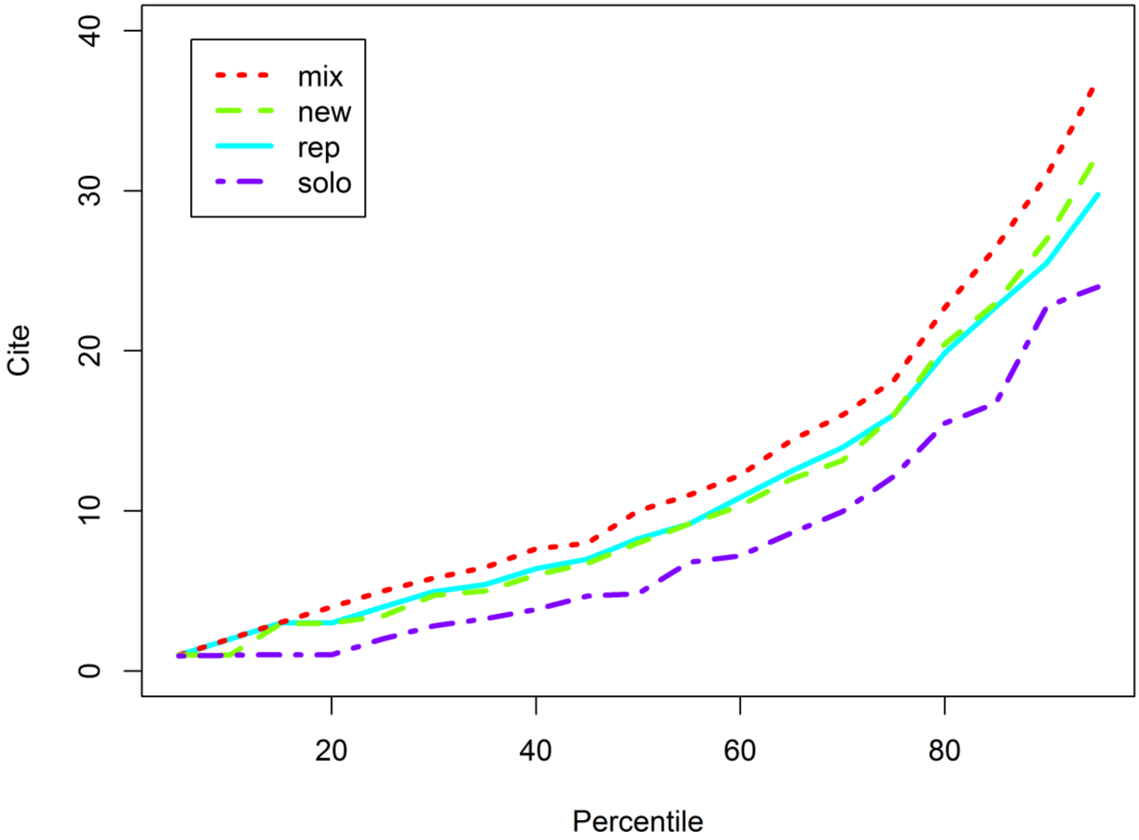
	1									
	ln(Cite+1)									
	Linear Quantile Mixed Model									
	10%		25%		50%		75%		90%	
Type-rep	0.00	(0.07)	0.07 *	(0.04)	0.02	(0.04)	0.00	(0.04)	-0.02	(0.04)
Type-solo	0.00	(0.28)	-0.28 **	(0.11)	-0.33 **	(0.12)	-0.23 *	(0.13)	-0.18	(0.15)
Type-mix	0.00	(0.08)	0.21 ***	(0.04)	0.14 ***	(0.04)	0.10 **	(0.04)	0.13 ***	(0.04)
Field-CHEM	0.00	(0.13)	0.05	(0.08)	0.00	(0.13)	0.08	(0.11)	-0.01	(0.09)
Field-CS	-0.69 ***	(0.10)	-0.45 ***	(0.14)	-0.61 ***	(0.10)	-0.54 ***	(0.08)	-0.55 ***	(0.09)
Field-EAS	0.00	(0.10)	0.18	(0.11)	-0.10	(0.08)	0.00	(0.08)	-0.03	(0.09)
Field-EE	-0.69 ***	(0.09)	-0.44 ***	(0.11)	-0.46 ***	(0.12)	-0.40 ***	(0.09)	-0.33 ***	(0.09)
Rank-Assoc	0.00	(0.05)	-0.14	(0.09)	-0.36 ***	(0.09)	-0.05	(0.09)	-0.23 **	(0.10)
Rank-Full	0.00	(0.05)	-0.22 ***	(0.08)	-0.24 **	(0.09)	-0.15 *	(0.08)	-0.22 ***	(0.08)
Gender-Female	0.00	(0.05)	0.08	(0.07)	0.12 *	(0.06)	0.08	(0.05)	0.12	(0.08)
Race-White	0.00	(0.03)	-0.02	(0.08)	0.10	(0.09)	0.02	(0.08)	-0.04	(0.07)
Year dummies	Yes		Yes		Yes		Yes		Yes	
likelihood ratio	370 ***		304 ***		284 ***		254 ***		197 ***	

Number of observations (papers): 7408

Number of groups (egos): 1102

* p < .10, ** p < .05, *** p < .01

The dependent variables are all *ln* transformed. The intercepts are random and not reported. The reference type group is *Type-new*.



The model specified in Table 12 is used for estimation. Fitted percentiles are: 5th, 10th, 15th, … , 95th, (i.e., 19 different percentiles). The estimated 19 percentiles of *ln(Cite+1)* for four different types are transformed to the original scale of Cite, and then plotted.

Figure 7. Creativity vs. Cost: Quantile Regression Estimations

4.4. A Replicate Study

As discussed in Section 3.3, there are several limitations of using coauthorship as proxies for collaboration and the number of coauthored papers as a measure of tie strength. Therefore, this section uses survey data to test the validity of the bibliometric measure of tie strength and some of the previous findings. First, a factor analysis using survey data is implemented to extract a survey-based tie strength factor. Second, the association between survey-based- and bibliometric- measures (i.e., survey-based tie strength factor and the number of coauthored papers) is investigated. Third, the relationship between survey-based measure of tie strength and citations of the tie is also investigated.

4.4.1. Tie Strength Factor Analysis

Granovetter (1973) defined tie strength as "a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie." Therefore, I treat tie strength as a latent variable and several survey items as measures of this latent variable. These survey items include: how long the ego has known the alter, frequency of their personal contact, whether the ego thinks the alter as a close friend, level of the ego's understanding of the alter's area of expertise, career development resources provided by the alter to the ego, whether the ego seeks advices from the alter, and whether the alter talks with ego regularly about university/department-related issues. Details about these survey items are reported in **Table 15**. 5,185 alters with complete relevant survey data are used for the factor analysis. Descriptive statistics are reported in **Table 16**.

All these survey measures are ordinal and positively correlated. A structural equation model is fitted, in which the tie strength is treated as the only latent variable, and all survey items as measures of this latent variable. The "sem" package available in R is used for model fitting, which provides appropriate treatments for ordinal data instead of treating them as continuous. It uses the polychoric correlations between ordinal variables for model construction and uses bootstrapping to estimate standard errors (Fox, 2006). The latent variable is extracted as the tie strength factor. Correlations between this factor and survey items are reported in **Table 16**, and the variables factor map is presented in **Figure 8**, which uses the principal component analysis method and pretends that the survey measures are continuous.

4.4.2. Survey-based Tie Strength Factor and The Number of Coauthored Papers

For notation simplicity, I define *close-coauthors* as coauthors that are nominated by the ego in the survey as close collaborators, and *other-coauthors* as those not nominated. In other words, the set of *close-coauthors* is the intersection of the set of *coauthors* (identified from bibliometric data during a pre-specified period, e.g., 2005-2007 and 2005-2010) and the set of *alters* (collected from survey data). *Close-coauthor* is a subset of *coauthor*, specifically, *close-coauthors* are *coauthors* that are nominated as close collaborators (i.e., alters). *Close-coauthor* is also a subset of *alters*, specifically, *close-coauthors* are *alters* who had at least one coauthored papers with the ego within a pre-specified period.

To test whether the number of coauthored papers is a good measure of tie strength, the numbers of coauthored papers are compared (1) between close-coauthors and other-coauthors, and (2) among close-coauthors with survey-based tie strength factor available.

Between 2005 and 2007, 1,157 egos have 24,094 coauthors in total, and 2,169 of them are close-coauthors[7]. Since the survey asked egos to nominate their "closest" collaborators, presumably egos have strong ties with the 2,169 close-coauthors than with the 21,925 other-coauthors. Therefore, the first analysis is to compare the numbers of coauthored papers between close-coauthors and other-coauthors. As reported in **Table 19**, after controlling for ego fixed-effects, the expected number of papers of close-coauthors is $e^{.31} = 1.36$ times of that of other-coauthors.

The second study analyzes only close-coauthors. Here, for matching coauthors and alters, I extend the sample from 2005-2007 papers to 2005-2010 papers. After removing observations with missing values, there are 2,392 close-coauthors of 820 egos left for analysis. Regression results (**Table 21**) suggest that close-coauthors with higher values of the survey-based tie strength factor also have significantly more coauthored papers with the ego, this is true for both between-ego and within-ego comparisons. Details of this regression model will be introduced in the next section, together with analyses of citations.

In summary, I can conclude here that the bibliometric measure of tie strength is valid.

[7] These egos have 6,031 alters in total, so not all alters have coauthored papers with the ego. If I extend the sample to include papers published after 2007, there will be 2,827 alters having coauthored papers instead of 2,169. However, the 2005-2007 data are used for this analysis, because there will be much more coauthors in the 2005-2010 sample, and it is unclear whether egos have already known those new coauthors appeared after 2007, in which case, egos were not be able to evaluate the tie strength with these new coauthors at the time of survey. In addition, even if use the 2005-2010 sample, there are still a considerable number of alters without coauthored papers. This could be because they have coauthored publications other than WOS journal articles, such as papers in journals not indexed by WOS, review papers, book chapters, reports, and conference papers.

Table 15. Tie Strength Factor Analysis: Variable Descriptions

Variables	Measures
Age	length of time knowing individuals named "How long have you known the individuals you named?" 1=Less than three years, 2=3-6 years, 3=More than three years
Frequency	frequency of personal contact with individuals named "In the past academic year, how frequently were you in personal contact with these individuals?" 4 = at least daily, 3 = about weekly, 2 = about monthly, 1 = less often
Friend	alter is close friend 0=No, 1=Yes
Understanding	level of understanding of alters area of expertise "Recognizing that you interact with people who have different areas of specialization, please indicate the extent to which you understand their area of expertise:" 3 = detailed understanding, 2 = working understanding, 1 = little to no understanding
Resource	Provided assistance to career development, summation of the following three dummies: - reviewed your papers or proposals prior to submission - introduced you to potential collaborators outside of your university - invited you to join a grant proposal team
Advise	sought for professional/career advice 0=No, 1=Yes
Talk	regularly talk to you about university/department related issues 0=No, 1=Yes

The unit of analysis is alter-ego pair.

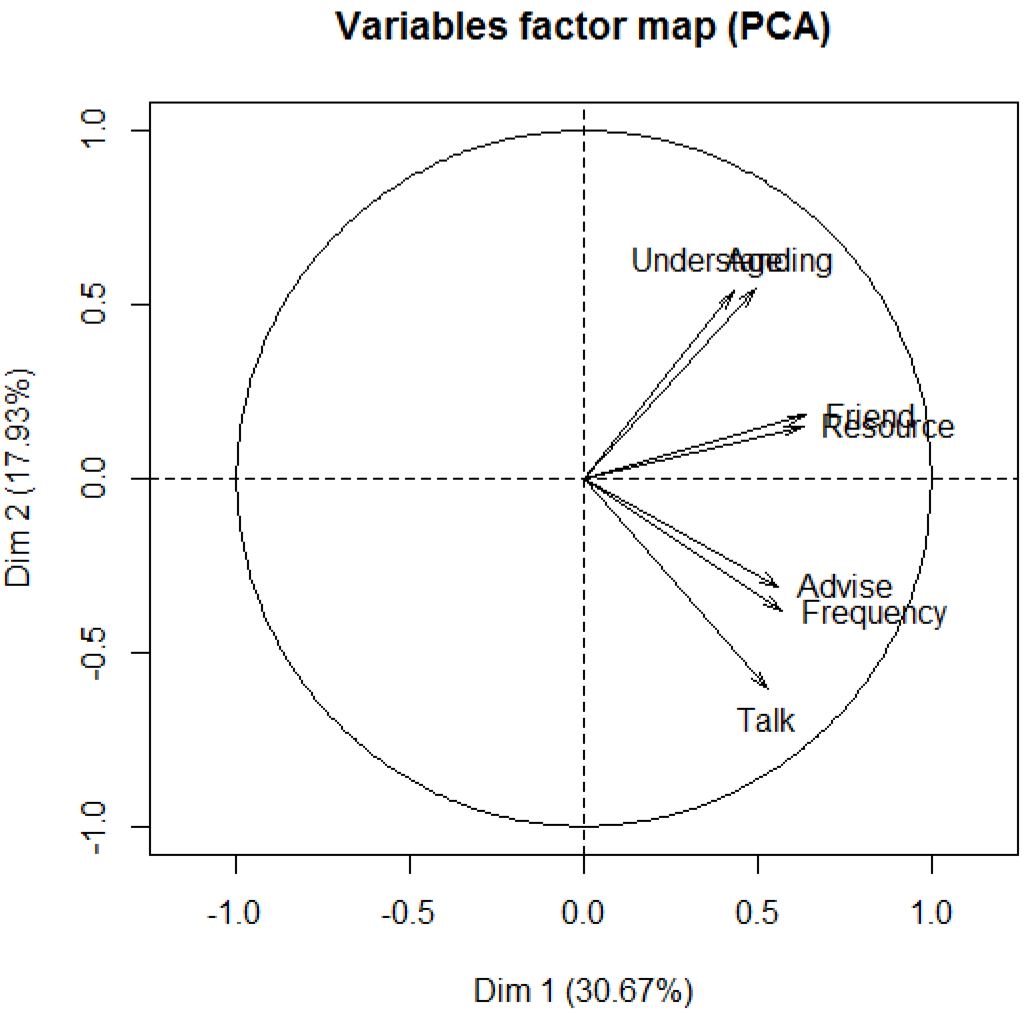


Figure 8. Tie Strength Factor Analysis: Variables Factor Map

Table 16. Tie Strength Factor Analysis: Descriptive Statistics

		mean	sd	median	min	max	1	2	3	4	5	6	7
1	Age	2.36	0.74	3	1	3							
2	Frequency	2.28	0.97	2	1	4	.09						
3	Friend	0.26	0.44	0	0	1	.30	.27					
4	Understanding	2.46	0.57	2	1	3	.26	.07	.18				
5	Resource	0.75	0.86	1	0	3	.21	.18	.28	.21			
6	Advise	0.10	0.30	0	0	1	.09	.18	.15	.12	.28		
7	Talk	0.12	0.32	0	0	1	.05	.37	.18	.01	.14	.33	
8	Tie Strength	0.01	0.78	-0.21	-1.01	2.98	.40	.67	.58	.35	.56	.50	.54

Number of observations (alter-ego pair): 5185.

