WHY DO PROJECT PARTICIPANTS WORK TOGETHER? AN INVESTIGATION OF THE ANTECEDENTS OF COLLABORATION TIE STRENGTH IN CYBERINFRASTRUCTURE PROJECTS

by

Qin Weng

B.S., Beijing Foreign Studies University, 1998

M.S., Virginia Commonwealth University, 2000

Submitted to the Graduate Faculty of

the Joseph M. Katz Graduate School of Business in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

THE JOSEPH M. KATZ GRADUATE SCHOOL OF BUSINESS

This dissertation was presented

by

Qin Weng

It was defended on

November 30, 2017

and approved by

Dennis F. Galletta, PhD
Professor of Business Administration
University of Pittsburgh

Narayan Ramasubbu, PhD
Associate Professor of Business Administration
University of Pittsburgh

Audrey J. Murrell, PhD
Associate Professor of Business Administration
University of Pittsburgh

Greg D. Moody, PhD
Assistant Professor of Information Systems
University of Nevada Las Vegas

Dissertation Advisor: Laurie J. Kirsch, PhD
Professor of Business Administration
University of Pittsburgh

**WHY DO PROJECT PARTICIPANTS WORK TOGETHER?
AN INVESTIGATION OF THE ANTECEDENTS OF COLLABORATION TIE STRENGTH IN CYBERINFRASTRUCTURE PROJECTS**

Qin Weng, PhD

University of Pittsburgh, 2018

A cyberinfrastructure (CI) project is a new form of large-scale distributed project that is different from other information systems projects, including open source software projects and distributed organizational information systems development projects. These projects may share some similarities but they also have many striking differences, including differences in goals, funding, participants, control types and coordination mechanisms.

A CI project aims to build a complex digital infrastructure to enable innovative and transformative research. Such digital infrastructure offers scientists and researchers a set of connected resources including laboratories, databases, computer hardware, software and people, so that they can conduct research that addresses complex questions that are beyond the capability of any individual person or institution. CI Projects are grand in scope and challenging to execute. Successfully building cyberinfrastructure requires intense and sustained collaborative efforts of people from diverse backgrounds and disciplines and from different organizations and institutions.

However, collaboration among key project participants is a complex phenomenon. In part, due to the different backgrounds and knowledge of the project participants, a number of factors may promote or hinder their collaboration. Furthermore, individuals associated with efforts to build cyberinfrastructure, unlike individuals involved in projects within traditional organizational settings, are free to choose with whom they want to collaborate. Therefore, collaboration in CI projects is not assigned as is typical in organizational projects but rather involves individual

choices to collaborate. Understanding the factors that promote collaboration will not only help us better understand individual behaviors, but also provide insights for the project management team in crafting better strategies to promote collaboration. This study intends to examine the antecedents of *collaboration tie strength* among CI project participants. The research question I address is: *what are the antecedents of collaboration tie strength in cyberinfrastructure projects?* Since a CI project is both technologically intensive and socially complex, I examine the antecedents of collaboration tie strength from both the technological perspective and the social perspective. More importantly, this study also examines: *how these social factors and technological factors interact with each other in predicting collaboration tie strength?*

The research site is a National Science Foundation (NSF) funded CI project, named The Global Environment for Network Innovations (GENI[1]). This project intends to build a digital infrastructure for networking scientists to explore the next generation Internet at scale. The major stakeholders involved in this project include computer scientists and researchers, government agencies, industrial professionals and students.

By the time of this study, GENI had gone through two phases: pre-planning and planning. The construction phase was ongoing. Pre-planning phase consists of events before 2004, by which time NSF agreed to support GENI and hired key individuals to drive the initiative. The planning phase began in 2004 and lasted until 2008. In this phase, groups of researchers and sponsors worked together to shape the idea of the GENI project in terms of its vision, goals and organization. The construction phase commenced in 2008. This phase featured development activities to build

[1] GENI project website: http://www.geni.net/

the specific CI technologies in different GENI technical clusters and to roll out the GENI projects to a larger scale.

The management team, GENI Project Office (GPO), adopted a spiral development approach (Boehm 1986), with each spiral involving steps of a complete project development lifecycle. A spiral began with the GPO setting out goals and allocating funds to different project teams. Each spiral lasted one year, so that each spiral ended after a year of complete project development cycle. At that point the performance of these projects was reviewed, and the evaluation results became, in part, the basis for funding decisions of the next spiral. In each spiral, the overall GENI project was divided into many smaller projects. A major form of collaboration among individuals involved in GENI, or interested in becoming involved in GENI, is through forming project teams. Each individual can choose to work with others on a specific project. An individual can also work on multiple projects with different people.

At the time of this study, the GENI project completed four spirals (1 to 4), and the project was still ongoing. My study focuses on the most recent spiral at the time of this study, i.e., project spiral 4, which includes around 126 projects with an average of five people on each project. The unit of analysis is the dyad of collaborative individuals. The dependent variable is collaboration tie strength, i.e., the number of projects two individuals both worked on in the GENI project spiral 4 weighted by the project size. The study examines the antecedents of collaboration tie strength from both the technological and social perspectives. From the technological perspective, *knowledge dependency*, *technical dependency* and *resource dependency* are considered to positively predict collaboration tie strength. From the social perspective, *power distance*, *social similarity* and *familiarity* are considered to positively predict collaboration tie strength.

Furthermore, the study examines the interactions between the three social factors and the three technological factors in predicting collaboration tie strength.

Three main sets of analyses are carried out to test the hypotheses with collaboration tie strength as the dependent variable. In the first set of analyses, collaboration tie strength, is measured as a binary variable and the full dataset is used for the analysis with logistic regression as the regression method. This set of analyses helps show how the different social and technological factors predict whether two people collaborate. In the second set of analyses, collaboration tie strength, is measured as a numeric count variable and the full data is used for the analysis with Poisson regression as the regression method. This set of analyses helps show how the different social and technological factors predict how many times two people collaborate. In the third set of analyses, collaboration tie strength, is still measured as a numeric count variable but only the partial dataset where collaboration tie strength is non-zero is used for the analysis with Poisson regression as the regression method. This set of analyses helps show how the different social and technological factors predict how many times two people collaborate for those who actually collaborated.

The results suggest that resource dependency, technical dependency and familiarity all significantly positively predict whether two people collaborate. Resource dependency and familiarity positively predict the number of times two people collaborate. Technical dependency and familiarity positively predict the number of times two people collaborate for those who actually collaborated. Overall, technological factors yield stronger positive prediction for collaboration tie strength than social factors.

Interactions between certain technological factors and social factors are also found to be significant, with all interaction coefficients being negative. In particular, similarity and familiarity both suppress the prediction of resource dependency on whether two people collaborate. Power distance suppresses the prediction of technical dependency on how many times two people collaborate for those who actually collaborated. Despite all the interaction effects, the prediction of all social and technological factors remains positive.

This study makes both theoretical and practical contributions. From the theoretical perspective, through the empirical study on the GENI project, the work not only contributes to the IS research on IT collaboration, but also answers the calls for more IS research on CI projects. From the practical perspective, the findings of this study suggest to CI project managers and fund providers that collaboration in CI projects is a very complicated phenomenon. It evolves, changes and depends on many factors. By providing a fine-grained view of how different social and technological factors interact and predict collaboration tie strength, this study may help project management in crafting better strategies to promote collaboration.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

To me, pursuing the doctoral studies at the Katz Graduate School of Business at the University of Pittsburgh is a long but rewarding journey. I felt honored when I was first admitted to the PhD program at Katz and I still feel blessed after many years of study. If asked what I love most about life here, I would say the people. My entire faculty of the doctoral program has been so supportive and helpful. They not only guided me as mentors with my study, but also helped me as friends with my life. There are many people I want to thank, but here I would like to express my gratitude to some of the most special people that helped me achieve my PhD degree.

Above all, I would like to express my deep gratitude to my advisor, Laurie Kirsch. Her help to me is tremendous and beyond my words. When I joined Katz doctoral program, she provided me the opportunity of working with her on the cyberinfrastructure project, which paved the foundation of my dissertation study eventually conducted here. Throughout the years, during my most difficulty times, she was always there encouraging, supporting and helping me. Even though she is busy with many other responsibilities, she always makes herself available whenever I need her. She is not only my mentor but also a role model to me as a great scholar and as a person.

I would like to give special thanks to our program director, Dennis Galletta. Over the years, I learned from him on how to do research through the many lectures taken from him, and I learned even more on how to be a good scholar through our day-to-day interactions. He is always kind, patient and caring. His wonderful personality will certainly influence me many years down the road in my academic life, when I teach, when I conduct research and when I work with colleagues.

I would like to thank the rest of my dissertation committee, Audrey Murrell, Narayan Ramasubbu and Greg Moody. They have provided me with valuable and insightful reviews and suggestions on my dissertation. They also made availability among their busy schedules to meet with me, fulfilling every step needed for the dissertation composition.

I would also like to thank my departmental staff, Carrie Woods and Chris Gursky for guiding me through the important procedures for fulfilling doctoral studies and providing me with the things needed to achieve my degree.

I want to extend my gratitude to the two professors who passed away during my PhD study, Sandra Slaughter and Kevin Kim. They were part of my committee in the early stages of my dissertation, and they both deeply touched my life. Sandy was extremely instrumental in helping me formulate the dissertation ideas and Kevin greatly helped me with developing my statistical skills needed for data analysis. Losing them was a tremendous tragedy that affected me greatly during my PhD study. They were greatly missed but they also motivated me to conquer all the difficulties and fulfill the requirements of my doctoral study.

Last but not the least, I want to express my deepest love and thanks to my family. They supported my decision to pursue the PhD degree after many years of my professional life. To become a mother that my children could be proud of is also something that has been inspiring me over the years and in the many years to come. My family is the joy of my life and the biggest blessing from God.

1.0 INTRODUCTION

Today's world is a connected world. Since the birth of the Internet, people have found ways to connect, to communicate and to collaborate through the cyberspace as was never previously possible. Like others, scientists learned to unleash the power of the Internet through collaborative work efforts using cyberinfrastructure (Atkins et al. 2003).

1.1 CYBERINFRASTRUCTURE AND CYBERINFRASTRUCTURE PROJECTS

Infrastructure, by definition, refers to "the basic physical and organizational structures and facilities needed for the operation of a society or enterprise."[2] Commonly known is public infrastructure that typically includes physical structures that support a society, such as roads, bridges, water supply systems, sewers, electrical grids and so forth. These interrelated physical components, as part of the public infrastructure, provide commodities and services essential to enable, sustain or enhance societal living conditions (Fulmer 2009). In contrast to physical infrastructure, cyberinfrastructure (CI) is an information technology system, with laboratories, computers, data and people linked by high-speed networks, that provides advanced and powerful

---

[2] *Infrastructure*, Online Compact Oxford English Dictionary, http://www.oxforddictionaries.com/definition/english/infrastructure?q=infrastructure (accessed July 12, 2014)

capabilities (Atkins et al. 2003). CI and public infrastructure are similar in that they both provide basic structures and resources, but are different in that they include different components and serve different purposes.

CI is an idea that emerged over the last decade from technological advances based on the convergence of basic technologies including the Internet, microchip technology and databases (Hart 2014). Some scientific research, because of its complexity and difficulty, is too hard and too complex for a single person or institution to conduct, and thus calls for collaborative efforts from a larger group of people using a variety of resources including data and technological instruments (Kirsch and Slaughter 2013a). By bringing together different resources through computer-supported systems (Finholt and Birnholtz 2006), CI provides a new environment to revolutionize scientific and engineering research, which has been witnessed in a variety of fields such as atmospheric science, oceanic science, environmental science, biology, computer science and social science (Atkins et al. 2003).

A CI project is a set of activities and efforts that aim to build a digital infrastructure that enables radical scientific findings and innovation within various scientific communities (Berman 2008, Edwards et al. 2009, Bietz et al. 2010, Kirsch and Slaughter 2013a). For example, the George E. Brown, Jr., Network for Earthquake Engineering Simulation (NEES) project is an 89 million dollar CI project funded by the engineering directorate of NSF that aims to build a national scale grid resource that helps earthquake engineers to conduct research to better understand earthquakes, along with their causes and effects (www.nees.org). The Global Environment for Network Innovations (GENI) project is another multi-million dollar CI project funded by NSF to develop a virtual laboratory that supports network engineers to conduct at-scale networking experimentation and distributed systems research and education (www.geni.net).