Table 17. Coauthor Analysis: Variable Descriptions

Variables	Descriptions
Dependent Variables	
Pub	Number of coauthored papers of the coauthor-ego pair.
C3.AVG	The average of three-year time window citation counts.
C5.AVG	The average of five-year time window citation counts.
IF.AVG	Average journal impact factors of the papers coauthored by the coauthor-ego pair.
C3.MAX	The maximum of three-year time window citation counts.
C5.MAX	The maximum of five-year time window citation counts.
IF.MAX	Maximum journal impact factors of the papers coauthored by the coauthor-ego pair.
AU.MIN	The minimum author numbers of papers coauthored by the coauthor-ego pair.
Age	Number of years since the first time that the coauthor-ego pair coauthored to 2007.
Independent Variables	
Close-Coauthor	Dummy: 1 if the coauthor matches to an alter, 0 otherwise.
Tie Strength	The tie strength factor extracted from the factor analysis using survey data.
Str.AVG	Ego level variable, average tie strength of the ego.
Str.D	The tie strength centered at the ego average, i.e, *Tie Strength – Str.AVG*
Alter Org	Type of organizations that the alter is affiliated to: same department, same university, other US universities, foreign universities, and others.
Alter Seniority	Seniority of the alter to the ego: senior, peer, and junior.
Alter Female	Gender of the alter: 1 if female and 0 if male.

The unit of analysis is coauthor-ego pair.

Publication and citation counts are based on 2005-2007 papers for coauthor-alter comparisons, and 2005-2010 papers for alter-alter comparisons. The last six independent variables are only available for coauthors that are matched to alters, and used for alter-alter comparisons.

Table 18. Close- vs. Other-Coauthor: Descriptive Statistics

		mean	sd	median	min	max	1	2	3	4	5	6	7
1	Close-Coauthor	0.09	0.29	0.00	0.00	1.00							
2	ln(Pub)	0.23	0.48	0.00	0.00	3.43	.17						
3	ln(C5.AVG+1)	2.99	1.54	2.83	0.00	6.83	-.11	-.05					
4	ln(IF.AVG+1)	1.56	0.69	1.50	0.03	3.67	-.07	-.02	.44				
5	ln(C5.MAX+1)	3.09	1.55	2.94	0.00	6.83	-.08	.09	.99	.44			
6	ln(IF.MAX+1)	1.61	0.72	1.50	0.03	3.67	-.04	.12	.43	.98	.44		
7	ln(AU.MIN)	2.62	1.71	2.08	1.10	7.61	-.17	-.20	.71	.20	.68	.16	
8	ln(Age)	0.82	0.73	0.69	0.00	3.33	.22	.34	-.25	-.08	-.20	-.03	-.37

Number of observations (coauthors): 24094
Number of groups (ego): 1157

Table 19. Close- vs. Other-Coauthor: LSDV Models

	1	2	3	4	5	6	7
	ln(Pub)	ln(C5.AVG+1)	ln(IF.AVG+1)	ln(C5.MAX+1)	ln(IF.MAX+1)	ln(AU.MIN)	ln(Age)
	LSDV	LSDV	LSDV	LSDV	LSDV	LSDV	LSDV
CloseCoauthor	0.31 *** (0.01)	0.09 *** (0.02)	0.04 *** (0.01)	0.23 *** (0.02)	0.1 *** (0.01)	-0.21 *** (0.01)	0.49 *** (0.02)
Ego Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R2 adj	0.22	0.71	0.61	0.69	0.58	0.92	0.3

Number of observations (coauthors): 24094
Number of groups (ego): 1157
* p < .10, ** p < .05, *** p < .01
The dependent variables are all *ln* transformed. The intercepts are fixed and not reported. The reference group is non-alter-coauthors.

Table 20. Close-Coauthor Analysis: Descriptive Statistics

		mean	sd	median	min	max	1	2	3	4	5	6	7	8	9
1	ln(Pub+1)	1.16	0.54	1.10	0.69	4.22									
2	ln(C3.AVG+1)	1.85	0.91	1.80	0.00	5.48	0.19								
3	ln(C5.AVG+1)	2.33	1.00	2.35	0.00	6.01	0.17	0.95							
4	ln(IF.AVG+1)	1.38	0.59	1.37	0.03	3.43	0.07	0.59	0.58						
5	ln(C3.MAX+1)	2.14	1.05	2.08	0.00	5.70	0.46	0.94	0.89	0.54					
6	ln(C5.MAX+1)	2.64	1.14	2.64	0.00	6.83	0.44	0.91	0.95	0.53	0.95				
7	ln(IF.MAX+1)	1.54	0.70	1.47	0.05	3.47	0.33	0.59	0.57	0.93	0.62	0.61			
8	Tie Strength	0.08	0.79	-0.14	-1.01	2.98	0.26	-0.07	-0.08	-0.08	0.01	0.00	0.00		
9	Str.AVG	0.08	0.48	0.02	-0.96	2.98	0.10	-0.11	-0.12	-0.12	-0.07	-0.08	-0.07	0.61	
10	Str.D	0.00	0.63	-0.01	-2.33	2.78	0.24	0.00	0.00	-0.01	0.07	0.07	0.06	0.79	0.00
11	Ego Field	BIOL 448, CHEM 517, CS 356, EAS 761, EE 310													
12	Ego Position	Assistant Prof 599, Associate Prof 644, Full Prof 1149													
13	Ego Female	Female 1103, Male 1289													
14	Ego White	Non-Hispanic White 2011, Others 381													
15	Alter Org	Same Department 656, Same University 379, Other US University 845, Foreign University 282, Others 230													
16	Alter Seniority	Senior to ego 887, Peer to ego 879, Junior to ego 626													
17	Alter Female	Female 626, Male 1766													

Number of observations (alters/close-coauthors): 2392

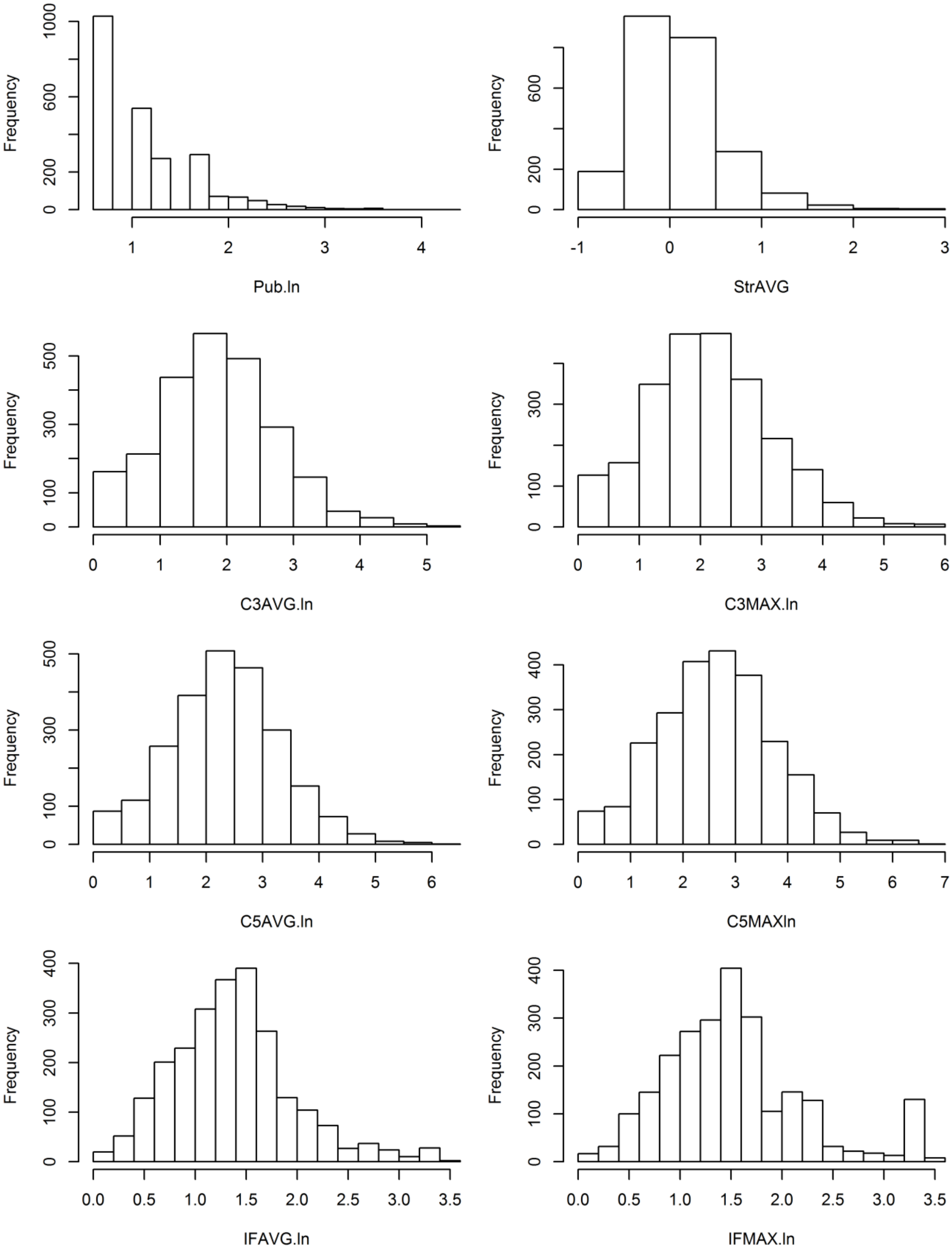


Figure 9. Close-Coauthor Analysis: Histograms

4.4.3. Survey-based Tie Strength Factor and Citations

The same dataset containing 2,392 close-coauthors of 820 egos is used to validate some of the findings based on bibliometric data reported in previous sections. Because this dataset only has information about the "closest" collaborators, I could not replicate all the previously reported studies using bibliometric data. Instead, I focus on the relationship between survey-based tie strength factor and publication/citation counts at the dyadic level.

I construct a hierarchical linear model based on the close-coauthor data. This model consists of two equations: First, within egos, I have the regression of Y (the dependent variable, i.e., publication and citation statistics) on *StrD*, which is a close-coauthor-level variable and centers *Str* (i.e., survey-based tie strength factor) at the ego average (i.e., $StrD_{ij} = Str_{ij} - StrAVG_{i\cdot}$). This approach makes regression results easier to interpret: The intercept for each ego represents the average level of *Y* for each ego, and the coefficient on *StrD* represents the within-ego tie strength effect.

The close-coauthor-level equation for close-coauthor *j* of ego *i* is

$$Y_{ij} = \alpha_{0i} + \alpha_{1i} \cdot StrD_{ij} + \varepsilon_{ij}$$

At the ego level, I allow both the intercept (α_{0i}) and the slope (α_{1i}) to depend on the average tie strength (*StrAVG*) of the egos:

$$\alpha_{0i} = \gamma_{00} + \gamma_{01} \cdot StrAVG_i + u_{0i}$$
$$\alpha_{1i} = \gamma_{10} + \gamma_{11} \cdot StrAVG_i + u_{1i}$$

Plugging the ego-level equations into the close-coauthor-level equation gives

$$Y_{ij} = \gamma_{00} + \gamma_{01} \cdot StrAVG_i + \mu_{0i} + (\gamma_{10} + \gamma_{11} \cdot StrAVG_i + \mu_{1i}) \cdot StrD_{ij} + \varepsilon_{ij}$$

Re-arranging terms gives

$$Y_{ij} = \gamma_{00} + \gamma_{01} \cdot StrAVG_i + \gamma_{10} \cdot StrD_{ij} + \gamma_{11} \cdot StrAVG_i \cdot StrD_{ij} + u_{0i} + u_{1i} \cdot StrD_{ij} + \varepsilon_{ij}$$

where the γ's are fixed effects, and the u's are random effect. Finally, rewriting the model in the previously used style gives

$$Y_{ij} = \beta_1 + \beta_2 \cdot StrAVG_i + \beta_3 \cdot StrD_{ij} + \beta_4 \cdot StrAVG_i \cdot StrD_{ij} + b_{i1} + b_{i2} \cdot StrD_{ij} + \varepsilon_{ij}$$

and the change is purely notational, using β's for fixed effects and b's for random effects. The variances and covariances of random effects are the same as specified in the METHOD section.

Therefore, the model is specified as

$$Y_{ij} = \beta_1 + \beta_2 \cdot StrAVG_i + \beta_3 \cdot StrD_{ij} + \beta_4 \cdot StrAVG_i \cdot StrD_{ij} + b_{i1} + b_{i2} \cdot StrD_{ij} + \varepsilon_{ij}$$

where

- Y_{ij} is the value of the dependent variable of close-coauthor *j* of ego *i*.
- $StrAVG_i$ is the average tie strength for the *i*-th ego, i.e., $StrAVG_i = \overline{Str_{i\cdot}}$ where Str_{ij} is the survey-based tie strength factor for close-coauthor *j* of ego *i*.
- $StrD_{ij}$ is the tie strength for the *j*-th close-coauthor of the *i*-th ego centered at the ego average, i.e., $StrD_{ij} = Str_{ij} - StrAVG_i$

In addition, the dependent variables are: number of publications (*Pub*), average five-year time window citation counts (*C5.AVG*), average three-year time window citation counts (*C3.AVG*), average journal impact factor (*IF.AVG*), maximum five-year time window citation counts (*C5.MAX*), maximum three-year time window citation counts (*C3.MAX*), and maximum journal impact factor (*IF.MAX*). *C3* and *IF* are also used as dependent variables because the 2005-2010 paper sample is used and papers after 2007 do not have a complete five-year period for accumulating citations. In addition, journal impact factor uses the 2007 version because the survey was conducted in 2007. All variables are listed and described in **Table 17**, descriptive statistics are reported in **Table 20**, and histograms of focal variables are plotted in **Figure 9**.

Two sets of models are fitted: linear mixed-effect models in which the dependent variables are *ln* transformed and negative binomial mixed-effect models in which the original integer values of the dependent variables are used. Estimation results are reported in **Table 21** – **Table 27**. In terms of publication counts, stronger ties have more publications, and this is true for both between- and within- ego comparisons.

Results of *C3*, *C5*, and *IF* converge. For between-ego comparison, egos with higher average tie strength have lower average and maximum citations. This finding is in line with previous finding of low creativity associated with high average tie strength at the egocentric network level. In addition, I argued for an inverted U-shaped relationship between network average tie strength and creativity when studying egocentric networks consisting of all coauthors, but argue for a negative linear effect here because the networks studied here consist of only the “closest” collaborators. Exploratory plots of citation statistics and ego average tie strength also suggest that a linear relationship is sufficient (**Figure 10**).

In addition, for within-ego comparison, there is no significant difference in average citations between strong and weak ties, but stronger ties do have higher maximum citations. This result might provide some evidence to support the cost-reduction hypothesis, that is, very strong ties indeed pursue more experimental research because of reduced transaction costs. This cost-reduction effect is not found from general coauthors but only here from the “closest” collaborators. The reason might be that experimental research really requires an extremely strong relation and possibly many other conditions, such as co-location (Catalini, 2012) and friendship. However, this result needs to be interpreted with great caution because I cannot control for the selection effect here.

At least, I can conclude here that the finding of low creativity associated with very strong tie-strength is also confirmed in survey data. There might be some evidence of a cost-reduction effect, but future research controlling the selection effect is needed for a more reliable conclusion.

4.4.4. Controlling for Contextual Factors

Another issue with bibliometric data is the lack of contextual information about the tie, so it is unclear whether the observed effects are independently caused by tie strength or actually caused by some omitted variables. Therefore, I control for a number of contextual factors to check if findings would be significantly changed. Specifically, I add a set of ego- and close-coauthor-level contextual variables to the model fitted in Section 4.4.3.

$$\begin{aligned} Y_{ij} = \beta_1 + \beta_2 \cdot StrAVG_i + \beta_3 \cdot StrD_{ij} + \beta_4 \cdot Field_i + \beta_5 \cdot Rank_i + \beta_6 \cdot Gender_i + \beta_7 \\ \cdot Race_i + \beta_8 \cdot Org_{ij} + \beta_9 \cdot Seniority_{ij} + \beta_{10} \cdot Gender_{ij} + \beta_{11} \cdot Rank_i \\ \cdot Seniority_{ij} + \beta_{12} \cdot Gender_i \cdot Gender_{ij} + b_{i1} + b_{i2} \cdot StrD_{ij} + \varepsilon_{ij} \end{aligned}$$

where

- $Field_i$ is the research field of the i-th ego: BIOL, CHEM, CS, EAS, and EE.
- $Rank_i$ is the academic rank of the i-th ego: Assistant, Associate, and Full Professor.
- $Gender_i$ is the gender of the i-th ego: female and male.
- $Race_i$ is the race of the i-th ego: non-Hispanic White and minorities.
- Org_{ij} is the type of institution that close-coauthor j of ego i is affiliated to: same department, same university, other US universities, foreign universities, and others.

- $Seniority_{ij}$ is the seniority relationship between close-coauthor j and ego i: senior to ego, peer of ego, and junior to ego.
- $Gender_{ij}$ is the gender of close-coauthor j of the i-th ego: female and male.
- $Rank_i \cdot Seniority_{ij}$ tests the interaction effect between ego rank and alter (i.e., close-coauthor) seniority, presumably the effect of alter seniority varies for egos at different academic ranks.
- $Gender_i \cdot Gender_{ij}$ tests the interaction effect between ego gender and alter gender, presumably collaborating with female have different effects for men and women.
- $StrAVG_i \cdot StrD_{ij}$ is removed from the model because its effect is insignificant. Interaction effects between $StrD_{ij}$ and all the new added contextual variables were also considered, allowing the effect of $StrD_{ij}$ to vary by different types of egos and alters. However, these interaction effects were insignificant and not included in the end.

Estimation results are reported in **Table 21** – **Table 27**. The size of the $StrAVG_i$ effect shrink after adding these control variables but remain significant, so $StrAVG_i$ is partially confounded with these contextual factors. On the other hand, the size of the $StrD_{ij}$ effects seems to have some increase after controlling for these contextual factors. In general, the conclusions would not change, so based on this analysis, I am more confident that the previous findings about tie strength effects using bibliometric data are reliable even though I did not control for many contextual factors.

Table 21. Close-Coauthor Analysis: PUB Models

	1	2	3	4
	ln(Pub)	ln(Pub)	Pub	Pub
	Linear Mixed Model	Linear Mixed Model	Negative-binomial Mixed Model	Negative-binomial Mixed Model
Str.AVG	0.18 *** (0.04)	0.21 *** (0.04)	0.19 *** (0.04)	0.22 *** (0.04)
Str.D	0.30 *** (0.03)	0.30 *** (0.03)	0.36 *** (0.03)	0.35 *** (0.03)
Str.AVG*Str.D	-0.05 (0.06)		-0.10 (0.06)	
Ego Level				
Field-CHEM		0.09 (0.06)		0.13 * (0.07)
Field-CS		-0.11 * (0.06)		-0.16 ** (0.08)
Field-EAS		-0.03 (0.05)		-0.04 (0.07)
Field-EE		0.13 ** (0.07)		0.18 ** (0.08)
Rank-Assoc		-0.17 ** (0.08)		-0.23 ** (0.09)
Rank-Full		-0.09 (0.07)		-0.08 (0.08)
Ego Female		0.01 (0.04)		0.03 (0.05)
Ego White		-0.06 (0.05)		-0.10 (0.06)
Alter Level				
Org-Same Uni		0.03 (0.05)		0.01 (0.06)
Org-US Univ		0.05 (0.04)		0.04 (0.05)
Org-ForeignU		0.08 (0.05)		0.04 (0.06)
Org-Others		0.13 ** (0.05)		0.12 * (0.06)
Alt Senior		-0.12 * (0.06)		-0.15 ** (0.08)
Alt Junior		-0.12 (0.11)		-0.12 (0.13)
Alt Female		0.04 (0.06)		0.05 (0.07)
Ego * Alter				
Senior*Assoc		0.15 * (0.09)		0.18 * (0.11)
Senior*Full		0.16 * (0.08)		0.25 ** (0.10)
Junior*Assoc		0.19 (0.13)		0.23 (0.16)
Junior*Full		0.21 * (0.11)		0.22 (0.14)
Ego Female * Alt Female		-0.18 ** (0.07)		-0.23 ** (0.09)
Log-likelihood	-2516	-2532	-4763	-4737

Number of observations (alter-ego pair): 2392

Number of groups (ego): 820

* p < .10, ** p < .05, *** p < .01

The dependent variables are *ln* transformed for linear mixed models (column 1&2), and use integer values for negative binomial mixed models (column 3&4). Random intercept and random effect of *Str.D* are included in all models but not reported. The reference group for ego field is BIOL, for ego rank is Assistant Professor, for alter org type is same department, for alter seniority is peer of ego.

Table 22. Close-Coauthor Analysis: C3.AVG Models

	1		2		3		4	
	ln(C3.AVG+1)		ln(C3.AVG+1)		C3.AVG		C3.AVG	
	Linear Mixed		Linear Mixed		Negative-		Negative-	
	Model		Model		binomial Mixed		binomial Mixed	
					Model		Model	
Str.AVG	-0.20 ***	(0.05)	-0.15 ***	(0.05)	-0.23 ***	(0.05)	-0.15 ***	(0.04)
Str.D	0.00	(0.03)	0.04	(0.02)	-0.02	(0.04)	0.01	(0.03)
Str.AVG*Str.D	0.02	(0.05)			-0.03	(0.07)		
Ego Level								
Field-CHEM			-0.14 *	(0.08)			-0.16 **	(0.07)
Field-CS			-0.84 ***	(0.08)			-0.68 ***	(0.08)
Field-EAS			-0.24 ***	(0.07)			-0.23 ***	(0.06)
Field-EE			-0.45 ***	(0.09)			-0.38 ***	(0.08)
Rank-Assoc			-0.17 *	(0.09)			-0.11	(0.09)
Rank-Full			-0.04	(0.08)			0.03	(0.08)
Ego Female			0.00	(0.05)			-0.02	(0.05)
Ego White			0.06	(0.07)			0.06	(0.06)
Alter Level								
Org-Same Uni			0.00	(0.05)			0.01	(0.05)
Org-US Univ			0.20 ***	(0.04)			0.20 ***	(0.04)
Org-ForeignU			0.06	(0.06)			0.05	(0.06)
Org-Others			0.15 **	(0.06)			0.19 ***	(0.06)
Alt Senior			0.01	(0.07)			0.05	(0.07)
Alt Junior			-0.07	(0.12)			-0.18	(0.12)
Alt Female			-0.12 *	(0.06)			-0.14 **	(0.06)
Ego * Alter								
Senior*Assoc			0.06	(0.10)			-0.02	(0.09)
Senior*Full			-0.02	(0.09)			-0.07	(0.09)
Junior*Assoc			0.11	(0.15)			0.15	(0.15)
Junior*Full			0.05	(0.13)			0.07	(0.13)
Ego Female * Alt Female			0.04	(0.08)			0.11	(0.08)
Log-likelihood	-2910		-2862		-7527		-7667	

Number of observations: 2392

Number of groups (egos): 820

* p < .10, ** p < .05, *** p < .01

The dependent variables are *ln* transformed for linear mixed models (column 1&2), and use integer values for negative binomial mixed models (column 3&4). Random intercept and random effect of *Str.D* are included in all models but not reported. The reference group for ego field is BIOL, for ego rank is Assistant Professor, for alter org type is same department, for alter seniority is peer of ego.

Table 23. Close-Coauthor Analysis: C5.AVG Models

	1	2	3	4
	ln(C5.AVG+1) Linear Mixed Model	ln(C5.AVG+1) Linear Mixed Model	C5.AVG Negative-binomial Mixed Model	C5.AVG Negative-binomial Mixed Model
Str.AVG	-0.25 *** (0.05)	-0.18 *** (0.05)	-0.31 *** (0.06)	-0.23 *** (0.06)
Str.D	-0.02 (0.03)	0.05 (0.03)	-0.03 (0.04)	0.03 (0.03)
Str.AVG*Str.D	0.06 (0.06)		0.02 (0.07)	
Ego Level				
Field-CHEM		-0.18 ** (0.08)		-0.20 ** (0.10)
Field-CS		-0.93 *** (0.09)		-1.02 *** (0.11)
Field-EAS		-0.25 *** (0.08)		-0.28 *** (0.09)
Field-EE		-0.54 *** (0.10)		-0.57 *** (0.11)
Rank-Assoc		-0.20 ** (0.10)		-0.21 * (0.12)
Rank-Full		-0.06 (0.09)		0.00 (0.10)
Ego Female		-0.01 (0.06)		-0.01 (0.07)
Ego White		0.03 (0.08)		0.08 (0.09)
Alter Level				
Org-Same Uni		0.03 (0.06)		0.07 (0.06)
Org-US Univ		0.26 *** (0.05)		0.29 *** (0.05)
Org-ForeignU		0.11 (0.06)		0.13 * (0.07)
Org-Others		0.21 *** (0.07)		0.27 *** (0.07)
Alt Senior		-0.01 (0.08)		0.03 (0.08)
Alt Junior		-0.10 (0.14)		-0.12 (0.15)
Alt Female		-0.10 (0.07)		-0.13 * (0.08)
Ego * Alter				
Senior*Assoc		0.11 (0.11)		0.09 (0.12)
Senior*Full		0.01 (0.10)		-0.05 (0.11)
Junior*Assoc		0.13 (0.16)		0.13 (0.18)
Junior*Full		0.13 (0.15)		0.09 (0.16)
Ego Female * Alt Female		0.04 (0.09)		0.10 (0.10)
Log-likelihood	-3152	-3098	-8692	-8609

Number of observations: 2392

Number of groups (egos): 820

* p < .10, ** p < .05, *** p < .01

The dependent variables are *ln* transformed for linear mixed models (column 1&2), and use integer values for negative binomial mixed models (column 3&4). Random intercept and random effect of *Str.D* are included in all models but not reported. The reference group for ego field is BIOL, for ego rank is Assistant Professor, for alter org type is same department, for alter seniority is peer of ego.

Table 24. Close-Coauthor Analysis: IF.AVG Models

	1	2	3	4
	ln(IF.AVG+1) Linear Mixed Model	ln(IF.AVG+1) Linear Mixed Model	IF.AVG Negative-binomial Mixed Model	IF.AVG Negative-binomial Mixed Model
Str.AVG	-0.13 *** (0.03)	-0.07 ** (0.03)	-0.22 *** (0.06)	-0.13 ** (0.05)
Str.D	-0.01 (0.02)	0.01 (0.01)	-0.02 (0.03)	0.01 (0.03)
Str.AVG*Str.D	0.01 (0.03)		0.01 (0.06)	
Ego Level				
Field-CHEM		-0.06 (0.05)		-0.11 (0.08)
Field-CS		-0.85 *** (0.05)		-1.58 *** (0.10)
Field-EAS		-0.33 *** (0.04)		-0.53 *** (0.08)
Field-EE		-0.67 *** (0.05)		-1.19 *** (0.10)
Rank-Assoc		-0.18 *** (0.05)		-0.30 *** (0.10)
Rank-Full		-0.08 * (0.05)		-0.12 (0.09)
Ego Female		0.07 ** (0.03)		0.12 ** (0.06)
Ego White		0.02 (0.04)		0.09 (0.08)
Alter Level				
Org-Same Uni		0.06 * (0.03)		0.12 ** (0.06)
Org-US Univ		0.06 *** (0.02)		0.10 ** (0.05)
Org-ForeignU		0.01 (0.03)		0.03 (0.06)
Org-Others		0.11 *** (0.03)		0.18 *** (0.07)
Alt Senior		-0.10 *** (0.04)		-0.15 ** (0.07)
Alt Junior		-0.21 *** (0.07)		-0.41 *** (0.14)
Alt Female		-0.01 (0.04)		-0.02 (0.07)
Ego * Alter				
Senior*Assoc		0.16 *** (0.05)		0.26 ** (0.10)
Senior*Full		0.03 (0.05)		-0.03 (0.10)
Junior*Assoc		0.24 *** (0.08)		0.46 *** (0.16)
Junior*Full		0.16 ** (0.07)		0.29 ** (0.15)
Ego Female * Alt Female		0.00 (0.05)		0.04 (0.09)
Log-likelihood	-1668	-1511	-5089	-4895

Number of observations: 2392

Number of groups (egos): 820

* p < .10, ** p < .05, *** p < .01

The dependent variables are *ln* transformed for linear mixed models (column 1&2), and use integer values for negative binomial mixed models (column 3&4). Random intercept and random effect of *Str.D* are included in all models but not reported. The reference group for ego field is BIOL, for ego rank is Assistant Professor, for alter org type is same department, for alter seniority is peer of ego.