Many questions that scientists try to address call for CI. Addressing these questions is challenging but can provide transformational results and has far-reaching impact on a national, or even international, level (Finholt and Birnholtz 2006, de la Flor et al. 2010). Sometimes the questions may not be apparently critical or even relevant in the current society but may have far-reaching impact on humanity in the future. Sometimes the questions are grand in nature but are beyond the capabilities of individual scientists to address. For example, how to predict tsunami and earthquakes and how to prevent them? How can we better capture oceanic wave patterns with new technologies to more accurately predict weather changes? To answer these types of questions, the resources and abilities of a small group of people are normally far from sufficient. On the other hand, a cyberinfrastructure can provide scientists and researchers a network of resources, including not only hardware and software but also people and data, making it feasible for them to investigate questions that are grand in scope and too complicated or difficult for an individual to address.

1.2 CI PROJECTS AND OTHER IS PROJECTS

CI projects differ from other IT projects in a variety of ways. Kirsch and Slaughter (2013a) thoroughly distinguished CI projects (or "cyber projects") from other IT projects, including open source software (OSS) projects, R&D projects, information systems development (ISD) projects and mega-infrastructure projects along a variety of dimensions. For example, compared to traditional IT projects, CI projects are large-scale and distributed, composed of data, computer hardware and software, people, and facilities in numerous locations. Furthermore, CI projects normally involve governmental intervention because they are typically sponsored or funded by government agencies. Such projects are characterized with a variety of participation roles and management hierarchy. Finally, CI projects are long-term in nature, often accomplished over the course of many years, or even decades, vastly exceeding the scope, size and budget of most traditional IT projects.

Building upon their findings, this paper will compare CI projects to two types of IT projects with similar characteristics, i.e., OSS projects and distributed organizational ISD projects, along the dimensions proposed by Mintzberg (1979). He suggested that the major variables for organization structure include parts of the organization, coordination mechanisms and design parameters. Based on his theory, Table 1 compares the three types of projects from the perspectives of participants, goal, funding, control forms and coordination mechanisms. The comparison of these features helps us to understand the context for collaboration in each type of project, and why collaboration is germane to the overall success of CI projects.

Table 1. Comparison of Projects

	Open Source Software Projects	**CI Projects**	**Distributed Organizational ISD Projects**
Goal	Open source software, semi-clearly defined, bounded	Scientific collaboration and innovation, not clearly defined, unbounded	Organizational revenues, clearly defined, bounded
Funding	Sponsoring companies	Government and external agencies	Company budgets
Participants	Society IT professionals	Researchers, scientists, students, professionals, government	Organizational IT professionals
Control forms	Peer-based behavioral control	Formal organizational and peer-based behavioral control	Formal organizational and supplementary informal control
Coordination mechanisms	Merit-based, self-organizing	Funding allocation, self-organizing, choice in collaboration, personnel appointment, conferences	Project planning, organizational structure, organizational meetings

First, CI projects differ from OSS projects and organizational ISD projects in terms of project goals. Distributed organizational ISD projects have clearly defined and bounded goals; most of which are business profit driven. Organizational ISD projects have clear specifications in terms of features and functions to be accomplished. In contrast, OSS projects are more targeted toward developing non-proprietary software, and most of these projects are non-profit driven (Aksulu & Wade 2010). OSS projects have comparatively clearly defined goals regarding the functionality and components of the software to be developed. However, in most cases, OSS projects do not impose strict schedule and feature requirements as with organizational ISD projects. Goals for CI projects are vaguely defined. CI projects have an over-arching goal of transforming science and research, however, how to achieve this goal and the specific end-state, are unknown and unknowable (Kirsch and Slaughter 2013a). Goals for CI projects are unbounded in nature because it is not possible to articulate the specific requirements of the targeted outcome.

The source of funding and its impact on participants are also very different for the three types of projects. This aspect of organizational ISD projects tends to be straightforward, as projects are typically funded through internal business budgets. Employees are paid to work on these projects and failure to implement the projects as required might risk their jobs. By contrast, OSS projects are self-organized by IT professionals and many of them offer their service without pay. Some OSS projects also obtain funding from sponsoring companies, but the project performance has much less impact on the participants than with organizational ISD projects because most of the OSS project participants have their own income source through professional employment. CI projects are sponsored and largely funded by government and external agencies. People working on CI projects would get grant funding through the CI projects, which would cover part of the salaries, but also would be funded by their universities or research centers. Project performance to a large extent will affect future funding opportunities, because the external funding agencies will periodically examine the project performance and decide whether to fund the project further. Therefore, project performance does have financial impact on project participants, although in a different way than with organizational ISD projects.

In terms of project participants, the three types of projects are very different as well. Participants of OSS projects and distributed organizational ISD projects are mostly IT professionals and they share similar professional backgrounds, while participants of CI projects are much more diverse in their educational and professional backgrounds (Kirsch and Slaughter 2013a). CI projects involve researchers, scientists, students, professionals and government personnel from locations across the world (Lee et al. 2006). In addition, participants of the three types of projects communicate and connect in different ways. Participants of CI projects are connected with each other through a less articulated goal but they also meet and communicate through conferences, meetings, emails and social media; participants of organizational ISD projects are connected by clear organizational roles and objectives; participants of OSS projects are mostly connected by software dependencies and in the form of online communities.

These three types of projects also feature differences in control mechanisms. Control mechanisms ensure work is carried out according to established plans and criteria, and generally involves two types of control, formal and informal (Orlikowski 1991). Formal control consists of mechanisms that are officially sanctioned and codified, such as written rules and managerial directives, while information control is composed of unwritten mechanisms, such as shared values and norms (Moody et al. 2016). Complex ISD projects are controlled through a portfolio of formal and informal control mechanisms, rather than through a single mode of control (Boland 1979, Henderson and Lee 1992, Gregory et al. 2013). Organizational ISD projects heavily rely on formal control mechanisms to help ensure that project goals, schedules and budgets are met (Kirsch 1997). On the other hand, informal control mechanisms, such as shared values and ideology (Stewart & Gosain 2006), are commonly observed in OSS project communities (Markus 2007). CI projects, like organizational ISD projects, utilize a portfolio of formal and informal controls (Moody et al. 2016). However, many factors make control complicated for CI projects. The high level of uncertainty at the initial stage of CI projects makes it difficult to articulate clear goals. Too much formal control may also stifle the innovativeness of CI projects (Kirsch and Slaughter 2013b). Therefore, CI project managers might supplement formal control with informal control to achieve the transformational goals. However, CI projects typically lack a defined group with shared norms and values when the project commences. The complex social relationships among the participants and the changing project subgroups also make it difficult to nurture stable culture and norms. Therefore, both purposeful formal control as seen in organizational ISD projects, and peer-based informal control as seen in OSS community, seem insufficient for CI projects. A recent study on control for CI projects demonstrates the emergence of a new form of control, "field control" (Moody et al. 2016). The field refers to "all of the individuals or collective entities that subscribe to the general purpose of the project." (p.324). The CI projects in the study exhibited an interchanging focus on field control and authority-based formal control at different phases of CI projects.

Furthermore, the three types of projects feature different coordination mechanisms. OSS projects are merit-based and self-organizing, while organizational ISD projects are coordinated through a variety of mechanisms including project planning, organizational structure and organizational meetings. A CI project utilizes a mixture of coordination mechanisms. On one hand, it features self-organization and choice in collaboration; on the other hand, it utilizes funding allocation, personnel appointment, conferences and meetings as coordination mechanisms (Kirsch and Slaughter 2013a).

The comparison illustrated here highlights an important fact: CI projects, illustrative of a new form of work (Moody et al. 2016), are different from traditional ISD projects in a variety of ways. Managing these projects is not only challenging but also different from managing other ISD projects. Traditional wisdom on project management may not apply well in this new context. The significance and uniqueness of CI projects call for more research attention from IS scholars (Kirsch and Slaughter 2013a). CI projects require creativity to foster exploration and innovation. Structured management techniques may not be suitable for these projects because too many rules and restrictions may possibly stifle the creativity of scientists and researchers. CI project participants have the freedom to self-organize and self-manage in many ways, such as choosing the people to collaborate with and selecting research topics to work on. Meanwhile, due to the large scale, as well as schedule and budget constraints, it is necessary to infuse discipline and standards to keep things under control (Kirsch and Slaughter 2013a). For example, the management team may use funding as a control mechanism to affect future participation of certain projects (Moody et al. 2016). Projects that perform well may be selected to receive further funding while unsuccessful ones may be dropped. A central challenge of CI projects is to provide scientists and researchers an environment with necessary resources and enough freedom to conduct collaborative and exploratory research while keeping things under control, with an end goal to bring about transformational and innovative research findings.

1.3 COLLABORATION IS ESSENTIAL FOR CI PROJECTS

A primary goal of CI is to revolutionize how research is done within a specific domain (Finholt and Birnholtz 2006, Berman 2008). Traditionally, research is conducted by individuals, or by researchers collaborating with a few known and trusted colleagues from similar areas (Atkins 2003). With the advent of the CI, a given academic community can have not only a state-of-the-art platform on which to conduct research, but also a platform that will enable wide-scale collaboration and experimentation both within and across disciplines (de la Flor et al. 2010).

Building a complex digital infrastructure, such as CI, requires collaborative work by groups of people with diverse knowledge backgrounds and skill sets. To illustrate, consider that most scientific research relies on technologies such as high-speed computers or software that supports various calculation and graphical needs. Therefore, domain scientists such as earthquake engineers, who bring content research knowledge and expertise, also need to work with IT professionals, who bring knowledge and expertise related to various technologies. The success of a CI project depends on the collaboration of its stakeholders.

1.4 INDIVIDUAL LEVEL OF COLLABORATION AS A BUILDING BLOCK OF THE CI COMMUNITY

Besides being essential, collaboration in CI projects is also unique. Most people consider that collaboration naturally exists in CI projects. In some sense, this is true because in the majority of cases people have to work together one way or another to complete CI projects. However, another factor is not everyone works with everyone else. Given the diverse backgrounds of the CI project participants, collaboration is dynamic and complex in nature. For example, collaboration depends on people's willingness to share data, and furthermore, cross-discipline collaboration may result in conflict (Kirsch and Slaughter 2013b). Therefore, it stands to reason that people have preferences about possible collaborators. Meanwhile, CI participants have the freedom to choose with whom they want to work; they can self-organize collaboration teams as needed. However, the choice cannot be random because possible collaborators will impact to a large extent the project performance, thus impacting future funding opportunities. Therefore, one question that remains unanswered is why do some people work together while others do not?

Furthermore, it seems likely that people will collaborate with some more frequently than with others by working on multiple projects during a period of time. A CI project features long duration and large scale, and it is common that a CI project is composed of many smaller sub-projects. In addition, each phase of the CI project has its own specific focus and short-term goals. CI participants may collaborate on multiple sub-projects during a specific time period of a CI project, forming a strong collaboration tie. Furthermore, various factors may prompt CI participants to continue their collaboration or drop it in the next project period. Therefore, another question that remains unanswered is why do some people collaborate more often than with others?

It is important to understand why individual project participants collaborate because the individual level of collaboration is the building block of a CI project community. Kirsch and Slaughter (2010) used social network diagrams to illustrate the communication and interaction patterns of a CI project community by linking people who participate in the same events, such as meetings, milestones, and email communications.[3] They also illustrated how a CI project

[3] For details, refer to the Kirsch & Slaughter presentation on GENI Engineering Conference 7, http://groups.geni.net/geni/wiki/Gec7Agenda/

community evolves by linking the people who attend the same conferences using social network diagrams (2013b). Following the similar methodology, by linking the individuals who participate in same projects, I use the following social network diagram to illustrate how the individual level of collaboration is the building block of the CI project community.

Figure 1 is the social network diagram constructed with data from the current study that demonstrates the collaborative relationship among project participants.