Table 25. Close-Coauthor Analysis: C3.MAX Models

	1		2		3		4	
	ln(C3.MAX+1) Linear Mixed Model		ln(C3.MAX+1) Linear Mixed Model		C3.MAX Negative-binomial Mixed Model		C3.MAX Negative-binomial Mixed Model	
Str.AVG	-0.13 **	(0.06)	-0.07	(0.05)	-0.16 *	(0.07)	-0.09	(0.07)
Str.D	0.14 ***	(0.04)	0.17 ***	(0.03)	0.19 ***	(0.04)	0.20 ***	(0.03)
Str.AVG*Str.D	-0.03	(0.07)			-0.08	(0.08)		
Ego Level								
Field-CHEM			-0.12	(0.09)			-0.12	(0.11)
Field-CS			-0.87 ***	(0.10)			-0.97 ***	(0.12)
Field-EAS			-0.27 ***	(0.08)			-0.29 ***	(0.10)
Field-EE			-0.38 ***	(0.10)			-0.35 ***	(0.12)
Rank-Assoc			-0.22 **	(0.11)			-0.30 **	(0.12)
Rank-Full			-0.04	(0.10)			-0.03	(0.11)
Ego Female			0.03	(0.06)			0.03	(0.07)
Ego White			0.03	(0.08)			0.06	(0.10)
Alter Level								
Org-Same Uni			0.00	(0.06)			0.00	(0.07)
Org-US Univ			0.22 ***	(0.05)			0.23 ***	(0.06)
Org-ForeignU			0.11	(0.07)			0.14 *	(0.08)
Org-Others			0.24 ***	(0.07)			0.28 ***	(0.08)
Alt Senior			-0.01	(0.08)			0.00	(0.09)
Alt Junior			-0.09	(0.14)			-0.15	(0.16)
Alt Female			-0.09	(0.07)			-0.10	(0.08)
Ego * Alter								
Senior*Assoc			0.10	(0.11)			0.13	(0.13)
Senior*Full			0.03	(0.11)			0.02	(0.12)
Junior*Assoc			0.18	(0.17)			0.28	(0.19)
Junior*Full			0.12	(0.15)			0.17	(0.17)
Ego Female * Alt Female			-0.08	(0.10)			-0.06	(0.11)
Log-likelihood	-3265		-3227		-8358		-8294	

Number of observations: 2392

Number of groups (egos): 820

* p < .10, ** p < .05, *** p < .01

The dependent variables are *ln* transformed for linear mixed models (column 1&2), and use integer values for negative binomial mixed models (column 3&4). Random intercept and random effect of *Str.D* are included in all models but not reported. The reference group for ego field is BIOL, for ego rank is Assistant Professor, for alter org type is same department, for alter seniority is peer of ego.

Table 26. Close-Coauthor Analysis: C5.MAX Models

	1	2	3	4
	ln(C5.MAX+1)	ln(C5.MAX+1)	C5.MAX	C5.MAX
	Linear Mixed Model	Linear Mixed Model	Negative-binomial Mixed Model	Negative-binomial Mixed Model
Str.AVG	-0.19 *** (0.06)	-0.10 * (0.06)	-0.20 ** (0.07)	-0.12 * (0.07)
Str.D	0.13 *** (0.04)	0.19 *** (0.04)	0.18 *** (0.04)	0.22 *** (0.03)
Str.AVG*Str.D	0.00 (0.08)		-0.06 (0.08)	
Ego Level				
Field-CHEM		-0.17 * (0.10)		-0.17 (0.11)
Field-CS		-0.95 *** (0.10)		-0.98 *** (0.12)
Field-EAS		-0.28 *** (0.09)		-0.29 *** (0.10)
Field-EE		-0.46 *** (0.11)		-0.42 *** (0.12)
Rank-Assoc		-0.29 ** (0.12)		-0.35 *** (0.13)
Rank-Full		-0.10 (0.10)		-0.07 (0.11)
Ego Female		0.02 (0.06)		0.02 (0.07)
Ego White		0.01 (0.09)		0.04 (0.10)
Alter Level				
Org-Same Uni		0.03 (0.07)		0.05 (0.07)
Org-US Univ		0.30 *** (0.05)		0.29 *** (0.06)
Org-ForeignU		0.16 ** (0.07)		0.18 ** (0.08)
Org-Others		0.29 *** (0.08)		0.33 *** (0.08)
Alt Senior		-0.07 (0.09)		-0.02 (0.09)
Alt Junior		-0.14 (0.15)		-0.16 (0.16)
Alt Female		-0.06 (0.08)		-0.07 (0.08)
Ego * Alter				
Senior*Assoc		0.18 (0.12)		0.18 (0.13)
Senior*Full		0.10 (0.12)		0.06 (0.12)
Junior*Assoc		0.24 (0.19)		0.27 (0.20)
Junior*Full		0.24 (0.17)		0.22 (0.18)
Ego Female * Alt Female		-0.09 (0.11)		-0.07 (0.11)
Log-likelihood	-3467	-3422	-9746	-9678

Number of observations: 2392

Number of groups (egos): 820

* p < .10, ** p < .05, *** p < .01

The dependent variables are *ln* transformed for linear mixed models (column 1&2), and use integer values for negative binomial mixed models (column 3&4). Random intercept and random effect of *Str.D* are included in all models but not reported. The reference group for ego field is BIOL, for ego rank is Assistant Professor, for alter org type is same department, for alter seniority is peer of ego.

Table 27. Close-Coauthor Analysis: IF.MAX Models

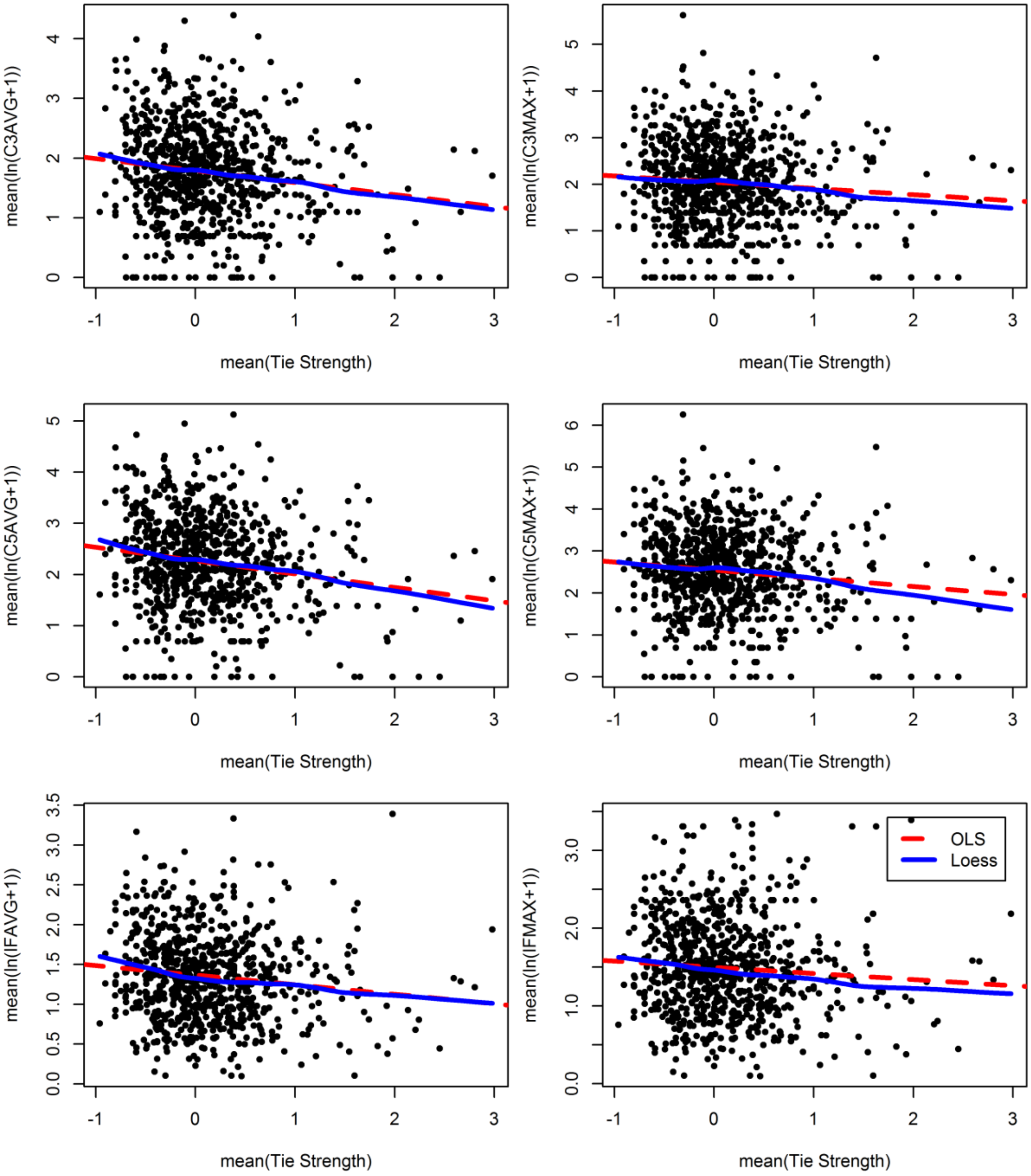
	1		2		3		4	
	ln(IF.MAX+1) Linear Mixed Model		ln(IF.MAX+1) Linear Mixed Model		IF.MAX Negative-binomial Mixed Model		IF.MAX Negative-binomial Mixed Model	
Str.AVG	-0.08 **	(0.04)	-0.02	(0.03)	-0.14 *	(0.06)	-0.03	(0.06)
Str.D	0.08 ***	(0.02)	0.08 ***	(0.02)	0.11 **	(0.04)	0.14 ***	(0.03)
Str.AVG*Str.D	-0.04	(0.04)			-0.03	(0.07)		
Ego Level								
Field-CHEM			-0.05	(0.06)			-0.08	(0.09)
Field-CS			-0.89 ***	(0.06)			-1.49 ***	(0.11)
Field-EAS			-0.37 ***	(0.05)			-0.53 ***	(0.09)
Field-EE			-0.68 ***	(0.06)			-1.07 ***	(0.11)
Rank-Assoc			-0.23 ***	(0.07)			-0.39 ***	(0.11)
Rank-Full			-0.11 *	(0.06)			-0.14	(0.10)
Ego Female			0.09 **	(0.04)			0.16 **	(0.06)
Ego White			0.01	(0.05)			0.07	(0.09)
Alter Level								
Org-Same Uni			0.06 *	(0.04)			0.11 *	(0.06)
Org-US Univ			0.09 ***	(0.03)			0.14 ***	(0.05)
Org-ForeignU			0.05	(0.04)			0.09	(0.07)
Org-Others			0.15 ***	(0.04)			0.23 ***	(0.07)
Alt Senior			-0.13 ***	(0.05)			-0.18 **	(0.08)
Alt Junior			-0.30 ***	(0.09)			-0.57 ***	(0.15)
Alt Female			0.00	(0.04)			0.00	(0.07)
Ego * Alter								
Senior*Assoc			0.17 **	(0.07)			0.26 **	(0.11)
Senior*Full			0.06	(0.07)			0.05	(0.11)
Junior*Assoc			0.38 ***	(0.10)			0.72 ***	(0.18)
Junior*Full			0.26 ***	(0.09)			0.49 ***	(0.16)
Ego Female * Alt Female			-0.05	(0.06)			-0.07	(0.10)
Log-likelihood	-2157		-2030		-5855		-5694	

Number of observations: 2392

Number of groups (egos): 820

* p < .10, ** p < .05, *** p < .01

The dependent variables are *ln* transformed for linear mixed models (column 1&2), and use integer values for negative binomial mixed models (column 3&4). Random intercept and random effect of *Str.D* are included in all models but not reported. The reference group for ego field is BIOL, for ego rank is Assistant Professor, for alter org type is same department, for alter seniority is peer of ego.



One point is one ego. The observations of dyads are grouped by egos. The X-axis the ego average tie strength, and the Y-axes are average citation statistics across dyads.

Figure 10. Close-Coauthor Analysis: Between-Ego Comparison

4.4.5. Summary

Based on this replicate study, I conclude that

- The bibliometric measure of tie strength concurs with the survey-based tie strength factor, and is therefore a valid measure.
- The finding of low creativity associated with very high network average tie strength is shared by both bibliometric data analysis and survey data analysis.
- There might be some evidence of cost-reduction effects, which were rejected by bibliometric data, but more research is needed for a reliable conclusion, particularly, the selection effect needs to be controlled.
- The lack of contextual variables is unlikely to cause severe problems that may overthrow previous findings based on bibliometric data.

4.5. The Knowledge Diffusion Hypothesis

Hypothesis 6: A paper gets more citations if the author has more collaborators, after controlling for prestige effect and intellectual relevance.

This section tests the knowledge diffusion hypothesis and studies whether an ego’s current collaboration network raises the awareness of his previously published papers. These papers are not products of the current collaboration network and have already been published, so the creativeness of these papers is not affected by the current networks. This design allows me to control for paper heterogeneities and focus on the knowledge diffusion effect.

4.5.1. Model Specification

The equation for citations of the *i*-th paper in its *t*-th year after publication is

$$Cite_{it} = \varphi_i + \lambda_t + \beta_1 \cdot Coll_{it} + \beta_2 \cdot Pres_{it} + \beta_3 \cdot Intl_{it} + \varepsilon_{it}$$

where

- $Cite_{it}$ is the number of times that the *i*-th paper is cited in the *t*-th year after it was published. *ln* transformed.
- φ_i is the paper fixed effect and controls for paper heterogeneities. A set of paper dummies are used for model fitting.
- λ_t is the age fixed effect. A set of age (years after publication) are used.
- $Coll_{it}$ is the number of collaborators that the ego of the *i*-th paper had in the preceding three years. *ln* transformed.
- $Pres_{it}$ is the prestige of the ego of the *i*-th paper.
- $Intl_{it}$ is the intellectual relevance of the *i*-th paper in this year.

Further descriptions of the incorporated variables are presented in **Table 28**.

Table 28. Knowledge Diffusion: Variable Descriptions

Variables	Descriptions
Dependent Variables	
Cite	Number of times that the paper is cited in its *t*-th year after publication.
I(Cite>0)	Dummy: 1 if the paper is cited in year *t*, 0 if not cited.
Independent Variables	
Coll	Number of collaborators of the ego in the preceding three years, *ln* transformed.
Prestige A	Total number of citations in the preceding three years received by all papers of the ego, *ln* transformed.
Prestige B	Total number of citations in the preceding three years received by the ego's papers excluding the focal solo-authored paper, *ln* transformed.
Prestige C	Total number of citations in the preceding three years received by the ego's papers excluding the focal solo-authored paper and other papers citing or cited by the focal solo-authored paper, *ln* transformed.
Relevance	Average citation of the reference papers, and these reference papers are those (a) published within a three-year interval before and after the publication year of the focal solo-authored paper, AND [(b1) cited by this focal solo-authored paper OR (b2) citing this focal paper OR (b3) cocited with this focal paper within three years after this focal paper was published] , *ln* transformed.

The unit of analysis is paper-year.

Prestige. The prestige of an ego is measured as the total number of citations he received in the preceding three years. If there is no prestige effect, then the ego's further prestige development should not affect the citations of this focal paper, since the focal paper has already been published. However, there might be some intellectual connections between this focal paper and further research conducted by this ego, and this connection would lead to correlations between citations of this focal paper and the ego's other papers. People citing the ego's other papers might also be interested in this focal paper which is intellectually related, and people may even search through reference lists of the ego's other papers and discover this focal paper, if it is in the reference lists. Therefore, three different measures are constructed to gradually eliminate the intellectual connections. They are: total number of citations received in the preceding three years by (A) all the papers published by this ego, (B) papers other than this focal paper, and (C) papers excluding this focal paper and other papers citing or cited by this focal paper. These measures are natural logarithm transformed.

Intellectual Relevance. To measure the intellectual relevance of the focal paper, I firstly identify a set of reference papers that are intellectually close to this paper, and then count the average citations of these papers in the same year as the dependent variable. The first criterion for reference papers pertains to year of publication: (a) only papers published within a three-year range before or after this focal paper are considered. The second criterion pertains to intellectual connections: reference papers are those: ($b1$) cited by this focal paper, or ($b2$) citing this focal paper, or ($b3$) cocited with this focal papers within three years after the focal papers was published. In other words, the criteria for reference papers are: $a\ and\ (b1\ or\ b2\ or\ b3)$. This measure is also natural logarithm transformed.

4.5.2. Sample Restriction

Only solo-authored papers are used for the analysis. The diffusion of coauthored papers will be affected by several egocentric networks collectively, making it difficult to estimate collaborator-marketing and prestige effects. Therefore, only solo-authored papers are analyzed. One potential issue is that solo-authored papers might be abnormal papers nowadays because of the increased dominance of team-production. However, the research question here is not about the production of papers, but about the diffusion of papers after controlling for paper heterogeneities. The production process might be very different between solo-authored papers and other "normal" papers, but the effects of collaborator-marketing, prestige, and intellectual-relevance on diffusion might not be so different. Therefore, findings from these solo-authored papers should be generalizable.

The unit of analysis is paper-year. I take a three-year time window to identify reference papers, so observations of the first three years for each paper are excluded. In this case, I am actually analyzing the declining segment of the citation life-cycle, while the early stage of citation maturing/increasing is left out (**Figure 11**). This is one big limitation of this analysis.

In total, I have 14,457 observations of 1,279 papers for analysis.

4.5.3. Findings

Descriptive statistics are reported in **Table 29**, and histograms in **Figure 12**. First, LSDV models are fitted using the ln of citations as the dependent variable (**Table 30** column 8-14). However, about half of the observations have 0 citations (**Figure 12** plot 1), so logistic mixed-effect models are fitted, in which the dependent variable is a dummy: 1 if cited and 0 otherwise. Since the LSDV strategy for controlling paper heterogeneities only works for OLS regressions, I make another modification for logistic models: replace the paper fixed effect φ_i by a random intercept (i.e., paper random effect). Regression results are reported in **Table 30** column 1-7.

There are no significant collaborator-marketing effects after controlling for prestige and intellectual-relevance effects, while significant prestige and intellectual-relevance effects are confirmed. In addition, after adding intellectual-relevance to the model, the size of the coefficients on prestige halves, which is in line with previous discussion that prestige and intellectual-relevance effects are closely related. In addition, comparing the size of the coefficients on three different measures of prestige, from A, B, to C, the size of coefficients also decreases. This is as expected, from A, B, to C, the intellectual connection between the focal paper and author prestige is gradually eliminated, so shrinks the estimated prestige effect.

In addition, this operationalization gives a relatively conservative estimation of the prestige effect, some of which might be absorbed by the fixed or random paper effect. Therefore, pooled-OLS models are fitted, that is, removing paper dummies but controlling for the research field. Results are reported in **Table 31**. The estimated prestige effects are much bigger than results in **Table 30**. However, this pooled-OLS may have serious endogeneity problems. Therefore, the true prestige effect should be somewhere between these two different estimations, in other words, the true prestige effect is larger than estimated in **Table 29**.

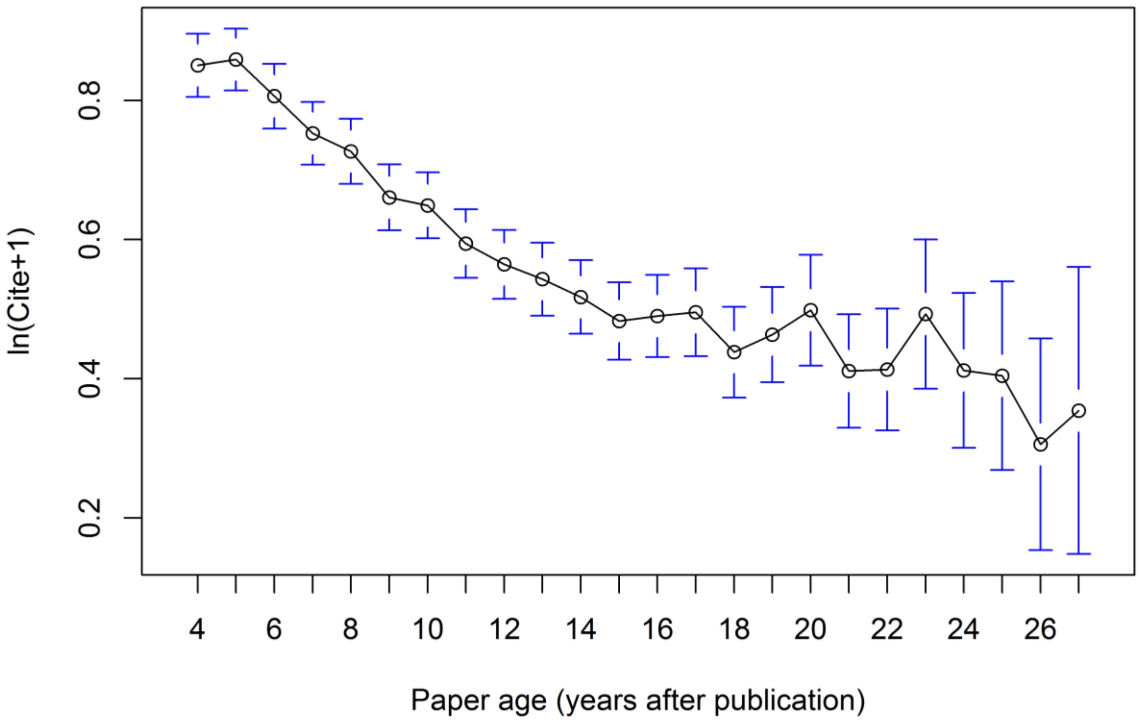


Figure 11. Knowledge Diffusion: Citation Ageing Trend

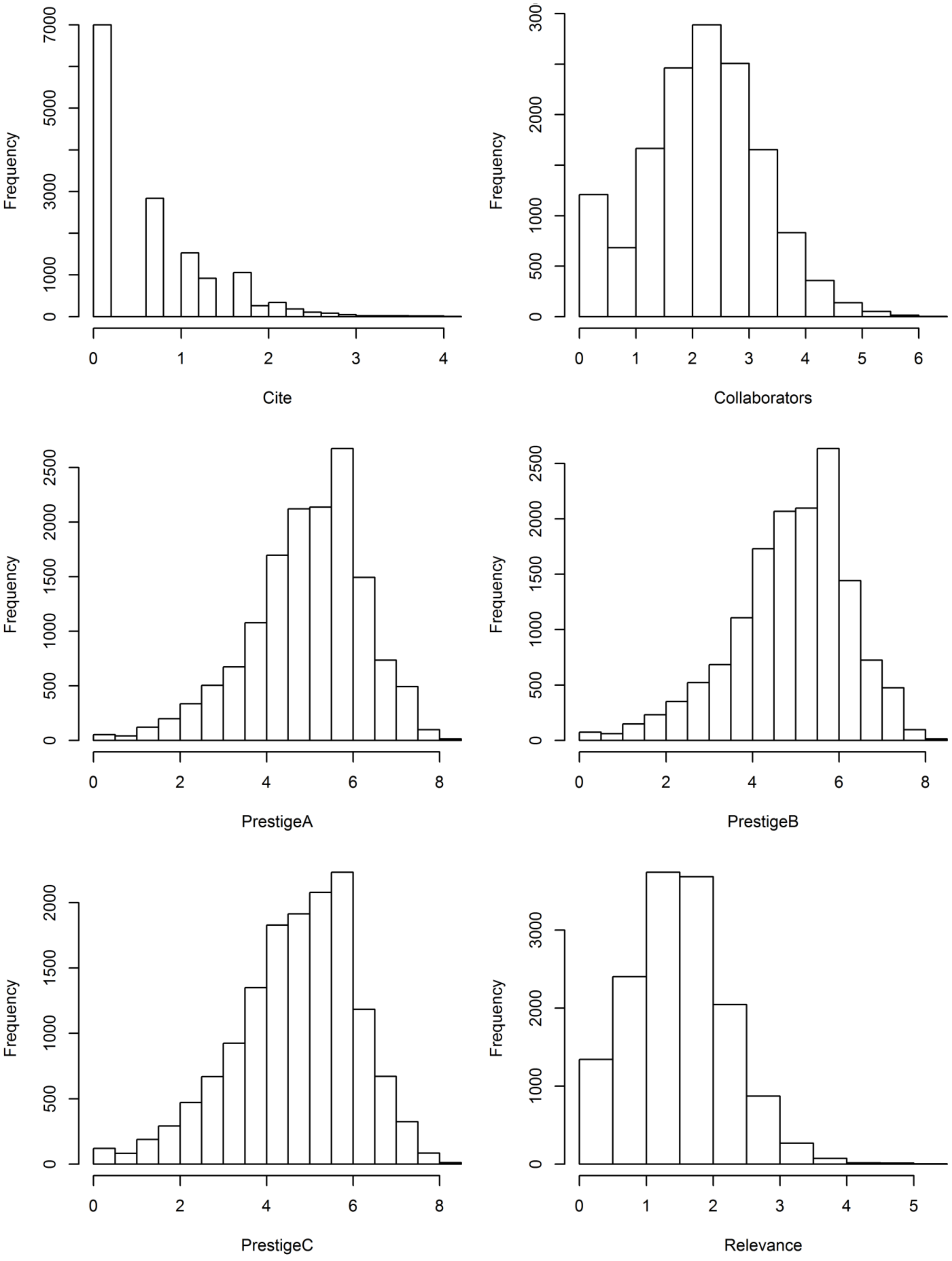


Figure 12. Knowledge Diffusion: Histograms

Table 29. Knowledge Diffusion: Descriptive Statistics

		n	mean	sd	median	min	max	1	2	3	4	5	6
1	ln(Cite+1)	14457	0.64	0.76	0.69	0	4.11						
2	I(Cite>0)	14457	0.52	0.50	1.00	0	1.00	.82					
3	Coll	14457	2.17	1.09	2.20	0	6.04	.12	.08				
4	Prestige A	14457	4.95	1.31	5.09	0	8.43	.23	.17	.68			
5	Prestige B	14457	4.90	1.35	5.05	0	8.43	.21	.15	.68	1.0		
6	Prestige C	14419	4.68	1.42	4.84	0	8.43	.12	.07	.67	.96	.97	
7	Relevance	14457	1.48	0.75	1.47	0	5.20	.30	.27	.18	.31	.30	.23

Unit of analysis: paper-year

Table 30. Knowledge Diffusion: Models A

	1	2	3	4	5	6	7
	I(Cite>0)	I(Cite>0)	I(Cite>0)	I(Cite>0)	I(Cite>0)	I(Cite>0)	I(Cite>0)
	Logistic Mixed Model	Logistic Mixed Model	Logistic Mixed Model	Logistic Mixed Model	Logistic Mixed Model	Logistic Mixed Model	Logistic Mixed Model
Coll	0.16 *** (0.03)	-0.03 (0.04)	-0.02 (0.04)	-0.01 (0.04)	0.00 (0.04)	0.04 (0.04)	0.04 (0.04)
Prestige A		0.50 *** (0.05)	0.30 *** (0.05)				
Prestige B				0.43 *** (0.04)	0.24 *** (0.05)		
Prestige C						0.28 *** (0.04)	0.12 *** (0.04)
Relevance			1.08 *** (0.08)		1.11 *** (0.08)		1.17 *** (0.08)
Paper dummies	No	No	No	No	No	No	No
Age dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Log-likelihood	-7165	-7109	-7009	-7120	-7015	-7118	-7001
Observations	14457	14457	14457	14457	14457	14419	14419
Groups	1279	1279	1279	1279	1279	1278	1278

	8	9	10	11	12	13	14
	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)
	LSDV	LSDV	LSDV	LSDV	LSDV	LSDV	LSDV
Coll	0.00 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	0.00 (0.01)	0.00 (0.01)
Prestige A		0.06 *** (0.01)	0.03 *** (0.01)				
Prestige B				0.04 *** (0.01)	0.02 * (0.01)		
Prestige C						0.02 ** (0.01)	0.00 (0.01)
Relevance			0.27 *** (0.02)		0.28 *** (0.02)		0.28 *** (0.02)
Paper dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Log-likelihood	0.68	0.68	0.69	0.68	0.69	0.68	0.69
Observations	14457	14457	14457	14457	14457	14419	14419
Groups	1279	1279	1279	1279	1279	1278	1278

* p < .10, ** p < .05, *** p < .01. The dependent variables are *ln* transformed for LSDV models (column 8-14), and use dummy (cited or not) for logistic mixed models (column 1-7). Intercepts of LSDV models are fixed and not reported, and intercepts of logistic

mixed models are random and not reported. Models using Prestige C (column 6,7,13,& 14) have fewer observations because some observations have missing *Prestige C* values.