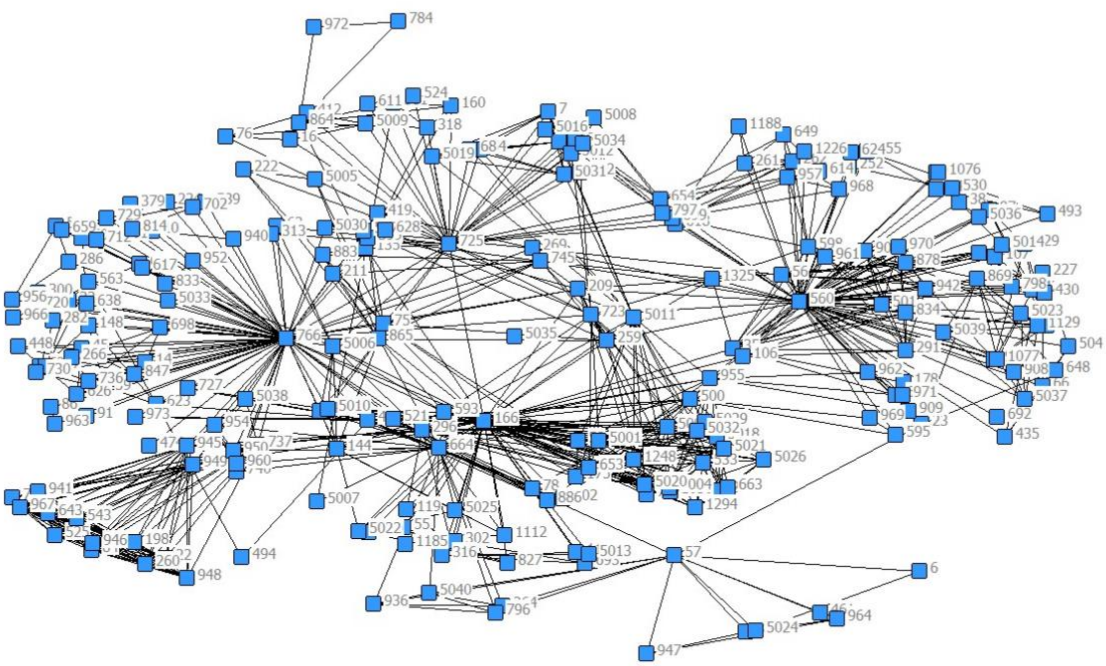

Figure 1. Collaboration Network of CI Project Participants

Each node on this network diagram is a project participant. Each line connects two people on the same project team. Two people may be on multiple projects. The thicker the line, the more projects they work on together. Not only is it intriguing for researchers to understand how this network forms, but understanding the evolution of a project network may also lead to better strategies to promote collaboration. This study intends to find out, as shown in the social network diagram, why the collaboration ties form and why the ties show different thickness. The thickness of the lines represents the dependent variable in this study, collaboration tie strength, or the number of times that two CI project participants work jointly on a project team during a specific time period of a CI project.

In short, collaboration in CI projects is essential and unique at the same time. It not only affects the success of individual project teams, but also the overall success of the CI project. Achieving collaboration is not an easy task given the diverse background of the project team members, their expertise and their distributed locations. Whereas we know a great deal about how collaboration plays out in traditional organizational settings, we know very little about how collaboration plays out in the complex and dynamic context of CI projects. A good understanding of what brings about collaboration is essential for management to better understand the motivations behind the collaboration and therefore utilize tools and methods to promote collaboration. The focus of this study is to examine, at the individual dyadic level, how different factors contribute to collaboration tie strength in CI projects. So one question I address is: *What are the antecedents of the collaboration tie strength between two CI project participants?* I examined the antecedents from both the social and technological perspective, so as my second research goal, I am especially interested in knowing: *how do the social and technological factors interact with each other in predicting collaboration tie strength?*

1.5 OVERVIEW OF CHAPTERS

The next chapters of the study are outlined below. Chapter 2 is the literature review on collaboration, collaboration antecedents and collaboration tie strength. Chapter 3 presents the theoretical model and hypotheses. Chapter 4 describes the methodology, including descriptions of the research site, data collection and measurement. Chapter 5 presents the analysis and results, and finally discussion and conclusion are presented in chapter 6.

2.0 LITERATURE REVIEW

2.1 DEFINITION OF COLLABORATION

The word collaboration originates from the Latin words *com,* meaning *together,* and *laborare,* meaning *to work* (Kotlarsky and Oshri 2005). Merriam-Webster (2014) defines collaboration as "two or more individuals work jointly on an intellectual endeavor."[4] Aram & Morgan (1976) define collaboration in R&D work settings as "the presence of mutual influence between persons, open and direct communication and conflict resolution, and support for innovation and experimentation." (p.1127). They identified three dimensions of collaboration, which include problem solving through support and integration; open, authentic communication; and knowledge-based risk taking. Hardy et al. (2005) defines collaboration as "cooperative, inter-organizational action that produces innovative, synergistic solutions and balances divergent stakeholder concerns." (p.72). These definitions are consistent in nature because they both involve aspects of joint work and common goals.

Meanwhile, collaboration can happen within a group of people on a team (inter-team) or between groups of people on different teams (intra-team). The teams could either be co-located or dispersed. Intra-team collaboration refers to interactions between individuals from two teams through cooperative activities, such as emails, documents and meetings that entail knowledge sharing between the two teams (Banker et al. 2006). In CI projects, both inter-team and intra-team

[4] Retrieved Feb 26, 2014, from http://www.merriam-webster.com/dictionary/collaboration

collaborations are important for the overall project success (Kirsch and Slaughter 2013a). Inter-team collaboration is critical to bring about outcomes to individual project teams, which is part of the larger CI project; while intra-team collaboration is important to bring about at-scale integration of research and experimental findings across project teams. At the initial stage of a CI project, individuals form project teams to work on a specific research topic. Most collaboration happens within a project team. Later on, these project teams need to roll out their findings on a larger scale so they can be integrated with other projects' findings and eventually transitioned to the intended scientific community. Therefore in the later stage of a CI project, intra-team collaboration becomes increasingly more common.

This study focuses on collaboration between individuals. A CI project is composed of many smaller sub-projects or project teams. People from different organizations and institutions choose the people with whom they want to work as collaboration partners. They work on the same project teams to achieve specific project goals in a certain time frame. These project teams often are funded and overseen by government agencies to ensure output is provided on a regular basis. Therefore, once these individuals form a project team, they work jointly to achieve a common goal.

In the context of CI projects, although there are many ways people could work together, such as co-authoring papers, the most common and fundamental way people collaborate is through working on the same project teams because CI projects are the fundamental building blocks of the overall CI project. The fact that two people choose to work on the same project is an expression of collaboration: because CI participants have the freedom to choose with whom they work, the action of forming project teams involves personal choice and a decision to collaborate. Therefore, in this study, I consider that collaboration in the CI context is manifested in individuals working jointly on a project team. The unit of analysis is a *collaboration dyad* consisting of two people working jointly on a project team.

2.2 ANTECEDENTS OF COLLABORATION

Past studies have examined antecedents of collaboration in a variety of project settings and found that many factors may promote collaborative activities.

Earlier work utilized resource dependency perspective to explain why collaboration occurs (Wood and Gray 1991). Specifically, teams or organizations enter into a collaborative relationship to gain access to resources, to obtain efficient use of resources and to form collective rules to govern resource use.

Early research on collaboration found that the major obstacles for successful collaboration are social boundaries and physical distance because they make it difficult to establish shared identity and practice (Levina and Vaast 2005). More often, social boundaries and physical distance are not mutually exclusive. For example, dispersed teams face more challenges to form shared understandings due to reduced frequency of interactions, and thus face more social boundaries with each other. Factors that were proposed to affect dispersed collaboration include, but are not limited to: coordination costs (Boh et al. 2007); social ties (Kotlarsky and Oshri 2005, Boh et al. 2007); mutual knowledge (Cramton 2001); knowledge sharing (Kotlarsky and Oshri 2005, Santoro et al. 2006); cultural and status differences (Levina and Vaast 2008).

Gradually, with the advancing and widely available collaborative technologies, physical distance becomes less of a barrier to collaboration between distributed teams. In fact, collaboration technologies have been found to contribute to the collaboration success for virtual teams (Wainfan and Davis 2004). Computer-mediated collaboration represents a rapidly expanding form of work and is the subject of numerous studies on collaboration (Katz and Te'eni 2007). These technologies include, but are not limited to: calendaring systems, chat rooms, desktop videoconferencing, email, e-meeting systems, information and knowledge repositories, newsgroups, personal information managers, project management systems, telephone conferencing, video whiteboards, and workflow systems (Santoro et al. 2006). In recent years, a wiki is widely used in companies or teams as a collaboration technology to facilitate conversational knowledge creation (Wagner 2004, Wilkinson and Huberman 2007).

With the growing popularity of the virtual environment, another strain of research looks at the collaboration in virtual teams. Trust has widely been identified as a predictor for effective collaboration among virtual teams (Brown et al. 2004, Hossain and Wigand 2004, Peters and Manz 2007). Furthermore, different forms of trust (self-interest, ability, empathy and integrated) interact and influence the performance of virtual collaborations (Paul and McDaniel 2004). Besides trust, other factors influencing collaboration success include interpersonal traits, complementarity (Brown et al. 2004), as well as relationship depth (Peters and Manz 2007).

Another important aspect of collaboration is its sustainability. Studies have suggested that past collaboration experience influences future collaboration decisions. For example, Schwab and Miner (2008) studied collaboration in the American movie industry and identified prior collaboration success, project relevance and organizational control as important predictors for repeated and future collaborations.

The aforementioned antecedents in extant literature broadly fall into two categories: social factors and technological factors. Given the unique characteristics of CI projects illustrated in the previous chapter, in combination with the findings of extant research, this study will examine the antecedents of collaboration tie strength that are relevant to the research context of CI projects from both the social and technological perspectives. Resource dependency, technical dependency, social concerns such as shared identity, interpersonal relationships, and prior collaboration will all be included in the study. However, the study did not include spatial and temporal factors because as stated in the research, with today's advancement in technologies that support activities of virtual teams, physical distance is less of an obstacle for collaboration. In the context of CI studied in this research, the participants are all familiar with virtual technologies and heavily rely on technology to collaborate.

Another factor not included in the study is trust. The major reason for excluding trust as one of the antecedents is that some antecedents included in the study are directly relevant to trust, making it an intermediary variable, not included in the proposed model. For example, extant research suggested that past collaboration experience, shared identity, and interpersonal relationships may all affect trust (Brown et al. 2004), which may in turn affect the collaboration choice (Paul and McDaniel 2004). Although trust is not directly included in the study, some of the important antecedents of trust suggested by past research will be included.

2.3 COLLABORATION TIE STRENGTH

Past literature examined collaboration from a variety of dimensions, including collaboration choice, collaboration degree and collaboration success. Collaboration choice pertains to who is selected as the collaboration partner. Collaboration degree could be examined from a variety of aspects, such as frequency, intensity and sustainability (Brown et al. 2004). Collaboration success pertains to both the subjective and objective outcomes brought about by collaboration. For example, previous studies have found that teams that collaborate effectively are more innovative, productive and satisfied than teams that do not collaborate (Schrage 1990, Peters and Manz 2007). Previous studies have not only used product success and personal satisfaction (Kotlarsky and Oshri 2005), but also performance improvement (Paul and McDaniel 2004) to measure collaboration success.

The concept of interest to my study, *collaboration tie strength*, focuses on the aspect of collaboration frequency. The unit of analysis is a dyad of collaborators. In network studies, tie strength is a concept used to describe the significance of intensity of a relationship (Sosa 2010, Aral and Walker 2014). It is a complex construct that involves the multiple dimensions such as relationship duration, communication frequency, emotional intensity and emotional closeness (Granovetter 1973, Marsden and Campbell 1984, Burt 1992). Network studies use various dimensions of tie strength mentioned above (e.g. Hansen 1999, Reagns and McEvily 2003, Sosa 2010, Aral and Walker 2014). Empirically, some dimensions of tie strength, such as communication frequency and emotional closeness, are often correlated (Hansen 1999; Reagans and McEvily 2003; Tortoriello 2012). When the intensity and duration are similar for dyadic pairs, collaboration frequency is equivalent to tie strength, as seen in the study of McFadyen et al. (2008) in which the average tie strength is measured by the frequency of interaction between two knowledge partners.

Tie strength in network studies is used to describe a variety of relationships. Some are social connections between people, and some are professional or work related connections. In this study, I am interested in the connections people establish to collaborate on projects, which are work related. To be accurate about the nature of the tie strength under study, I use collaboration tie strength to describe the number of times that two individuals work on the same project team. Collaboration here indicates that the two people form a collaborative relationship through working on the same project team, and tie strength captures the significance of such a collaborative relationship by forming the relationship multiple times. In network research, tie strength is a complex construct that involves the duration, frequency and intensity of a relationship. In this study, the project duration is the same for all projects, which lasts one year. To address the concern on the variance in project size, collaboration tie strength in this study is weighted by the project size. Therefore, holding the duration and intensity constant, the more projects two people work on together, the higher the collaboration tie strength; this conceptualization adopted in this study is similar to the conceptualization of tie strength in the work by McFadyen et al. (2008).

A question central to this study is *what are the factors that predict collaboration tie strength in CI projects*? To answer this question, this study examines the collaboration tie strength of CI participants at the dyadic level. One important characteristic of CI projects is the complexity of the project participants because they work in different organizations and have various educational and cultural backgrounds. Meanwhile, CI project participants are connected because most of them are part of the academic or scientific communities. Furthermore, CI project participants are different or similar in a variety of ways, such as age, ethnicity, education and other demographic factors. What promotes the collaboration between these comparatively independent yet connected people? How do the various social and technological factors relate to their collaboration patterns? Examining the relationship from a dyadic level provides a fine-grained view on the behavior of project participants and helps disentangle more precisely the effects among factors pertaining to a relationship such as similarities, differences and shared experiences of two people.

CI projects have their own unique features that set them apart from traditional ISD projects. This calls for a customized theory that takes into consideration their special characteristics in terms of structure and resource allocation. It is not clear how the dynamics of CI projects and the complexity of social relationships of project participants play out in collaboration. In addition, although past research has suggested that resource dependency, technical dependency, knowledge dependency, social factors and prior collaboration experience play an important role in predicting collaboration, it is not clear how these factors integrate together in explaining collaboration. Past network studies focus mostly on the social relationships between people and often use tie strength as a predictor; however few studies have examined what predicts tie strength and even fewer have examined collaboration tie strength in the complex environment of CI projects. In this study, I ground my proposition in considerations of resource dependency, technical dependency, knowledge dependency and social attractiveness, and aim to explain collaboration tie strength from the technological and social perspectives, examining each aspect with CI project related variables. In the following section, I will elaborate on the research model and the theoretical backgrounds.

3.0 RESEARCH MODELS AND HYPOTHESES

The theoretical framework for my study leverages the social and technological view of information systems to integrate the extant research findings with specific considerations for CI projects. In this study, I examine the dyadic collaborative relationship between two CI project participants. Here, CI project participants refer to individuals who participate in the overall CI project through becoming a member of any project teams. In particular, I incorporate the technological considerations of CI projects with social considerations to explain the factors that predict collaboration tie strength. Figure 2 illustrates my theoretical model.

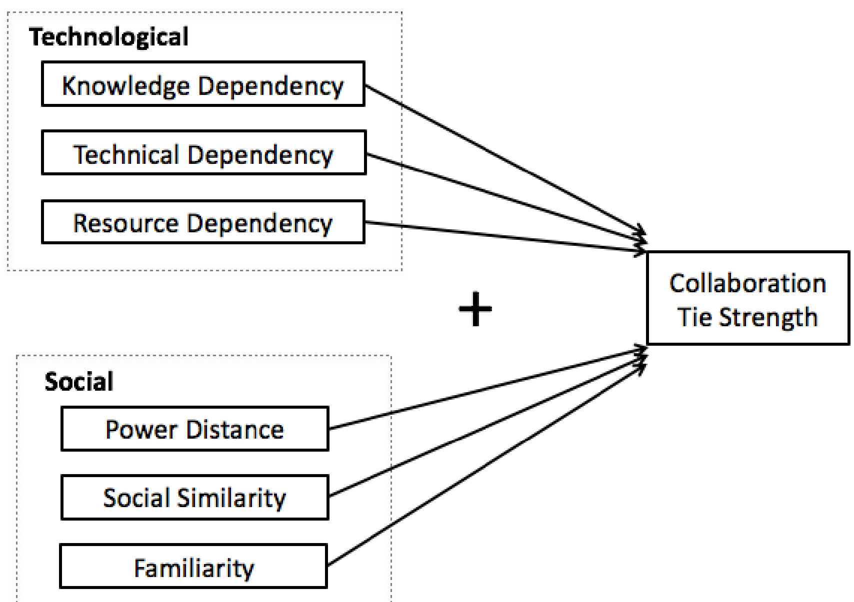

Figure 2. Theoretical Model (Main Effects)

3.1 A SOCIAL & TECHNOLOGICAL VIEW OF COLLABORATION

A long tradition in IS research suggests that IT phenomena can be examined from two aspects: social and technological (Bostrom and Heinen 1977, Trist 1981). I consider this view suitable for studying CI projects because these projects are both technologically intensive and socially complex. On one hand, technology is inseparable from CI projects because it not only provides tools and resources for conducting research, but also facilitates the sharing of ideas and knowledge, as well as the management of projects by coordinating individual behaviors. On the other hand, CI projects are socially complex because they not only involve geographically distributed experts from various walks of life but also are organized in a unique way that attempts to balance discipline and freedom in order to encourage creativity and innovation (Kirsch and Slaughter 2013a). The collaboration among the participants of CI projects may be largely influenced by both technological and social considerations. Therefore, in this study, I examine the dyadic collaborative relationship from these two perspectives. The unit of analysis is the collaborative dyad between individuals on the same project team. In general, I propose that collaboration tie strength is dependent on the technological needs and social attractiveness between the two individuals. I ground the technological needs perspective in the concepts of knowledge dependency, technical dependency, as well as resource dependency, and I ground the social attractiveness perspective in dimensions derived from social influence theory, i.e., power distance, social similarity and familiarity.

As my second goal of this study, I am specifically interested in knowing how the social factors and technological factors interact with each other in predicting collaboration tie strength, as illustrated in Figure 3.

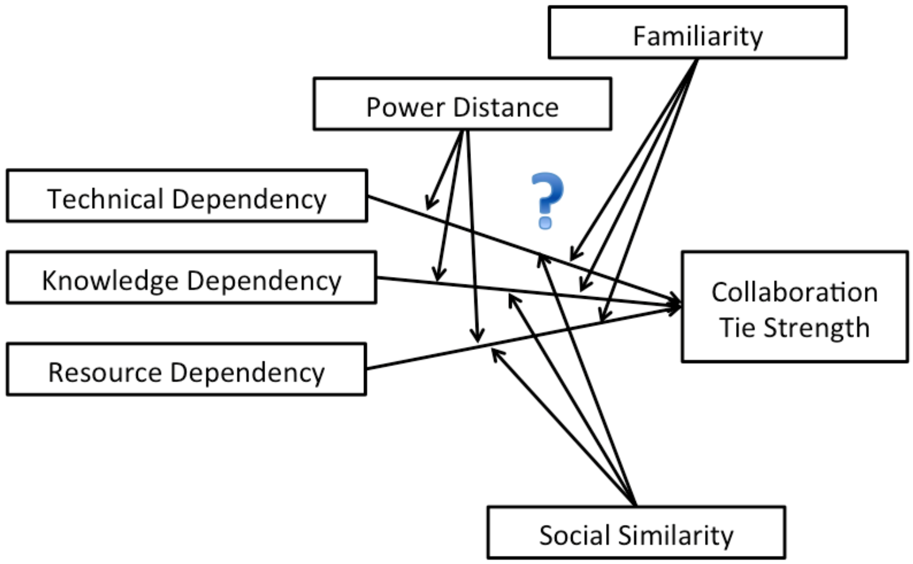

Figure 3. Theoretical Model (Interaction)

3.2 TECHNOLOGICAL NEEDS

First, I examine collaboration tie strength from the perspective of technological needs. In contrast to the social perspective that will be discussed in later sections, technological needs reflect a variety of dependencies between CI participants related to project development. I examine technological needs from three perspectives: knowledge dependency, technical dependency and resource dependency.

3.2.1 Knowledge dependency

The end goal of scientists and researchers entering a collaborative relationship is to transfer or create knowledge (Tortoriello et al. 2012). Past research has found that a certain level of knowledge overlap is essential for scientific researchers to form an effective collaborative relationship (Polanyi 1966). As Kuhn (1962) asserted, overlapping knowledge is important in the pursuit of “normal science.” People with overlapping knowledge, when interacting with each other, make it possible for knowledge to converge and combine, and for new ideas to emerge (McFadyen et al. 2008). Such convergence is extremely important at the early stage of knowledge creation because it bridges knowledge between collaborating partners (Nonaka 1994, McFadyen et al. 2008). Therefore, this study posits that knowledge dependency is an important predictor for collaborative relationships.

Knowledge can be either tacit or explicit (Nonaka 1994). Tacit knowledge, sometimes referred to as “know-how” (Brown and Duguid 1998), is mostly based on experience and can be difficult to describe (Polanyi 1966), while explicit knowledge is codified, in tangible form and can be easily stored and retrieved (Wellman 2009) in either paper or digital form. For the purpose of this study, I focus on explicit knowledge because it is more objective to measure and evaluate.

People demonstrate knowledge dependency in a variety of ways. For example, if a person seeks answers to a question by posting it to a general forum and someone responds to it, it is reasonable to believe that the former has a knowledge dependency on the latter.

Hypothesis 1A: Collaboration tie strength is positively related to knowledge dependency between two CI project participants.

3.2.2 Technical dependency

Technical dependency is a concept broadly used in project management, especially for large complex projects. The idea behind technical dependency is that in order to successfully complete a project, it requires different people working on different parts of the project in a specific order within a certain time frame and budget. Because of such requirements, a project participant is not a stand-alone worker, but rather, an inseparable part of a larger group. They all, to a certain degree, technically depend on others' work to carry out their tasks. Of course, such technical dependency has multiple facets.

One facet concerns task execution, i.e., the execution order of different task components. As noted in prior research, distributed teams working collaboratively on complex tasks need to resolve task component dependency issues (Espinosa et al. 2007a). The idea is that in carrying out projects, people need to perform various tasks (Boh et al. 2007), and many tasks are related to each other. The action needed to complete one task may be dependent on the status of another. For example, whether to start a task is dependent on the successful completion of another, or how to carry out a task is dependent on the outcome of another. The concept of task dependency is closely related to the nature of the task and implies sequence of task execution. Task dependency could be caused by an established task schedule or scarce shared resource (Espinosa et al. 2007b). In project management, people use different ways to express task dependency, such as Gantt charts, network diagrams or calendars. A critical step in project management is to identify the relevant tasks and then create links or dependencies among these tasks. A similar concept in software engineering is software dependency. For example, in software engineering, the outcome – software – is generally designed as interacting components or modules (Shaw and Garlan 1996).

Task dependency is often simplified to: it only exists when one task cannot be carried out unless another task is finished. However, in project management, that view is far from sufficient; in fact, many other task dependency types exist. A popular method to examine task dependency is to look at the timing required by the two tasks using the Finish-Start method, in which “F” represents “Finish” and “S” represents “Start,” as expressed in Table 2.

Table 2. Task Dependency Types

Task Dependency Types	Examples
FS – T2 cannot start until T1 is finished.	Coding must wait on design to be finished
SS – T1 and T2 must start at same time.	Documentation and coding start together
FF – T1 and T2 finish at same time.	Configuration management ends when testing is done
Lag – T2 cannot start until a given amount of time after T1 is done.	Start on-site training one week after final demo to customer management group.
Soft – T2 should start after T1, but it is not completely essential that T1 be finished.	Populate the new database after the database has been built. If the new database is not completely finished we can still populate the few finished tables.

Suppose in a CI project that aims to build a digital infrastructure to facilitate atmospheric and oceanic research, an earthquake engineer, after coming up with a theoretical model and designing the algorithms, needed people with strong programming skills to translate the theoretical model into computer programs. This engineer decides to form collaboration with a computer programmer. We could say that in this case, the earthquake engineer and computer programmer collaborate because of task dependency. The programming and algorithm designing forms a Finish-Start (FS) type of task dependency because the programmer cannot start programming until the engineer completes the designing of the algorithm, and further, the engineer cannot process the data and carry out the analysis until the programmer finishes programming. This type of dependency is a form of strong task dependency, with intensive collaboration between the engineer and programmer.

The above outlines one aspect of technical dependency, which pertains to task execution. Another important facet of technical dependency pertains to the integration of different task components. This is a prominent issue in large complex software development projects. When a group of people works on different parts of a project, in order for the end product to function properly, different parts of the end product need to eventually integrate with each other; therefore, each individual's work is technically dependent on another's work. Such technical dependency is extremely important when planning complex projects, such as CI projects. As described earlier, the end goal of a CI project is to build a digital infrastructure that enables scientists and researchers to conduct innovative and transformative research. In order to achieve this goal, each project team, while working on their own portions of the CI project, must keep the broad picture in mind because later on, the technologies developed by these project teams will need to integrate with each other and roll out to the meso-scale. Eventually, the end product needs to be transitioned to the intended user community. Collaboration with other project teams is an important way to achieve such integration due to technical dependency. This study posits that the more a person is technically dependent on another, the more likely these two individuals will collaborate.

Hypothesis 1B: Collaboration tie strength is positively related to technical dependency between two CI project participants.

3.2.3 Resource dependency

Past research found that people enter into a collaborative relationship to obtain resources (Wood and Gray 1991). Resource dependence theory originates from organizational research and states that organizations obtain power by rationally adapting to changes in the external environment (Pfeffer and Salancik 1978, Ulrich and Barney 1984). The theory posits that actors seek to establish relationships with others who can provide needed resources, i.e., anything perceived valuable (Tillquist et al. 2002).