Table 31. Knowledge Diffusion: Models B

	1	2	3	4	5	6	7
	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)	ln(Cite+1)
	pooled OLS	pooled OLS	pooled OLS	pooled OLS	pooled OLS	pooled OLS	pooled OLS
Coll	0.08 *** (0.01)	-0.08 *** (0.01)	-0.07 *** (0.01)	-0.06 *** (0.01)	-0.04 *** (0.01)	0.01 (0.01)	0.02 ** (0.01)
Prestige A		0.21 *** (0.01)	0.17 *** (0.01)				
Prestige B				0.18 *** (0.01)	0.13 *** (0.01)		
Prestige C						0.09 *** (0.01)	0.05 *** (0.01)
Relevance			0.17 *** (0.01)		0.18 *** (0.01)		0.22 *** (0.01)
Paper dummies	No	No	No	No	No	No	No
Age dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field-CHEM	-0.15 *** (0.02)	-0.19 *** (0.02)	-0.16 *** (0.02)	-0.19 *** (0.02)	-0.16 *** (0.02)	-0.18 *** (0.02)	-0.14 *** (0.02)
Field-CS	-0.21 *** (0.02)	-0.02 (0.02)	0.00 (0.02)	-0.05 ** (0.02)	-0.02 (0.02)	-0.14 *** (0.02)	-0.09 *** (0.02)
Field-EAS	0.12 *** (0.02)	0.19 *** (0.02)	0.20 *** (0.02)	0.18 *** (0.02)	0.19 *** (0.02)	0.15 *** (0.02)	0.17 *** (0.02)
Field-EE	-0.26 *** (0.02)	-0.18 *** (0.02)	-0.12 *** (0.02)	-0.19 *** (0.02)	-0.13 *** (0.02)	-0.24 *** (0.02)	-0.15 *** (0.02)
R2 adj	0.09	0.15	0.17	0.13	0.16	0.10	0.14
Observations	14457	14457	14457	14457	14457	14419	14419
Groups	1279	1279	1279	1279	1279	1278	1278

An observation is a paper-year

A group is a paper

* p < .10, ** p < .05, *** p < .01.

The dependent variables are *ln* transformed. Intercepts are fixed and not reported. Models using Prestige C (column 6 & 7) have fewer observations because some observations have missing *Prestige C* values. The reference group for field is BIOL.

Table 32. Hypothesis Testing Summary

	Hypothesis	?	Operationalization
1a	Knowledge spillover (new→ rep)	Y	• Section 4.1 • Data: panel data of ego's citations and publications by year. • Dependent variable: citations of rep/new papers • Independent variable: number of new/rep papers
1b	Knowledge spillover (rep→ new)	N	
2	Curvilinear effect of tie strength	Y	• Section 4.2 • Data: cross-sectional data of ego's coauthorship networks and citations • Dependent variable: citations • Independent variable: tie.strength, tie.strength^2, skew, skew* tie.strength, skew*tie.strength^2
3	Effect of skewness	Y	
4	Moderating effect of skewness	Y	
5a	Creativity-decline effect	Y	• Section 4.3 • Design 1: paired test on citations of rep and new papers • Design 2: quantile regression to model citation distributions of rep and new papers
5b	Cost-reduction effect	N	
6	Collaborator-marketing effect	N	• Section 4.5 • Data: panel data of solo-authored papers' citation history • Dependent variable: citations • Independent variable: coauthors, prestige, and intellectual relevance

CHAPTER 5

CONCLUSION

5.1. Intellectual Structure

The motivation of this dissertation was primarily to understand the relationship between collaboration networks and scientific creativity. As a first step, I integrated cognitive- and social- psychology literature about small groups and sociology literature about social networks. A theoretical framework emerges from this integration, in which the creative-process perspective provides a micro-foundation to understand causes of creativity and bridges the structural characteristics of the network and observed creativity produced from this network.

The unit of analysis in this intellectual quest is an egocentric network, while the prevailing model in the literature is analysis at the team level, so does the egocentric-network-level analysis make any sense at all? I argued that because of the fluidness of teams in scientific collaboration and significant knowledge-spillover between teams, bounding the study within an arbitrary team may miss a considerable amount of external transactions that are important to the creative product. Therefore, it is legitimate to adopt a holistic egocentric network perspective and position the knowledge production process in a network of connected collaborators instead of in a single and closed team. This knowledge-spillover thesis is empirically tested using panel data of ego's publications and citations. I found significant knowledge spillover from collaborations with new collaborators to collaborations with other repeated collaborators.

After justifying the legitimateness of this egocentric-network approach, I started to investigate effects of tie strength and tie strength skewness. How does tie strength affect creative process and creative outcomes? I argued that weak ties are beneficial to idea generation but not idea convergence and implementation, while strong ties are the opposite. Therefore, there is an inverted U-shaped relationship between tie strength and creativity. Furthermore, I investigated the configurations of weak and strong ties in a network, instead of taking a simple dichotomy between them. Specifically, I argued that a more skewed network is more creative when the network average tie strength is high, because of its mixture of strong and weak ties. In addition, tie strength skewness moderates the effect of network average tie strength, that is, a more skewed network is less sensitive to changes in network average tie strength. These three effects were supported by cross-sectional data of egos' citations and coauthorship networks.

One important thesis in my study of tie strength effect is that products of strong-tie-collaborations have low creativeness because of the path dependency or low creative capacity associated with strong-tie-collaborations. This explanation is from a psychologist's perspective. However, economists may provide an alternative explanation, that is, reduced transaction costs. Because strong-tie-collaborations have lower costs, it is "profitable" for scientists to do trivial or experimental research with strong-tie-collaborators. These two competing hypotheses were tested in two different ways: (1) paired tests (at the ego level) on citations of new- and repeated- collaborations, and (2) quantile regressions (at the paper level) to model citation-distribution differences between new- and repeated- collaborations. The creativity-reduction but not the cost-reduction hypothesis was supported.

My analyses are primarily based on bibliometric data, which have a variety of limitations, specifically, is the number of coauthored papers between two collaborators a good measure of tie strength between these two collaborators, and would my findings still hold if I use a "better" tie strength measure or control for many contextual factors that are not available in the bibliometric data? To address these concerns, I did a replicate study using survey data. I did a factor analysis to extract the tie strength factor from a variety of survey measures. Subsequently, I compared publication and citation counts (1) between coauthors nominated as the “closest” collaborators and coauthors not nominated and (2) within the group of the “closest” collaborators for whom the survey-based tie strength factor and other contextual information are available. It seems that number of coauthored papers is a valid measure of tie strength, and I also validated some of my findings from bibliometric data.

Finally, I changed the context from knowledge creation to diffusion, asking the question: will the current collaboration network help to raise the awareness (and therefore citations) of a previously published paper? Here the network of interest is not the venue where the paper is produced, but a channel to diffuse knowledge. However, parsing out the collaborator-marketing effect is very tricky because of two important confounding effects: the prestige effect and the intellectual-relevance effect. Using panel data about the citation history of solo-authored papers, I found no evidence of collaborator-marketing effect but significant prestige effect even after controlling for the intellectual relevance.

5.2. Findings

Hypotheses and corresponding operationalization and results are summarized in **Table 32**. Major findings of this dissertation are:

- The number of new-collaborations has a positive effect on the average but not the maximum citations of repeated-collaborations, while the number of repeated-collaborations has no effect on the average or maximum citations of new-collaborations. This suggests knowledge spillover from new-collaborations to repeated-collaborations.
- There is an inverted U-shaped relationship between network average tie strength and creativity. Furthermore, when the average tie strength is high, a more skewed network has higher creativity. In addition, a more skewed network is less sensitive to changes in average tie strength. The sizes of both the initial positive

and the later negative effect of increases in tie strength are smaller in a more skewed network.

- Compared with the citation distributions of new-collaborations, repeated-collaborations have more citations at low-percentiles but fewer citations at high-percentiles. This finding suggests significant selection effect (i.e., abandoning unpleasant collaborations), supports the creativity-decline hypothesis, and rejects the cost-reduction hypothesis.
- Compared with other coauthors, those nominated by egos as the "closest" collaborators have significantly more papers. In addition, among these "closest" collaborators, those with stronger tie-strength (survey-based measure) have more papers. Therefore, the number of coauthored papers seems to be good proxy for tie strength.
- The study of these "closest" collaborators found that egos with higher average tie strength (survey-based) have lower average and maximum citations at the ego level. However, within-ego comparison found that stronger ties have no different average but higher maximum citations. This finding may provide some evidence of the cost-reduction effect. However, further studies are needed to properly control for the selection effect and also explore why this effect is only found from the "closest" collaborators.
- Number of coauthors does not lead to higher citations of previously published papers, but preceding citations received by the author's papers that have no intellectual connections with the focal paper have significant positive effects on citations of this focal paper, even after controlling for citations of a set of reference papers that are very similar to this focal paper. Therefore, there is evidence of the prestige effect but not the collaborator-marketing effect.

5.3. Theoretical Contributions

This dissertation makes the following three theoretical contributions. First, given the fluidness of teams in scientific collaboration and significant knowledge spillovers across teams, the ongoing fashion of team-study should be re-assessed. Given that (1) studies of corporate teams have proven useful and fruitful and (2) the model of scientific production is increasingly team-based (Wuchty et al., 2007), studying collaboration teams should contribute to better understanding of scientific creativity and correspondingly more effective policies to promote creativity. However, the organization of modern sciences has some distinct features, and these features calls for different perspectives instead of simply transplanting previous team-study frameworks from the corporate environment to the world of science.

Specifically, in scientific collaborations, teams are fluid and fuzzy, and an unambiguous entity as a "team" to be credited for a scientific product does not usually exist. This causes difficulties in identifying teams as objects to be studied and makes the social construct of "team" less useful for studies of science.

In addition, scientists participate in multiple teams simultaneously, and there are significant knowledge spillovers across these teams. In other words, creativity of a team depends on not only the internal process but also the external activities, and the knowledge creation process of a team cannot be isolated from its external environment. Therefore, to study scientific creativity at the team level, the external environment should also be taken into account. Furthermore, the external activity is more important to team performance than acknowledged by the corporate R&D management literature (Ancona & Caldwell, 1992; Bresman, 2010; Edmondson, 2002; Wong, 2004), and the distinction between internal and external learning is blurred in the context of scientific collaboration.

Second, integrating the structure- and the process- perspectives is very helpful for understanding the relationships between structures and creativity. For sociologists interested in structural determinants, there is often a gap between the structure and its predicted effects, and some psychology perspectives have proven useful to fill the gap and provide a micro-foundation to explain structural effects. For example, in his effort to bridge weak ties (i.e., structure) and the access to non-redundant information (i.e., effect), Granovetter cited the theory of cognitive balance (Granovetter, 1973) and the theory of homophily (Granovetter, 1983). Similarly, this dissertation has also proven that the cognitive- and social- psychology literature about creative process can serve as a useful micro-foundation to explain effects of tie strength. The process of idea generation, convergence, and implementation is affected by interpersonal relations, such as cognitive differences, shared understandings, and mutual trust. This perspective explains why and how does tie strength affect creative process and product.

In addition to providing a micro-foundation to explain structural effects, integrating the structure- and the process- perspectives also helps to reconcile competing predictions provided by different network theories. There have been long-standing debates between the weak tie theory (Granovetter, 1973, 1983) and the strong tie theory (Krackhardt, 1992; Uzzi, 1996, 1997), and between the structural hole theory (Burt, 1992, 2005) and the network closure theory (Coleman, 1988, 1990). These theories provide competing predictions about which type of networks are more advantageous to performance or other outcomes of interest, and both sides are supported by a number of empirical studies. A process-perspective might be one possible solution to integrate these competing theories for a more coherent and comprehensive network theory. For example, tie strength has different effects at different stages of the creative process. Therefore, we may observe evidence favoring weak ties in some cases and evidence favoring strong ties in other cases, depending on which stage we are studying. Moreover, at the aggregated level, we observe an inverted U-shaped relationship between tie strength and creativity, because the mixed effects of tie strength at different stages of the creative process. In addition, there seems to be an emerging interest in using the process approach to reconcile competing network theories in recent network studies (Fleming et al., 2007; Lavie & Drori, 2012; Obstfeld, 2005).

Third, the configuration of ties rather than a simple dichotomy between weak and strong ties should be investigated. Theories of weak and strong ties are at the dyadic level, and many studies have adopted a simple aggregation approach to explore network-level effects of tie strength, that is, the tie strength effect at the network level is a simple aggregation of effects at the dyadic level (Abbasi et al., 2011; Gabbay, 1997; Hansen, 1999; McFadyen & Cannella, 2004; McFadyen et al., 2009; Reagans & McEvily, 2003). Underlying this simple aggregation approach is an implicit assumption that networks have homogeneous ties.