By entering a relationship with someone who possesses the needed resource, a person either formalizes the access to, or takes possession of, the resource (Tillquist et al. 2002). For example, an earthquake engineer has a research idea about how to identify the location of a specific type of earthquake using oceanic and atmospheric data; however, the earthquake engineer does not have the access to such data. Therefore, the engineer enters into a collaborative relationship with an oceanic scientist who knows how to obtain the needed data. Overall, to carry out the research, the earthquake engineer may find it necessary to obtain different resources from a variety of people, such as high-speed computers and software tools to do the analysis.

Due to its complex nature, a CI project requires the effort and input from diverse groups of people; it is common for CI project participants to enter into a collaborative relationship to obtain needed resources to carry out project tasks. The more resources a person needs from others, the more this person's work is dependent on others, and the more likely this person will collaborate with the others. When this translates to a dyadic collaborative relationship, it means that the more relevant resources two people can provide to each other, the more the two will collaborate. A relevant resource is defined as a resource that is possibly needed and useful to the other's work.

Hypothesis 1C: Collaboration tie strength is positively related to the resource dependency between two CI project participants.

3.3 SOCIAL ATTRACTIVENESS

In the sections above, I examine collaboration tie strength from the technological perspective. However, past research on project management has found that most project problems are social instead of technological (Markus 1984, Brooks 1995), suggesting that the social perspective is also relevant in studying collaboration in CI projects. Social factors are broad and complex. Network research has suggested that a variety of social factors may affect tie strength. For example, people who have known each other for a long time frequently communicate and feel emotionally attached to each other, and tend to form strong ties (Marsden and Campbell 1984, Reagans 2011).

Here, I ground the social perspective of collaboration in social influence theory (Kelman 1958), using this theory to examine collaboration tie strength between CI project participants. Each project team is composed of individuals who are situated in the broader CI project community. They interact with and influence each other in this complex ecosystem. It is logical to posit that collaboration is, to a certain degree, influenced by their social relationships. In this study, I aim to find out what influences people's collaboration behaviors, or what factors predict the collaboration behaviors between pairs of CI participants. I also strive to operationalize each construct with measurements relevant to the CI context, based on their underlying theoretical meanings. These ideas are explored below.

In sociology, social influence refers to the processes whereby people directly or indirectly influence the thoughts, feelings and actions of others (Allport 1985, Turner 1991). This theory is widely used to examine how one person is socially influenced by another in making decisions or taking actions. This study draws from social influence theory to examine a person's collaboration choice and behavior as a result of social influence.

Early research on social influence follows the categorization of normative influence and informational influence (Deutsch and Gerard 1955). Normative influence happens when people agree outwardly but may not inwardly. Informational influence is based on knowledge and subjective validity and leads to internal agreement in private and long-lasting attitude change. Kelman (1958) later provides a finer view of social influence and posits that there are three main types of influence processes demonstrated in social relationships: compliance, based on power and status; identification, based on social similarity; and internalization, based on information and knowledge. Turner (1991) integrates the above two theories and suggests that compliance and identification fall into the category of normative influence, while internalization falls into the category of informational influence.

In this study, I examine collaboration tie strength using the social influence categories outlined by Kelman (1958), i.e., compliance, identification and internalization. Based on his theory about social influence, I consider three categories of social factors that may affect the collaboration choice and behavior of CI project participants: power distance, social similarity and familiarity.

3.3.1 Power distance

First of all, drawing from the concept of compliance in social influence theory, two people may collaborate because they are attracted to each other on the basis of power and status.

According to Kelman (1958), an important factor that may influence a person's decisions is compliance. Compliance occurs when one is influenced because of power and status. It happens when a person adopts certain behavior "not because he/she believes in its content but because he/she expects to gain specific rewards or approval and avoid specific punishments of disapproval by conforming" (Kelman 1958, p. 53). The word "conforming" indicates that when influenced by compliance, a person adopts a certain behavior because of external forces such as power and status, rather than internal belief or possessed knowledge.

This reasoning may apply in the context of a CI project, in that, if a person has more power or holds a higher level of position or social status in a CI project, he/she may have more collaboration opportunities. Social status in the context of a CI project is mostly demonstrated in two forms. One is the power relationship produced by a work relationship, commonly referred to as a "role" on a CI project. A person who holds a management or team lead position, normally known as a principal investigator (PI), has a higher social status than a regular project worker. PIs have more access to resources or personnel on other teams. Therefore, it is reasonable to believe that PIs tend to have more collaborative opportunities than others. The other type of relationship in a CI project that indicates power and status could be the advisorship or sponsorship developed in the academic world. In such a relationship, an academic advisor is considered to have a higher level of power and status than his/her student. Therefore, the study also posits that people collaborate because one person was another's academic advisor. For the purpose of this study, I use power distance to represent the difference in power and status between two CI project participants, different from the definition by Hofstede (1983). I hypothesize that:

Hypothesis 2A: Collaboration tie strength is positively related to power distance between two CI project participants.

3.3.2 Social similarity

Drawing from the concept of identification in social influence theory, two people may collaborate because they are attracted to each other due to social similarity.

In social influence theory, identification suggests that when people put themselves into the same category, they are more likely to be influenced by each other. A person adopts a certain behavior because of desired relationship, or “classical relationship” (Kelman, 1958, p.53). Here, classical means belonging to a class or category. In other words, when influenced by identification, a person puts himself/herself into the same class as another, or takes the “role” (Kelman 1958, p.53) of another. Such classification is based on similarity, which could be similarity in demographic characteristics such as age and gender, or similarity in some implicit features such as cultural and educational background. Identification and compliance are the same in that both are caused by conforming to achieve desired relationship, thus both are normative influence (Deutsch and Gerard 1955). Meanwhile, they are different in that identification is based on similarity while compliance is driven by power difference.

Applied to the study on collaboration choice and behavior, the idea of identification is attractiveness due to similarity, i.e., people are attracted to work with people who have similar characteristics, such as age, major, gender, cultural background and location (Turner 1991). In the context of a CI project, people collaborate with certain people because of desired relationship based on similarity. This suggests:

Hypothesis 2B: Collaboration tie strength is positively related to the social similarity between two CI project participants.

3.3.3 Familiarity

Drawing from the concept of internalization in social influence theory, two people may also collaborate because they are attracted to each other based on familiarity.

As outlined in social influence theory, internalization is considered a type of "true" influence. When a person finds the information convincing and subjectively valid, this person is more likely to be influenced (Turner 1991). Therefore, it falls into the category of informational influence (Deutsch and Gerard 1955). Internalization differs from compliance and identification because it is based on internal belief and knowledge rather than on external factors such as power and relationship and social similarities. With internalization, a person adopts a certain behavior because "it is congruent with his value system" and "the satisfaction derived from internalization is due to the content of the new behavior" (Kelman 1958, p53).

The concept of internalization, when contextualized in CI projects, is in fact a form of trust in each other because they know or are familiar with each other. In a CI project, there are numerous ways for participants to become familiar with each other. For example, two people who graduated from the same school tend to know more about each other; familiarity could be developed if the two people share the same academic advisor. Another possible venue of two people developing familiarity is through past experience of working on the same project teams. On the surface, attending the same school or having the same academic advisors is a type of similarity. However, we should not confuse familiarity suggested by internalization with similarity suggested by identification because the former contributes to deep knowledge about each other while the latter does not. For example, having similar age or skin color does not equate to deep knowledge about each other.

To summarize, in social influence theory, internalization suggests that a person adopts a certain behavior because he/she believes doing so is correct based on his/her knowledge. In the context of a CI project, people choose to work with people who are helpful and competent in completing the project based on their knowledge about each other. This knowledge about each other is different from the knowledge dependency outlined previously, because knowledge dependency pertains to common knowledge on projects, whereas the knowledge outlined here is, from a social perspective, referring to the knowledge about people obtained through social interactions such as past acquaintances, work relationships or collaboration experiences. It is referred to in this study as familiarity. I hypothesize that the more familiar the participants are with each other, the more likely they are to collaborate.

Hypothesis 2C: Collaboration tie strength is positively related to the familiarity with each other between two CI project participants.

3.4 TECHNOLOGICAL NEEDS & SOCIAL ATTRACTIVENESS

Finally, technological needs and social attractiveness may interact with each other in predicting collaboration tie strength. In other words, the effect of social attractiveness on collaboration tie strength may depend on the level of technological needs or vice versa.

It is important for us to be aware of the complexity involved in the interactions between social and technological factors. As noted previously, this study adopts a social-technical view to examine collaboration in a CI project because a CI project is not only technology intensive but also socially complex. The social-technical view recognizes that both social and technological components coexist in an information system and this view has been broadly adopted to study IT phenomena or organizations with embedded IT artifacts, e.g., software networks (Bird et al. 2016), software development (Cataldo 2008, Tsay 2014), communication (Fairhurst), quality management (Manz 1997) and organizational groups (Manz 1978). Similarly, the current study follows the socio-technical framework and proposes a main-effect model that examines both the social and technological antecedents of collaboration.

The social-technical view, however, has a deeper implication for IS research, which goes beyond simply incorporating multiple facets when studying an IS phenomenon. In a complex social-ecosystem such as a CI project, interactions between social factors and technological factors become very complex and hard to predict. First, people may interpret the importance of technology, resource and knowledge differently when provided with different social choices. As noted earlier, the project participants have the freedom to choose with whom to work. When faced with a different combination of choices, people may come to different decisions to optimize the social and technological preferences. Second, the interaction between social factors and technological factors may be dynamic. Past research has pointed out that the social and technological are entangled and inextricably related, and they shape and reshape each other in their interactions (Leonardi 2012, Winter et al. 2014). Furthermore, technological artifacts are never stabilized and complete since they change and evolve over time, due to people's perceptions about them changing and evolving (Orlikowski 2000). Entities in a socio-technical system not only interact with each other, but they continuously "perform in a web of relations" (Cecez-Kecmanovic et al., p.809). Therefore, it is important to consider the possible dynamics involved with the interactions between social and technological factors when we study collaboration in a CI project.

It is not clear though how the social and technological factors interact with each other in predicting collaboration tie strength. Do they complement the effects of each other or substitute the effects of each other? If they complement each other, to what degree do they accentuate the prediction? If they substitute for each other, do they completely cancel out their main effects? No prior research, to the best of my knowledge, provides a clear answer to these questions. Past research examined how knowledge sharing and social ties together contribute to collaboration (Kotlarsky & Oshri 2005), but whether interactions exist between these factors has not been specifically investigated. Therefore, this study attempts to explore these interactions to unpack the relationships among technological and social factors.

There are two possible ways for the interaction, one being that they are complementary to each other in predicting collaboration tie strength, i.e., social attractiveness accentuates the prediction of collaboration tie strength by technological needs, the other being that they are substitutive to each other in predicting collaboration tie strength, i.e., social attractiveness suppresses the prediction of collaboration tie strength by technological needs.

In reality, both arguments may stand true. For example, when a person is more familiar with another, whether the other party possesses the needed resource, technology or knowledge may not matter much in making collaboration decisions. In this case, familiarity and technological factors are substitutive to each other. However, it may also be true that when a person is more familiar with another, it matters even more whether the other person could provide the needed resource, technology or knowledge. They may be demanding on what the other party could provide because they are familiar with each other. In this case, familiarity and technological factors are complementary to each other in predicting collaboration tie strength.

Competing arguments may also exist for the interactions between similarity and technological needs. On one hand, people may prefer working with those who are similar to them so that at a higher level of similarity it does not matter as much whether the other party could meet their technological needs. In this case, similarity and technological factors are substitutive to each other in predicting collaboration tie strength. On the other hand, it could also be true that people have preset expectations for people who are similar to themselves or they have prejudice about what the other party could do based on the similarity, causing them to demand more technological resources than the other party can provide. In this case, similarity and technological needs are complementary to each other in predicting collaboration tie strength.

Finally, competing arguments may also apply to the interactions between power distance and technological needs. Past research has demonstrated both substitutive (Kirkman 2009) and complementary (Wei et al 2016) effects of power distance in studying leadership. However, it is unclear what kind of interactions power distance may have with technological factors in CI projects because it could go both ways. For example, when the power distance between two people is high, a person may choose a party with more power because of the convenience and access this party can provide to complete projects, to such an extent that less importance is placed on whether the other party can provide the needed knowledge, resource or technology. In this case, power distance and technological needs are substitutive to each other in predicting collaboration tie strength. On the other hand, it is also possible that a higher power distance makes it even more important for the other party to possess the needed resources to be considered a collaboration partner. A person may feel pressured and constrained when working with someone with more power, and therefore may be more demanding on the resource, technology or knowledge the other party could provide in order to consider the other party as a collaboration partner. In this case, power distance and technological needs are complementary to each other in predicting collaboration tie strength.

Based on the above reasoning, I present both the hypotheses for the substitution effects between social and technological factors and the alternate hypotheses for the complementary effects between social and technological factors.

Hypothesis 3A: Knowledge dependency and power distance negatively interact with each other, so that the increase in power distance weakens the prediction of knowledge dependency on collaboration tie strength.

Hypothesis Alt-3A: Knowledge dependency and power distance positively interact with each other, so that the increase in power distance strengthens the prediction of knowledge dependency on collaboration tie strength.

Hypothesis 3B: Knowledge dependency and similarity negatively interact with each other, so that the increase in similarity weakens the prediction of knowledge dependency on collaboration tie strength.

Hypothesis Alt-3B: Knowledge dependency and similarity positively interact with each other, so that the increase in similarity strengthens the prediction of knowledge dependency on collaboration tie strength.

Hypothesis 3C: Knowledge dependency and familiarity negatively interact with each other, so that the increase in familiarity weakens the prediction of knowledge dependency on collaboration tie strength.

Hypothesis Alt-3C: Knowledge dependency and familiarity positively interact with each other, so that the increase in familiarity strengthens the prediction of knowledge dependency on collaboration tie strength.

Hypothesis 3D: Technical dependency and power distance negatively interact with each other, so that the increase in power distance weakens the prediction of technical dependency on collaboration tie strength.

Hypothesis Alt-3D: Technical dependency and power distance positively interact with each other, so that the increase in power distance strengthens the prediction of technical dependency on collaboration tie strength.

Hypothesis 3E: Technical dependency and similarity negatively interact with each other, so that the increase in similarity weakens the prediction of technical dependency on collaboration tie strength.

Hypothesis Alt-3E: Technical dependency and similarity positively interact with each other, so that the increase in similarity strengthens the prediction of technical dependency on collaboration tie strength.

Hypothesis 3F: Technical dependency and familiarity negatively interact with each other, so that the increase in familiarity weakens the prediction of technical dependency on collaboration tie strength.

Hypothesis Alt-3F: Technical dependency and familiarity positively interact with each other, so that the increase in familiarity strengthens the prediction of technical dependency on collaboration tie strength.

Hypothesis 3G: Resource dependency and power distance negatively interact with each other, so that the increase in power distance weakens the prediction of resource dependency on collaboration tie strength.

Hypothesis Alt-3G: Resource dependency and power distance positively interact with each other, so that the increase in power distance strengthens the prediction of resource dependency on collaboration tie strength.

Hypothesis 3H: Resource dependency and similarity negatively interact with each other, so that the increase in similarity weakens the prediction of resource dependency on collaboration tie strength.

Hypothesis Alt-3H: Resource dependency and similarity positively interact with each other, so that the increase in similarity strengthens the prediction of resource dependency on collaboration tie strength.

Hypothesis 3I: Resource dependency and familiarity negatively interact with each other, so that the increase in familiarity weakens the prediction of resource dependency on collaboration tie strength.

Hypothesis Alt-3I: Resource dependency and familiarity positively interact with each other, so that the increase in familiarity strengthens the prediction of resource dependency on collaboration tie strength.

4.0 METHODOLOGY

4.1 RESEARCH SITE

The research site for my study is the Global Environment for Network Innovation (GENI) project. The goal of the GENI project is to establish a virtual laboratory for exploring a future Internet at-scale.[5] In this virtual laboratory, users can perform at-scale experimentation and do deep programming on shared and heterogeneous GENI resources, and conduct innovative networking and distributed systems research and education.

The people who build GENI are mostly network engineers and computer scientists, but they also include researchers, educators, students and industry professionals from different institutions and organizations. These nearly 1,750 participants come from multiple knowledge domains with different needs and requirements.

By the time of this study, the GENI project had completed the pre-planning and planning phases, and the construction phase was ongoing. Pre-planning consists of events before 2004, by which time NSF agreed to support the GENI project and hired key individuals to drive the initiative. The planning phase began in 2004 and lasted until 2008. In this phase, groups of researchers and sponsors worked together to shape the idea of the GENI project in terms of its vision, goals and organization. The construction phase commenced in 2008. This phase featured

---

[5] For more information, see www.geni.net.

development activities to build the specific CI technologies in different GENI technical clusters and to roll out the GENI projects to a larger scale.

The management team is called the GENI Project Management Office (GPO). It adopted a spiral development approach (Boehm 1986). The GENI project went through several spirals, with each spiral involving steps of a complete development project lifecycle, i.e., planning, design, build, integration and use, as shown in Figure 4. A spiral began with the GPO setting out goals and allocating funds to different project teams. Each spiral ended after one year of a complete project development cycle. At that point, the performance of these projects was reviewed and the evaluation results became, in part, the basis for funding decisions of the next spiral. In this way, the GENI project provided flexibility needed for CI projects by retaining promising/successful projects and carrying them over to the next spiral, while dropping unpromising/unsuccessful projects from the overall development. Some projects deemed critical might have been kept for the next spiral regardless of the performance due to their strategic importance.

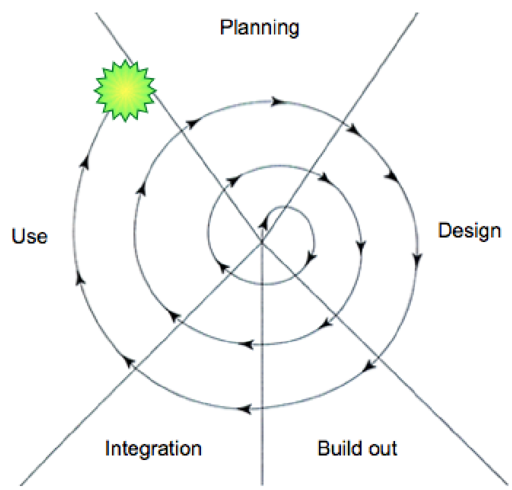

Figure 4. Ongoing Spiral Development and Prototyping (www.geni.net)

The GENI Engineering Conferences (GECs) served as an important mechanism for implementing the spiral project management approach. These conferences were held quarterly in different locations in the United States. Most locations were close to, or on the campuses of, universities participating in the GENI project. Stakeholders from current projects were required to attend the conferences. Major project decisions and plans were discussed and announced at the GECs. A variety of other activities happened at these conferences, which included, but were not limited to: GENI management teams providing project status updates, major project teams doing demonstrations of project prototypes, technical experts giving tutorials to end users, different project teams exchanging ideas and brainstorming to address current issues, and industrial partners hosting panel sessions to discuss key technologies.

At the time of this study, the construction phase included four spirals (1 to 4) and nine GECs (3 to 17). The GENI project has completed spiral 4, and three additional spirals are anticipated to finish building the GENI infrastructure.

The GENI development process described above is a spiral development approach. Next, I focus on the project structure in each GENI project spiral. Overall, the GENI project was divided into several project groups, with each group focusing on a specific type of technology. In the earlier spirals (1 to 3) these project groups are also called clusters. For example, spiral 1 is composed of five clusters (A, B, C, D and E). Each project group was further composed of smaller project teams as the working units. These project teams conduct a variety of innovative networking and distributed systems research in the overarching CI environment that provides shared data and networking tools. The project teams vary in size, but mostly involve five to 10 people. They have different foci, with some focusing on developing networking programs, some on wireless technologies and others on infrastructure development. Altogether, as of spiral 4, there are about 126 project teams.

Figure 5 depicts the overall structure of the GENI project.

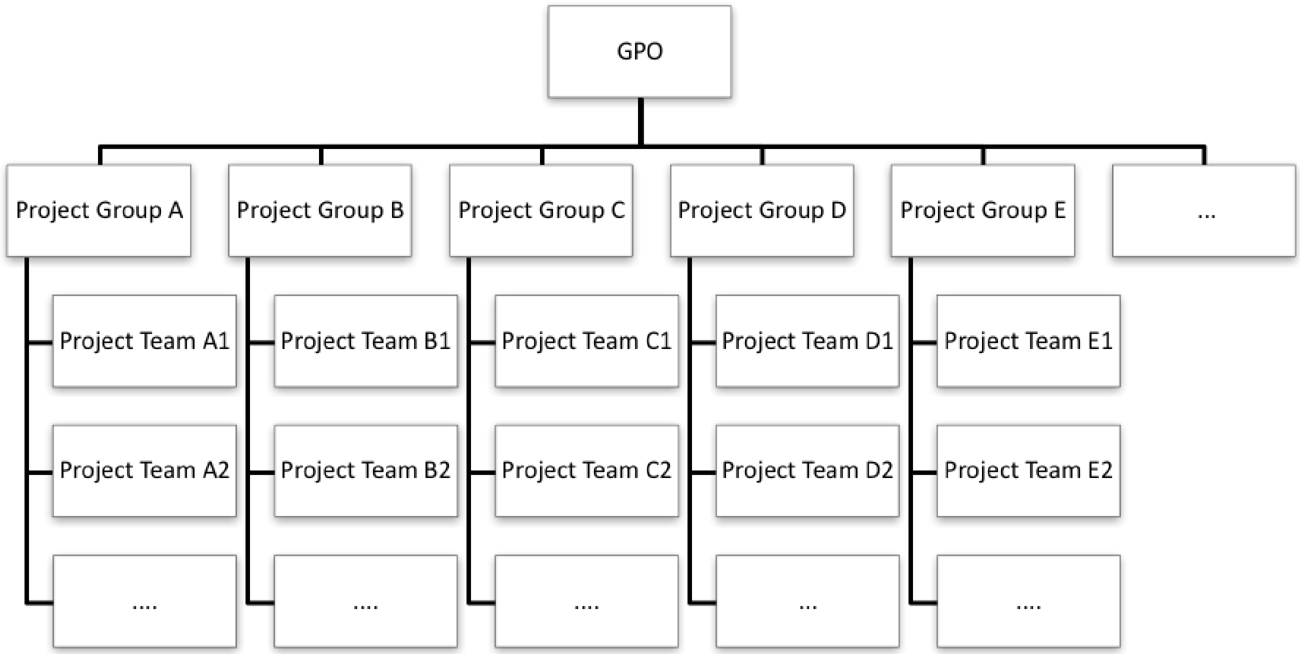

Figure 5. GENI Project Structure

Figure 6 further illustrates the structure by adding the CI project participants to the graph. People participate in the GENI project through forming project teams and working on projects. Some people may work on several projects, such as participant A (working on project teams A1, B1 and B2), participant B (working on project teams A1 and A2), and participant C (working on project teams A1 and B1).

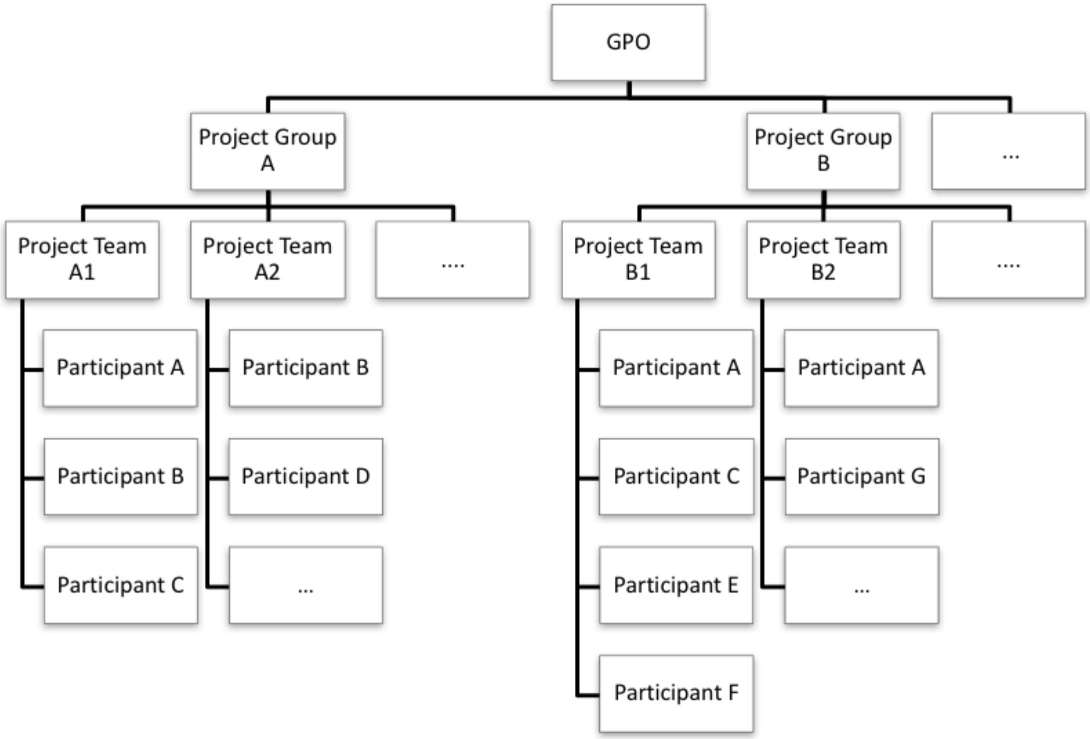

Figure 6. GENI Project Structure with Project Participants

For the purpose of this study, I examine the dyadic collaboration relationships between the GENI project participants during spiral 4. Since each person is not limited to working on one project at any time, a person could thus work with another person on multiple project teams during a certain time period. In Figure 6, for example, collaboration dyads in project team A1 include AB, AC and BC. The dyad AC also occurs in project team B1. The total number of dyad AC is two. If there is no other occurrence of AC in other project teams, the collaboration tie strength for AC will be two.