However, this dissertation points out that this implicit assumption should not be taken as granted but instead should be evaluated. In the world of science, scientists have heterogeneous ties, simultaneously keeping a small group of strong ties and a large number of weak ties. Furthermore, this complicated configuration characteristics should be studied, instead of simply studying the overall tie strength of the network. For example, this dissertation argued that skewness has its own effect on network creativity, and one finding of this dissertation is that a good mixture of strong and weak ties is better than a homogeneous network.

In addition, this dissertation also calls for special attentions in translating dyadic-level theories into network-level predictions. The network-level effect might not be simple summation of dyadic-level effects, while moderating effect of tie configuration or interaction effects between ties should also be taken into account.

5.4. Policy Implications

This dissertation has three major implications for science policy and research evaluation. First, the findings of this dissertation contribute to the discussion on how to better design/reform the science funding system. There are increasing concerns in the

United States that the current project-based and peer-review-based funding model may impede path-breaking discoveries, because it favors projects and investigators with prior successful records and projects confirming rather than challenging current norms. This dissertation contributes to this discussion: There is another potential risk of the current funding model, that is, it may encourage repeated collaborations. Because the current funding model favors proposals with prior successful records, it may provide incentives for scientists to keep exploiting previous successful ideas and collaborative relations instead of exploring novel ideas and new collaborative relations. However, this may impede creativity of the whole scientific system because repeated collaborations are less creative.

Furthermore, given the widely accepted notion that collaboration is good for productivity and creativity, many funding agencies in many countries have established special programs supporting collaborative research. However, not all collaborations are equally beneficial to scientific creativity. The policy goal of promoting creativity might not be achieved if the program is filled with repeated collaborations or other types of collaborations that are not so creative. Therefore, finer-grained program design is needed to differentiate between collaborations, and put more focus on collaborations that are more likely to generate creative products, such as new collaborations rather than repeated ones.

However, adding efforts to support new collaborations does not mean abandoning all repeated collaborations. Heinze et al. (2009) noticed significant growth in group size following the main creative event, as a result of continued efforts following up and capitalizing on the opportunities opened up by the highly creative event. Although these later research activities are less creative, they are critical for realizing the potential of the initial highly creative event. Heinze et al. (2009) discussed this intriguing dilemma pertaining to the relationship between group size and creativity: small group size is important for the development of highly creative outputs, but highly creative event leads to the growth of group size, and the growth of group size decreases creativity. Similarly Hollingsworth (2009) discussed another intriguing question in science: why successful departments in history are less able to make breakthrough discoveries? Prior success gives birth to a set of institutional norms and routines, which in turn constraints future research activities because of path-dependency. Both dilemmas are beyond the scope of this dissertation, and future research is needed for better understanding these issues and developing effective policies. Nevertheless, there is one policy recommendation: on top of the current funding system, additional programs should be design specially for supporting new collaborations. In this case, collaborations following up initial creative event which are less creative but important for realizing the potential of initial creative event are not abandoned, while more opportunities are provided to stimulate highly creative events. In other word, this added effort creates a weaker institutional environment which encourages the emergence of new collaborations (analogous to departments in Hollingsworth's discussion) and therefore reduces the path-dependency at the level of the whole science system.

Another implication to the funding system is at the individual level. This dissertation finds that a skewed collaboration network is better for scientific creativity at the individual level. Therefore, a funding strategy helping scientists to build such skewed networks would contribute to higher creativity at both the individual- and the system-level. One possible recommendation is to set up lab-based funding programs with a proportion of funds reserved for outreaching activities. On the one hand, this funding strategy facilitates the development of strong ties between lab members within the lab. On the other hand, it creates opportunities for lab members to establish weak ties outside of the lab. One essential component of this funding strategy is to encourage active interactions between lab members, instead of building a virtual lab pooling researchers' profiles online without substantial collaborations between them. In addition, compared with project-based funding strategy, this lab-based funding strategy provides researchers with the flexibility to allocate the funds, which is found to be beneficial to creativity (Heinze et al., 2009). Another essential component is the reserved funds to encourage researchers to establish weak ties through visiting other institutions, hosting visiting scholars, and organizing workshops and conferences.

Second, this dissertation also raises a fundamental issue in science and technology policy, that is, what’s the precise goal of science and technology polices? The policy goal of creativity and innovativeness is widely accepted around the world, while the meaning of creativity and innovativeness is not always clear. There are actually two distinct goals: (1) increasing the average performance and reducing the variance and (2) foster highly creative product while tolerating many failures. Sometimes, creativity is understood dichotomously: a product is either highly creative or not. Correspondingly, the policy goal is to foster breakthroughs while failures are inevitable prices of uncertainty. At other times, creativity is understood as a continuum: all products are creative while some are more creative than others. In this context, a high and reliable average creativity also has its policy values. One analogy can be made to education policies: is the goal to have every student well-educated (i.e., high average and low variance/inequality) or to foster genius while leaving the majority of students behind (i.e., high maximum and high failure rate)?

Furthermore, different policy goals require distinct policy instruments. This dissertation shows that the average and maximum citations respond differently to different factors, and the effects of collaboration networks vary at different percentiles of the citation distribution. However, distinctions between these two policy goals are often missing in the current science and technology policy discussions. This ambiguity in the understanding of creativity and policy goal leads to problematic policy designs. Therefore, more detailed discussions about these two competing goals are needed, which goal to pursue needs to be pre-specified, and different policies should be designed and implemented for different goals. Correspondingly, different indicators are needed for evaluating policies or research units.

Third, this dissertation contributes to research evaluation using citation-based indicators. Citation-based indictors have been widely used in many countries to evaluate science performance of countries, institutions, individuals, and papers. Well-designed evaluation projects usually clarify that citation is not an indicator of research quality but a partial measure of impact and acknowledge the limitations of citation-based indicators, such as incomparability across fields, errors of using short citation time windows, and most importantly the "halo effect." This dissertation found significant prestige effect even after controlling for the intrinsic quality and intellectual appropriateness of the paper, and this prestige effect reflects a systematic bias in favor of prestigious scientists. Therefore, this prestige effect needs to be controlled or discounted in research evaluations using citation-based indicators.

REFERENCES

Abbasi, A., Altmann, J., & Hossain, L. 2011. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. ***Journal of Informetrics***, 5(4): 594-607.

Aghion, P., Dewatripont, M., & Stein, J. C. 2008. Academic freedom, private-sector focus, and the process of innovation. ***The RAND Journal of Economics***, 39(3): 617-635.

Ahuja, G. 2000. Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study. ***Administrative Science Quarterly***, 45(3): 425-455.

Aizenman, J., & Kletzer, K. 2011. The life cycle of scholars and papers in economics - the 'citation death tax'. ***Applied Economics***, 43(27): 4135-4148.

Aksnes, D. W. 2003. A macro study of self-citation. ***Scientometrics***, 56(2): 235-246.

Amabile, T. M. 1982. SOCIAL-PSYCHOLOGY OF CREATIVITY - A CONSENSUAL ASSESSMENT TECHNIQUE. ***Journal of Personality and Social Psychology***, 43(5): 997-1013.

Amabile, T. M. 1983. THE SOCIAL-PSYCHOLOGY OF CREATIVITY - A COMPONENTIAL CONCEPTUALIZATION. ***Journal of Personality and Social Psychology***, 45(2): 357-376.

Ancona, D. G. 1990. OUTWARD BOUND - STRATEGIES FOR TEAM SURVIVAL IN AN ORGANIZATION. ***Academy of Management Journal***, 33(2): 334-365.

Ancona, D. G., & Caldwell, D. F. 1988. BEYOND TASK AND MAINTENANCE - DEFINING EXTERNAL FUNCTIONS IN GROUPS. ***Group & Organization Studies***, 13(4): 468-494.

Ancona, D. G., & Caldwell, D. F. 1992. BRIDGING THE BOUNDARY - EXTERNAL ACTIVITY AND PERFORMANCE IN ORGANIZATIONAL TEAMS. ***Administrative Science Quarterly***, 37(4): 634-665.

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

Barron, F., & Harrington, D. M. 1981. CREATIVITY, INTELLIGENCE, AND PERSONALITY. ***Annual Review of Psychology***, 32: 439-476.

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

Beaver, D. B., & Rosen, R. 1979a. Studies in scientific collaboration Part III. Professionalization and the natural history of modern scientific co-authorship. ***Scientometrics***, 1(3): 231-245.

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

Beyer, J. M., Chanove, R. G., & Fox, W. B. 1995. REVIEW PROCESS AND THE FATES OF MANUSCRIPTS SUBMITTED TO AMJ. ***Academy of Management Journal***, 38(5): 1219-1260.

Biagioli, M. 1999. Aporias of scientific authorship: credit and responsibility in contemporary biomedicine. ***The Science Studies Reader, edited by M. BIAGIOLI. New York: Routledge***.

Bordons, M., & Gomez, I. 2000. Collaboration networks in science. ***The web of knowledge: A festschrift in honor of Eugene Garfield***: 197–213.

Bresman, H. 2010. External Learning Activities and Team Performance: A Multimethod Field Study. ***Organization Science***, 21(1): 81-96.

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

Burt, R. S. 2005. ***Brokerage and closure : an introduction to social capital***. Oxford ; New York: Oxford University Press.

Catalini, C. 2012. Microgeography and the Direction of Inventive Activity. ***Available at SSRN 2126890***.

Cohen, M. D., March, J. G., & Olsen, J. P. 1972. A Garbage Can Model of Organizational Choice. ***Administrative Science Quarterly***, 17(1): 1-25.

Cohen, W. M., & Levinthal, D. A. 1990. Absorptive-Capacity - a New Perspective on Learning and Innovation. ***Administrative Science Quarterly***, 35(1): 128-152.

Cole, S., & Cole, J. R. 1967. SCIENTIFIC OUTPUT AND RECOGNITION - STUDY IN OPERATION OF REWARD SYSTEM IN SCIENCE. ***American Sociological Review***, 32(3): 377-390.

Coleman, J. S. 1988. Social Capital in the Creation of Human Capital. ***American Journal of Sociology***, 94: S95-S120.

Coleman, J. S. 1990. ***Foundations of social theory***. Cambridge, Mass.: Belknap Press of Harvard University Press.

Cozzens, S. 1989. What do citations count? the rhetoric-first model. ***Scientometrics***, 15(5-6): 437-447.

Crane, D. 1965. Scientists at Major and Minor Universities: A Study of Productivity and Recognition. ***American Sociological Review***, 30(5): 699-714.

Crane, D. 1969. Social Structure in a Group of Scientists: A Test of the "Invisible College" Hypothesis. ***American Sociological Review***, 34(3): 335-352.

Cronin, B. 2001. Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? ***Journal of the American Society for Information Science and Technology***, 52(7): 558-569.

Cronin, B. 2005. ***The Hand of Science: Academic Writing and Its Rewards***. Lanham, MD: Scarecrow Press.

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

De Bellis, N. 2009. ***Bibliometrics and citation analysis: from the Science citation index to cybermetrics***. Lanham, MD: Scarecrow Press.

de Solla Price, D. J. 1986. ***Little science, big science--and beyond***: Columbia University Press New York.

de Solla Price, D. J., & Beaver, D. 1966. Collaboration in an invisible college. ***American Psychologist***, 21(11): 1011-1018.

Drazin, R., Glynn, M. A., & Kazanjian, R. K. 1999. Multilevel theorizing about creativity in organizations: A sensemaking perspective. ***Academy of Management Review***, 24(2): 286-307.

Edge, D. 1979. Quantitative measures of communication in science: A critical review. ***History of Science***, 17(36): 102-134.

Edmondson, A. C. 2002. The local and variegated nature of learning in organizations: Aa group-level perspective. ***Organization Science***, 13(2): 128-146.

Fleming, L. 2001. Recombinant uncertainty in technological search. ***Management Science***, 47(1): 117-132.

Fleming, L., Mingo, S., & Chen, D. 2007. Collaborative brokerage, generative creativity, and creative success. ***Administrative Science Quarterly***, 52(3): 443-475.

Ford, C. M. 1996. Theory of individual creative action in multiple social domains. ***Academy of Management Review***, 21(4): 1112-1142.

Fox, J. 2006. Teacher's Corner: structural equation modeling with the sem package in R. ***Structural equation modeling***, 13(3): 465-486.

Fox, M. F. 1983. PUBLICATION PRODUCTIVITY AMONG SCIENTISTS - A CRITICAL-REVIEW. ***Social Studies of Science***, 13(2): 285-305.

Gabbay, S. M. 1997. ***Social capital in the creation of financial capital: The case of network marketing***. Champaign, IL: Stipes Publishing.

Garfield, E. 1973. CITATION AND DISTINCTION. ***Nature***, 242(5398): 485-485.

Geraci, M. 2013. lqmm: Linear quantile mixed models. ***R package version 1.03***.

Ghiglino, C. 2012. Random walk to innovation: Why productivity follows a power law. ***Journal of Economic Theory***, 147(2): 713-737.

Gilbert, G. N. 1977. REFERENCING AS PERSUASION. ***Social Studies of Science***, 7(1): 113-122.

Glänzel, W., Debackere, K., Thijs, B., & Schubert, A. 2006. A concise review on the role of author self-citations in information science, bibliometrics and science policy. ***Scientometrics***, 67(2): 263-277.

Glänzel, W., Schlemmer, B., & Thijs, B. 2003. Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. ***Scientometrics***, 58(3): 571-586.

Glänzel, W., & Schubert, A. 2005. Analysing scientific networks through co-authorship. ***Handbook of quantitative science and technology research***: 257-276.

Gordon, M. D. 1980. A critical reassessment of inferred relations between multiple authorship, scientific collaboration, the production of papers and their acceptance for publication. ***Scientometrics***, 2(3): 193-201.

Granovetter, M. 1973. Strength of Weak Ties. ***American Journal of Sociology***, 78(6): 1360-1380.

Granovetter, M. 1983. The strength of weak ties: A network theory revisited. ***Sociological theory***, 1(1): 201-233.

Granovetter, M. 1985. ECONOMIC-ACTION AND SOCIAL-STRUCTURE - THE PROBLEM OF EMBEDDEDNESS. ***American Journal of Sociology***, 91(3): 481-510.

Guilford, J. P. 1950. Creativity. ***American Psychologist***, 5(9): 444-454.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. 2005. Team assembly mechanisms determine collaboration network structure and team performance. ***Science***, 308(5722): 697-702.