The GENI spiral under study is spiral 4 (October 2011 to October 2012). The studied population is all GENI project participants in spiral 4: 257 individuals in total from 126 projects. Using this spiral to study collaboration behavior has several benefits. First, as illustrated above, different GENI spirals involve different sets of projects. Some projects may be carried over from a previous spiral because of their good performance; some may be dropped from the overall project due to an unfulfilled outcome; others may be added because of emergent needs or strategic importance. In spiral 4, the project structure is relatively mature. Data from a later stage like this reflect a more stable collaboration state. The GENI project utilizes a variety of mechanisms to socialize participants and help them become familiar with each other; these mechanisms include GEC conferences, listserv, etc. Since the study population is only 257, the small number makes it more likely that the participants are aware of the social characteristics of each other (Lawrence 2006), and even more likely after the iterations of several project phases. Second, using spiral 4 makes it possible to capture historical data by using data from prior spirals. For example, knowledge dependency and prior collaboration calls for data from spiral 3. Given these reasons, this study uses GENI spiral 4 data supplemented by spiral 3 data to study collaboration tie strength.

4.2 DATA COLLECTION

To empirically test the research model and hypotheses, I collected data from archival resources. The GENI wiki website includes lists of projects in each spiral, and the participants of each project. These data are used to construct a collaboration matrix, which is used to measure the dependent variable collaboration tie strength. To measure the independent variables, I collected demographic information and working histories of the GENI project participants from data publicly available, mostly from the Internet. Furthermore, I utilized the GENI developers' listserv, the GENI task repository and the attendance information from GENI Engineering Conferences (GECs) in the analysis.

The major resource of this study is archival data from the GENI wiki, GEC conference data, Internet resources, developer listserv and GENI task repository. The unit of analysis is a collaboration dyad, i.e., two people who work on the same project team.

In order to collect the project data, I started with the GENI wiki site for the project spiral under study, i.e., spiral 4.[6] Overall, there are 14 project groups in spiral 4. Each project group represents a high-level category of the GENI projects and includes a number of individual projects. Table 3 lists the project groups and the number of projects included in each project group, with 126 projects in total.

Table 3. Projects in GENI Project Spiral 4

Project Group	Project Description	Number of projects
1	GENI Meso-Scale Core Network Deployments	6
2	GENI Meso-Scale Regional Network Deployments	7
3	GENI Meso-Scale Campus OpenFlow Deployments	9
4	GENI WiMAX Meso-Scale Deployments	20
5	GENI Meso-Scale Rack Deployments	2
6	GENI Aggregates	22
7	GENI Control Infrastructure Software Development	9
8	GENI Instrumentation and Measurement (I&M) Tools	14
9	GENI Experimenter Tools, Support and Education	12
10	Experimentation with GENI	4
11	GENI Meso-Scale Operations	4
12	GENI Security Projects	6
13	GENI Study Projects	5
14	GENI Projects with Collaborations, Connectivity or Resources outside of the US	6
Total		126

Following the hyperlink for each project of spiral 4, I went to its individual project site and collected the project participants' data. Each project has on average approximately five participants, ranging from one to 15, with the majority having three to six participants. After

[6] http://groups.geni.net/geni/wiki/SpiralFour

identifying the projects and their participants, I looked for each individual's CV from online resources such as personal, institutional and conference websites. I also collected data from social networking sites (LinkedIn and Facebook) of these individuals to supplement and cross validate the data from their CVs, including demographic data (age, gender and race), educational information (schools attended, academic advisor and major), and working history (organizations). Also used in the analysis is the GEC attendance data (update to GEC 14), which were collected during the past GECs from the conference registration.

The first step of data collection is to record individual participant's data in a spreadsheet table, referred to as *participant table*. All data relevant to individuals, including project group arrangement, conference attendance and demographic data, are recoded as columns. After the table for individuals' data is completed, another table for collaboration dyads is created based on the individual data, referred to as *dyad table*. In this table, each pair of individuals within a project team is recorded as a dyad record with dimensions of independent and dependent variables listed as columns. The number of occurrences of the dyad, i.e., collaboration dyad, is counted to serve as the dependent variable, collaboration tie strength. For each dyad, the data from the participant table are compared and computed to generate the independent variables in the dyad table.

A thorough data cleaning was carried out before the analysis. In order to generate an accurate collaboration dyad matrix, each participant's data were validated to be complete. All project participants identified from the GENI project spiral 4 documents were included in the analysis. A database of the project participants and their relevant demographic information, the GENI project data and the GENI conference attendance data, were organized and stored in Excel and Microsoft SQL databases. I used Excel to handle the initial data collection for easy manual updates. Once all the data were collected, the data were organized into a relational database including roughly 10 tables, and then were imported into Microsoft SQL server so that participants' information could be compared and calculated to generate the results for the collaboration dyad. Over 1100 lines of SQL program code were written to clean and prepare the data for the final analysis.

4.3 MEASURES

4.3.1 Dependent variable – collaboration tie strength

The dependent variable in this is study is collaboration tie strength. This construct demonstrates whether or not two individuals collaborate and how often they collaborate. In the context of the GENI project, two individuals are considered to collaborate when they work on the same project team. Collaboration tie strength, in this study, focuses on the frequency perspective of a collaborative relationship. All project teams under study are bound by a time frame, i.e., the duration of the project spiral 4. To take into account differences in project size, the dependent variable is weighted by *project size*, which is measured by the number of people that the project has to which the dyad belongs. Therefore, collaboration tie strength is measured by the number of times two individuals work together on the same project team during a project spiral, weighted by project size.

For this study, I identified the participants of all project teams found in the GENI wiki and project documents for spiral 4. Then based on these data, I constructed a collaboration dyad matrix and calculated the number of collaboration instances for each dyad. For example, if person A and person C worked together on two project teams during spiral 4, the number of times they collaborate is 2. Suppose the one project size is 4 and the other project size is 5, then the weighted collaboration tie strength is: .25+.2=.45

4.3.2 Independent variables – technological needs: knowledge dependency, technical dependency & resource dependency

To evaluate knowledge dependency, technical dependency and resource dependency, I use various archival data stored on the GENI wiki. For knowledge dependency, I utilize the developer listserv data[7] and the task repository data[8] presented on the GENI wiki.

(1) Knowledge Dependency

In the context of the GENI project, there are a variety of ways or tools for a person to seek help from other GENI project participants, but two major tools are the developer listserv and the ticket system. The listserv data include discussions and correspondences among the project participants, and the task repository includes the requests from GENI project participants for help and support and the corresponding solutions provided by other GENI project participants. The GENI listserv and task repository are the major resources open to all GENI participants to communicate and exchange ideas.

When GENI project participants need help or have questions to carry out their tasks, they engage in listserv conversations or open tickets through the GENI ticket system. Because both

---

[7] http://lists.geni.net/pipermail/dev/

[8] http://groups.geni.net/geni/report/

systems are open to the general population and multiple people can respond to any posted question, there is no selection bias. Following the reasoning that a knowledge seeker is dependent on the knowledge of the person(s) who address(es) the question, it is logical to assume that people engaging in the same listserv discussions possess the knowledge each other needs. It is also reasonable to assume that if one provides a solution to the other, as shown in the task repository, the person who seeks the help is dependent on the other's knowledge to complete his/her tasks.

The GENI listserv is a communication tool used by GENI project participants to carry out discussions on project related topics and sharing ideas. Because in many cases GENI project participants are distributed, the listserv is a great tool to facilitate a centralized discussion visible and accessible to other participants. When an individual raises a topic of interest, people who have ideas to share about the topic would respond, and others could then respond to the responses, and thus a listserv thread would form. When many discussions of two individuals occur in the same listserv threads, it is a good indicator that the two people have a high-level of knowledge dependency.

The GENI ticket system is a task-tracking tool used by GENI project participants to post requests and seek help for project related questions. It is more need-based rather than interest-based, as is the case with the GENI listserv. When someone needs help from the broader GENI project population, without knowing who can resolve the issue, he/she may open a GENI ticket so others can offer solutions. Similar to the listserv, the ticket system is open to all GENI project participants, and people who have the solutions can follow the ticket until it is resolved and closed. Therefore, the more times two individuals occur in the same ticket, the higher the level of knowledge dependency between them.

Based on the above reasoning, knowledge dependency is measured by the number of listserv threads that involve the two people in a dyad. The GENI task repository was also considered as a potential measure for knowledge dependency. I measured and stored them into two columns of the dyad table as "knowledge dependency (listserv)" and "knowledge dependency (task)" respectively.

However, in examining the specific contents of the ticket system, I found that all the tickets fall into three major categories. One is reporting. In these tickets, project participants update the GPO with the status of their milestones and tasks. The second category is resource request. In these tickets, project participants request resource for their GEC demonstrations, such as hardware, network resources and rooms. A few others fall into the third category, notifications, which are tickets that notify others of network downtime, reboot and outages. None of these tickets are suitable to measure knowledge dependency because they are meant for coordinating tasks rather than exchanging knowledge. Therefore, tickets are excluded from the analysis as a measure for knowledge dependency, and only listserv threads are kept as the measure for knowledge dependency. Furthermore, to avoid a possible tautology, because the dyads under study have the tendency to generate more dialogs on listserv when they work on the same projects, I use the listserv threads from the previous spiral, spiral 3 (2010.11 to 2011.10), to measure knowledge dependency.

(2) Technical dependency

Technical dependency refers to the degree to which one can provide relevant technical support to the other. In the GENI project, each project team has a list of related projects. The related projects section shown on each project website lists all the other GENI projects that are technically related to this GENI project. Participants of these related GENI projects should be able to provide technical support to each other. It is thus reasonable to believe that participants of these related projects are technically dependent as well. The more times two people are participants of these related projects, the more technically dependent they are.

To measure technical dependency, I use the related projects data presented on each project's website in the GENI project spiral 4. In particular, the "related projects" section from a project page is used to first arrive at the number of pairs of projects. These projects dyads, based on the reasoning above, have technical dependency. Using the project participants data, we could obtain the count of project dyads that involve each collaboration dyad, which is used to measure the technical dependency for each dyad in this study.

(3) Resource dependency

Resource dependency refers to the degree to which one can provide relevant resources to the other. In the GENI project, each project team has a list of participating organizations. To measure resource dependency, I use the participating organizations data presented on each GENI project website. The participating organizations section shown on the project website lists all the organizations that are related to this GENI project. These organizations are either educational institutions, research centers or industrial companies. Participating organizations of the same project can provide resources for each other to develop the GENI projects. Therefore, these organizations have resource dependency on each other. Some of the GENI participants are affiliated with these organizations. It is reasonable to believe that people who are, or once were, affiliated with these organizations have the potential to provide relevant resources to each other. A person can be affiliated with many organizations. The more organizations two people are both affiliated with, the higher the resource dependency between them, and the more often they will collaborate in the GENI project.

Participating organizations provide resources needed to carry out this GENI project. Therefore, individuals from these participating organizations should be able to provide needed resources to each other. The more times two people are participants of these participating organizations, the more resource dependency they share. The study uses GENI spiral 4's project data to measure resource dependency. In particular, the participating organizations section is used to first arrive at the number of pairs of organizations. These organization dyads, based on the reasoning above, have resource dependency. Using the collected affiliation data for the individual participants, we could obtain the count of organization dyads that involve each collaboration dyad, which is used to measure the resource dependency for each dyad in this study.

4.3.3 Independent variables – social attractiveness: power distance, social similarity & familiarity

I identify a group of variables that measure social attractiveness as independent variables and also control for a number of social characteristics that could have predicted the collaboration tie strength, because these social characteristics have been examined in prior network research as predictors for collaboration behavior (Tortoriello et al. 2012).

As outlined in the previous chapter, people are more attracted to work with people who have more power and higher social status. In the context of the GENI project, I examine the power and status through each individual's project role. For example, a person could be a project manager or a PI, or he/she could represent higher management, such as those from the GPO. Overall, three types of roles are identified in the GENI project, i.e., GPO, PI, and Staff/Student, with rank order from high to low using numbers 3 to 1. Power distance between two individuals of the dyad is calculated with the absolute value of the difference between the values of the roles. For example, if one individual in the dyad is a GPO (role=3) and the other is a Student (role=1), the value for power distance is 2.