Gupta, A. K., Smith, K. G., & Shalley, C. E. 2006. The Interplay between Exploration and Exploitation. ***Academy of Management Journal***, 49(4): 693-706.

Hackman, I. R., & Morris, C. G. 1975. Group tosks, group interaction processes, and group performance effectiveness: A review and proposed integration. In L. Berkowitz (Ed.), ***Advances in experimental social psychology***: 47-99. New York: Academic Press.

Hansen, M. T. 1999. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. ***Administrative Science Quarterly***, 44(1): 82-111.

Heinze, T., Shapira, P., Rogers, J. D., & Senker, J. M. 2009. Organizational and institutional influences on creativity in scientific research. ***Research Policy***, 38(4): 610-623.

Helmreich, R. L., Spence, J. T., Beane, W. E., Lucker, G. W., & Matthews, K. A. 1980. Making it in academic psychology: Demographic and personality correlates of attainment. ***Journal of Personality and Social Psychology***, 39(5): 896-908.

Hicks, D., & Katz, J. S. 2011. Equity and Excellence in Research Funding. ***Minerva***, 49(2): 137-151.

Hollingsworth, R. 2004. Institutionalizing excellence in biomedical research: the case of Rockefeller University. In D. H. Stapleton (Ed.), ***Creating a Tradition of Biomedical Research***. New York: Rockefeller University Press.

Hollingsworth, R. 2009. Scientific discoveries: an institutionalist and path-dependent perspective. In C. Hannaway (Ed.), ***Biomedicine in the Twentieth Century: Practices, Policies, and Politics***: 317-353. Bethesda, MD: National Institutes of Health.

Huckman, R. S., & Staats, B. R. 2011. Fluid Tasks and Fluid Teams: The Impact of Diversity in Experience and Team Familiarity on Team Performance. ***M&Som-Manufacturing & Service Operations Management***, 13(3): 310-328.

Hwang, K. 2008. International collaboration in multilayered center-periphery in the globalization of science and technology. ***Science, Technology & Human Values***, 33(1): 101-133.

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

Katz, J. S., & Hicks, D. 1997. How much is a collaboration worth? A calibrated bibliometric model. ***Scientometrics***, 40(3): 541-554.

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

King, N., & Anderson, N. 1990. Innovation in working groups. In M. A. West, & J. L. Farr (Eds.), ***Innovation and creativity ai work***: 81-100. Chichester, England: Wiley.

Krackhardt, D. 1992. The strength of strong ties: The importance of philos in organizations. In N. Nohria, & R. G. Eccles (Eds.), ***Networks and organizations: Structure, form, and action***: 216-239. Boston, MA: Harvard Business School Press.

Krackhardt, D., & Porter, L. W. 1985. When Friends Leave: A Structural Analysis of the Relationship between Turnover and Stayers' Attitudes. ***Administrative Science Quarterly***, 30(2): 242-261.

Laband, D. N., & Piette, M. J. 1994. A CITATION ANALYSIS OF THE IMPACT OF BLINDED PEER-REVIEW. ***Jama-Journal of the American Medical Association***, 272(2): 147-149.

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

Landry, R., Traore, N., & Godin, B. 1996. An econometric analysis of the effect of collaboration on academic research productivity. ***Higher Education***, 32(3): 283-301.

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

Laudel, G. 2002. What do we measure by co-authorships? ***Research Evaluation***, 11(1): 3-15.

Lavie, D., & Drori, I. 2012. Collaborating for Knowledge Creation and Application: The Case of Nanotechnology Research Programs. ***Organization Science***, 23(3): 704-724.

Lee, S., & Bozeman, B. 2005. The impact of research collaboration on scientific productivity. ***Social Studies of Science***, 35(5): 673-702.

Levin, S. G., & Stephan, P. E. 1991. RESEARCH PRODUCTIVITY OVER THE LIFE-CYCLE - EVIDENCE FOR ACADEMIC SCIENTISTS. ***American Economic Review***, 81(1): 114-132.

Levine, J. M., & Moreland, R. L. 2004. Collaboration: The social context of theory development. ***Personality and Social Psychology Review***, 8(2): 164-172.

Lin, N., & Ensel, W. M. 1989. Life Stress and Health: Stressors and Resources. ***American Sociological Review***, 54(3): 382-382.

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

Machado, J. A. F., & Silva, J. M. C. S. 2005. Quantiles for Counts. ***Journal of the American Statistical Association***, 100(472): 1226-1237.

Manso, G. 2011. Motivating Innovation. ***The Journal of Finance***, 66(5): 1823-1860.

March, J. G. 1991. Exploration and Exploitation in Organizational Learning. ***Organization Science***, 2(1): 71-87.

Martin, B. R., & Irvine, J. 1983. ASSESSING BASIC RESEARCH - SOME PARTIAL INDICATORS OF SCIENTIFIC PROGRESS IN RADIO ASTRONOMY. ***Research Policy***, 12(2): 61-90.

McCrae, R. R. 1987. CREATIVITY, DIVERGENT THINKING, AND OPENNESS TO EXPERIENCE. ***Journal of Personality and Social Psychology***, 52(6): 1258-1265.

McDowell, J. M., & Melvin, M. 1983. The determinants of co-authorship: An analysis of the economics literature. ***The Review of Economics and Statistics***, 65(1): 155-160.

McFadyen, M. A., & Cannella, A. A. 2004. Social capital and knowledge creation: Diminishing returns of the number and strength of exchange relationships. ***Academy of Management Journal***, 47(5): 735-746.

McFadyen, M. A., Semadeni, M., & Cannella, A. A. 2009. Value of Strong Ties to Disconnected Others: Examining Knowledge Creation in Biomedicine. ***Organization Science***, 20(3): 552-564.

Mednick, S. A. 1962. THE ASSOCIATIVE BASIS OF THE CREATIVE PROCESS. ***Psychological Review***, 69(3): 220-232.

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

Melin, G., & Persson, O. 1996. Studying research collaboration using co-authorships. ***Scientometrics***, 36(3): 363-377.

Merton, R. K. 1957. PRIORITIES IN SCIENTIFIC DISCOVERY - A CHAPTER IN THE SOCIOLOGY OF SCIENCE. ***American Sociological Review***, 22(6): 635-659.

Merton, R. K. 1968. The Matthew Effect in Science. ***Science***, 159(3810): 56-63.

Merton, R. K. 1973. ***The sociology of science : theoretical and empirical investigations***. Chicago, IL: University of Chicago Press.

Merton, R. K. 1983. Foreword. In E. Garfield (Ed.), ***Citation indexing, its theory and application in science, technology, and humanities***: xiii, 274 p. Philadelphia, PA: ISI Press.

Moed, H. F., Burger, W., Frankfort, J., & van Raan, A. 1985. The application of bibliometric indicators: Important field- and time-dependent factors to be considered. ***Scientometrics***, 8(3): 177-203.

Moody, J. 2004. The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999. ***American Sociological Review***, 69(2): 213-238.

Mumford, M. D., & Gustafson, S. B. 1988. CREATIVITY SYNDROME - INTEGRATION, APPLICATION, AND INNOVATION. ***Psychological Bulletin***, 103(1): 27-43.

Nahapiet, J., & Ghoshal, S. 1998. Social Capital, Intellectual Capital, and the Organizational Advantage. ***Academy of Management Review***, 23(2): 242-266.

Nelson, R. R., & Winter, S. G. 1982. ***An evolutionary theory of economic change***. Cambridge, MA: Belknap Press of Harvard University Press.

NETWISE. 2007. NETWISE Project.

Newman, J. M., & Cooper, E. 1993. DETERMINANTS OF ACADEMIC RECOGNITION - THE CASE OF THE JOURNAL-OF-APPLIED-PSYCHOLOGY. ***Journal of Applied Psychology***, 78(3): 518-526.

Nilles, J. M. 1975. Interdisciplinary research management in the university environment. ***Journal of the Society of Research Administrators***, 6: 9–16.

Nonaka, I. 1994. A Dynamic Theory of Organizational Knowledge Creation. ***Organization Science***, 5(1): 14-37.

Obstfeld, D. 2005. Social Networks, the Tertius Iungens Orientation, and Involvement in Innovation. ***Administrative Science Quarterly***, 50(1): 100-130.

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

Perretti, F., & Negro, G. 2006. Filling Empty Seats: How Status and Organizational Hierarchies Affect Exploration versus Exploitation in Team Design. ***Academy of Management Journal***, 49(4): 759-777.

Perry-Smith, J. E., & Shalley, C. E. 2003. The social side of creativity: A static and dynamic social network perspective. ***Academy of Management Review***, 28(1): 89-106.

Pfeffer, J., & Salancik, G. R. 1978. ***The external control of organizations: a resource dependence perspective***. New York, NY: Harper and Row.

Pierce, S. J. 1999. Boundary crossing in research literatures as a means of interdisciplinary information transfer. ***Journal of the American Society for Information Science***, 50(3): 271-279.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. 2013. nlme: Linear and Nonlinear Mixed Effects Models. ***R package version 3.1-109***.

Podolny, J. M., & Baron, J. N. 1997. Resources and Relationships: Social Networks and Mobility in the Workplace. ***American Sociological Review***, 62(5): 673-693.

Porac, J. F., Wade, J. B., Fischer, H. M., Brown, J., Kanfer, A., & Bowker, G. 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): 661-678.

Presser, S. 1980. COLLABORATION AND THE QUALITY OF RESEARCH. ***Social Studies of Science***, 10(1): 95-101.

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): 869-881.

Reagans, R., & McEvily, B. 2003. Network structure and knowledge transfer: The effects of cohesion and range. ***Administrative Science Quarterly***, 48(2): 240-267.

Schumpeter, J. A. 1939. ***Business cycles; a theoretical, historical, and statistical analysis of the capitalist process*** (1st ed.). New York, London,: McGraw-Hill Book Company, inc.

Simonton, D. K. 1984. Artistic creativity and interpersonal relationships across and within generations. ***Journal of Personality and Social Psychology***, 46(6): 1273-1286.

Simonton, D. K. 1999. ***Origins of genius : Darwinian perspectives on creativity***. New York: Oxford University Press.

Simonton, D. K. 2004. ***Creativity in science : chance, logic, genius, and Zeitgeist***. Cambridge, UK ; New York: Cambridge University Press.

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

Skaug, H., Fournier, D., Nielsen, A., Magnusson, A., & Bolker, B. 2013. glmmADMB: generalized linear mixed models using AD model builder. ***R package version 0.7.4***.

Skilton, P. F., & Dooley, K. J. 2010. THE EFFECTS OF REPEAT COLLABORATION ON CREATIVE ABRASION. ***Academy of Management Review***, 35(1): 118-134.

Smart, J. C., & Bayer, A. E. 1986. AUTHOR COLLABORATION AND IMPACT - A NOTE ON CITATION RATES OF SINGLE AND MULTIPLE AUTHORED ARTICLES. ***Scientometrics***, 10(5-6): 297-305.

Star, S. L., & Griesemer, J. R. 1989. Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. ***Social Studies of Science***, 19(3): 387-420.

Stephan, P. E., & Levin, S. G. 1991. Inequality in Scientific Performance: Adjustment for Attribution and Journal Impact. ***Social Studies of Science***, 21(2): 351-368.

Tortoriello, M., & Krackhardt, D. 2010. Activating cross-boundary knowledge: The role of Simmelian ties in the generation of innovations. ***Academy of Management Journal***, 53(1): 167-181.

Uzzi, B. 1996. The sources and consequences of embeddedness for the economic performance of organizations: The network effect. ***American Sociological Review***, 61(4): 674-698.

Uzzi, B. 1997. Social structure and competition in interfirm networks: The paradox of embeddedness. ***Administrative Science Quarterly***, 42(1): 35-67.

Uzzi, B., & Spiro, J. 2005. Collaboration and creativity: The small world problem. ***American Journal of Sociology***, 111(2): 447-504.

Van Raan, A. F. J. 1998. The influence of international collaboration on the impact of research results. ***Scientometrics***, 42(3): 423-428.

Van Raan, A. F. J. 2004. Sleeping beauties in science. ***Scientometrics***, 59(3): 467-472.

van Rijnsoever, F. J., & Hessels, L. K. 2011. Factors associated with disciplinary and interdisciplinary research collaboration. ***Research Policy***, 40(3): 463-472.

van Rijnsoever, F. J., Hessels, L. K., & Vandeberg, R. L. J. 2008. A resource-based view on the interactions of university researchers. ***Research Policy***, 37(8): 1255-1266.

Venables, W. N., & Ripley, B. D. 2002. ***Modern Applied Statistics with S*** (Fourth ed.). New York: Springer.

Walsh, J. P., & Maloney, N. G. 2002. Computer Network Use, Collaboration Structures and Productivity. In P. Hinds, & S. Kiesler (Eds.), ***Distributed Work***: 433-458. Cambridge, MA: MIT Press.

Wang, J. 2013. Citation time window choice for research impact evaluation. ***Scientometrics***, 94(3): 851-872.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. 2012. A boosted-trees method for name disambiguation. ***Scientometrics***, 93(2): 391-411.

Weick, K. E. 1976. EDUCATIONAL ORGANIZATIONS AS LOOSELY COUPLED SYSTEMS. ***Administrative Science Quarterly***, 21(1): 1-19.

West, M. A. 2002. Sparkling fountains or stagnant ponds: An integrative model of creativity and innovation implementation in work groups. ***Applied Psychology-an International Review-Psychologie Appliquee-Revue Internationale***, 51(3): 355-387.

Whitley, R. 2000. ***The intellectual and social organization of the sciences*** (2nd ed.). Oxford England ; New York: Oxford University Press.

Williamson, O. E. 1981. THE ECONOMICS OF ORGANIZATION - THE TRANSACTION COST APPROACH. ***American Journal of Sociology***, 87(3): 548-577.

Wittenbaum, G. M. 2003. Putting communication into the study of group memory. ***Human Communication Research***, 29(4): 616-623.

Wong, S. S. 2004. Distal and local group learning: Performance trade-offs and tensions. ***Organization Science***, 15(6): 645-656.

Woodman, R. W., Sawyer, J. E., & Griffin, R. W. 1993. TOWARD A THEORY OF ORGANIZATIONAL CREATIVITY. ***Academy of Management Review***, 18(2): 293-321.

Wuchty, S., Jones, B. F., & Uzzi, B. 2007. The increasing dominance of teams in production of knowledge. ***Science***, 316(5827): 1036-1039.

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