Meanwhile, people are also attracted to work with people with whom they share similarities. Past network research noted that social similarity tends to produce strong interpersonal connections (McPherson et al. 2001). One aspect of social similarity includes similarity in demographic characteristics, such as age (Marsden 1988, Burt 1991), education (Yamaguchi 1990), gender (Brass 1985; Ibarra 1992, 1997), and race (Marsden 1987, Ibarra 1995, Moody 2001). Other social similarity includes organizational tenure (Zenger and Lawrence 1989) and organizational affiliations (McPherson and Smith-Lovin 1986, 1987). People who share similar social characteristics tend to share similar life experiences and attitudes, thus interacting more easily with each other and forming stronger ties (Laumann 1966, Byrne 1971, Schneider 1987, Reagans 2011). Measuring social similarity could be done by either calculating the commonality or calculating the difference, known as relational demography scores (Reagans 2011).

In this study, I examine social similarity through age, major, gender, race[9], organization and organization type. In order to calculate the similarity between two people, a similarity index

[9] For definition and measure of race, refer to the following websites:
http://www.census.gov/topics/population/race/about.html
http://www.pewresearch.org/fact-tank/2015/06/15/is-being-hispanic-a-matter-of-race-ethnicity-or-both/
http://www.usatoday.com/story/news/politics/2016/09/30/white-house-wants-add-new-racial-category-middle-eastern-people/91322064/

is calculated using the above six similarity dimensions. Each dimension is first transformed into a binary variable 0 and 1, with value 0 meaning not similar and 1 meaning similar. Social similarity in age is 0 if the difference in age of the two individuals is less than 10, because the mean of age difference is around 11 and the standard deviation is around 9 based on the value of the whole dataset. Social similarity in organizations is 1 if the number of the same organizations two individuals ever worked in is more than 0, because the mean value of common organizations is less than 1. Social similarity in major, gender, race and organization type, are all binary variables. Major is the primary field a person studied at school. Gender is categorized as male and female. Race is categorized based on the 1997 Office of Management and Budget (OMB) Standards: White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Hispanic, Middle Eastern and North African or MENA, and Indian. Organization type is the type of organization that an individual belongs to, such as academic institution or industry. For each of these dimensions, a 1 is placed on a dyad where the two persons share the same characteristic, and a 0 is placed on a dyad where the two individuals differ in that dimension. The social similarity in this study is calculated by adding all the binary values in the six similarity dimensions.

Finally, people tend to work with people with whom they are familiar. If a person is acquainted with another person, the degree of familiarity is higher and they are more likely to form a collaborative relationship. People become more familiar through attending the GENI conferences. Therefore, I measure familiarity using the number of common GECs attended throughout the GENI spirals up to spiral 4.

4.3.4 Control variables – same academic advisor, academic advisor & prior collaboration

There are several other variables identified in the GENI project that may predict collaboration tie strength. The variable same academic advisor – a familiarity proxy besides common GECs attended – is measured by the number of same academic advisors the two people ever had; the variable academic advisor – a power proxy besides role difference – is measured by a binary variable that indicates whether one person is the other's academic advisor. In addition, prior collaboration was found to be an important predictor for team formation in prior research (Hahn et al. 2008) and, therefore, is controlled in this study. It is measured using collaboration tie strength from the prior spiral (spiral 3).

The operationalization of all the variables is detailed in Table 4.

Table 4. Summary of Measures

Variable	Definition	Operational Definition	Source
Collaboration tie strength (DV)	The number of times that the two people work on the same project team in a GENI project spiral, weighted by project size.	Numeric variable, which equals the number of projects on which the two people worked together in spiral 4, weighted by the	GENI Wiki (project report)

		project size that the two people participated in.	
Project size (Weight)	The size of the GENI project for each dyad.	Numeric variable, which equals the number of the people in the project that the two people belong to in spiral 4.	GENI wiki
Knowledge dependency (IV)	The degree to which one person could provide relevant knowledge to the other.	Numeric variable, which equals the number of listserv threads that involve the two people in spiral 3.	GENI listserv
Technical dependency (IV)	The degree to which one person could provide relevant technical support to the other.	Numeric variable, which equals the number of "related projects" pairs that involve the two people in spiral 4.	GENI listserv
Resource dependency (IV)	The degree to which one could provide relevant resource to the other.	Numeric variable, which equals the number of "participating organizations" pairs that involve the two people in spiral 4.	GENI wiki (project report)
Power difference (IV)	The difference in power between the two people.	Numeric variable, which equals the power difference indicated by the roles of the two people in GENI: GPO=3, PI=2, Staff/Student=1.	GENI wiki (project report)
Social similarity (IV)	The degree of similarity between the two people based on their social features.	Numeric variable. A similarity index by aggregating the binary difference between the two people in Age, Major, Gender, Race, Organization and Organization Type.	
Age	The similarity between the two people in age.	Numeric variable, which equals the difference in age between the two people.	CV
Major	The similarity between the two people in age.	Binary variable, which equals 1 if the two people have same ultimate academic major, 0 otherwise.	CV
Gender	The similarity between the two people in gender.	Binary variable, which equals 1 if the two people have the same gender, 0 otherwise.	Internet
Race	The similarity between the two people in race. A modified race categorization based on the 1997 Office of Management and Budget (OMB) Standards classify race into: White (1), Black or African American (2), American Indian or Alaska Native (3), Asian (4), Native Hawaiian or Other Pacific Islander (5), Hispanic (6), Middle East and North African or MENA (7), Indian (8).	Binary variable, which equals 1 if the two people share the same race category, 0 otherwise.	Internet
Organization	The similarity between the two people through working for same organizations.	Numeric variable, which equals the number of same	CV& GENI wiki

		organizations the two people ever worked for.	
Organization type	The similarity between the two people through working for same type of organizations, including school or industry.	Binary variable, which equals 1 if the two people work for the same type of organization, 0 otherwise.	CV
Familiarity (IV)	The familiarity between the two people through past acquaintance.	Numeric variable, which is the number of GECs the two people both attended (up to GEC 14).	GEC attendance
Same academic advisor (control)	Whether the two people had the same academic advisor.	Binary variable, which equals 1 if the two people had the same academic advisor, 0 otherwise.	CV
Prior collaboration (control)	The number of times that two individuals work together on the same project team in the prior spiral.	Numeric variable, which equals the number of times the two people previously worked on the same project team in spiral 3.	GENI wiki
Academic Advisor (control)	Whether one person is another's academic advisor.	Binary variable, which equals 1 if one person is the academic advisor for the other.	CV

5.0 ANALYSIS AND RESULTS

5.1 VALIDATION OF MEASURES

Regression with R 1.0.153 is used to analyze the data, and the results are cross validated with SPSS Version 24 for consistency. A total of 32,896 dyads constructed from 257 participants are included in the analyses. Table 5 shows the descriptive statistics for the variables included in the analyses. As shown in the table, there is no missing data.

The dependent variable, *collaboration tie strength*, is measured by the number of projects on which two project participants work together in the GENI project spiral 4 weighted by *project size*. To explore the data thoroughly, additional analyses are conducted for non-weighted values of collaboration tie strength. The independent variables include three technological factors, i.e., *knowledge dependency*, *technical dependency*, and *resource dependency*, and three social factors, i.e., *power distance*, *familiarity* and *similarity*. Table 5 presents these variables and their measurements. In particular, similarity is an index calculated using the binary difference between two people in age, major, gender, race, organization and organization type. Three control variables are included in the analyses, including *academic advisor*, *same academic advisor* and *prior collaboration*.

Table 5. Descriptive Statistics of Variables (Full Dataset)

Variables	Measurement	Data Type	N	Min	Max	Mean	Std. Dev
CollaborationTieStrength (DV)	The number of projects on which the two people worked together in spiral 4	Numeric	32896	0	5	.034	.212
ProjectSize (Weight)	The number of people in the project to which the dyad belongs in spiral 4	Numeric	32896	0	15	.226	1.488
KnowledgeDependency (IV)	The number of listserv threads that involve the two people in spiral 3	Numeric	32896	0	5	.004	.100
TechnicalDependency (IV)	The number of "related projects" pairs that involve the two people in spiral 4	Numeric	32896	0	36	.131	.761
ResourceDependency (IV)	The number of "participating organizations" pairs that involve the two people in spiral 4	Numeric	32896	0	15	.198	.692
PowerDistance (IV)	The difference in project role between two people	Numeric	32896	0	2	.556	.553
Similarity (IV)	A similarity index by aggregating the binary difference between the two people in Age, Major, Gender, Race, Organization, and Organization Type	Numeric	32896	0	6	3.097	1.11
Age	The difference in age between the two people	Numeric	32896	0	56	11.316	8.706
Major	Whether the two people have same ultimate academic major	Binary	32896	0	1	.481	.500
Gender	Whether the two people have the same gender	Binary	32896	0	1	.742	.438
Race	Whether the two people share the same race category	Binary	32896	0	1	.531	.499
Organization	The number of same organizations where two people ever worked	Numeric	32896	0	4	.043	.215
OrganizationType	Whether the two people work for the same type of organization	Binary	32896	0	1	.806	.395
Familiarity (IV)	The number of GECs the two people both attended	Numeric	32896	0	14	1.243	2.146
AcademicAdvisor (control)	Whether one person is the academic advisor for the other	Binary	32896	0	1	.001	.027
SameAcademicAdvisor (control)	Whether the two people had the same academic advisor	Binary	32896	0	1	.001	.024
PriorCollaboration (control)	The number of times the two people previously worked on a same project team in spiral 3	Numeric	32896	0	4	.025	.184

5.2 ANALYSES FOR BASIC MODELS AND RESULTS

Three sets of analyses are carried out to thoroughly test the hypotheses. In the first set of analyses, the full dataset is used, and the dependent variable, collaboration tie strength, is measured as a binary variable. Values different than 0 are entered into the analyses with the value of 1. The result of this set of analyses helps show how different technological factors and social factors predict whether two people collaborate. For this set of analyses, both binary logistic regression and Firth logistic regression are carried out and compared.

In the second set of analyses, the full dataset is used, and the dependent variable, collaboration tie strength, is measured as a numeric count variable weighted by project size. The result of this set of analyses helps show how the technological factors and social factors predict the number of times two people collaborate. Based on the distribution of the dependent variable, collaboration tie strength, Poisson, zero-inflated Poisson, negative binomial regression, and negative binomial regression with maximum likelihood estimation (MLE) are carried out for this set of analyses, with project size as the offset variable. The results are compared and discussed.

In the third set of analyses, only records with non-zero collaboration tie strength are included. The result of this set of analyses helps show how the technological factors and social factors predict the number of times two people collaborate for those who actually collaborated. Poisson regression, negative binomial regression and negative binomial regression (MLE) are carried out with project size as the offset variable for this set of analyses. The results are compared and discussed.

Table 6 summarizes the dataset, dependent variable data type, purpose and regression methods for three sets of analyses. Each set of analyses is described in detail in the following sections.

Table 6. Summary of Analyses and Regression Methods

Analysis	Data set	DV	Purpose	Regression Method
1	Full Dataset	Collaboration tie strength as binary variable	Examine how the technical and social factors predict whether two people collaborate	Logistic; Firth Logistic
2	Full Dataset	Collaboration tie strength as numeric count weighted by project size	Examine how the technical and social factors predict the number of times two people collaborate	Poisson; Zero-inflated Poisson; Negative Binomial; Negative Binomial (MLE)
3	Partial Dataset (DV > 0)	Collaboration tie strength as numeric count weighted by project size	Examine how the technical and social factors predict the number of times two people collaborate for those who collaborate, because the dynamics could be different for first time collaboration and repeated collaboration.	Poisson; Negative Binomial; Negative Binomial (MLE)

In addition, two more sets of analyses are carried out for the non-weighted version of the dependent variable, collaboration tie strength, in which collaboration tie strength is measured by the number of times two people collaborate without using the project size as a weight variable. The results are presented after the weighted version.

5.2.1 Analyses using the full dataset with collaboration tie strength as a binary variable

This set of analyses utilizes the full dataset. The dependent variable, collaboration tie strength, is transformed into a binary variable, i.e., collaboration tie strength is either a value of 0, or a value of 1 if it is non-zero. The goal of this set of analyses is to find out how the technological factors and social factors work together in predicting whether two people collaborate. Logistic regression is considered for the analyses because the dependent variable is a binary variable. Firth logistic regression is also carried out for this analysis as a comparison because it is suitable for rare events.

The corresponding regression models for control variables, main effects and interaction terms are:

$P(Y| x_1 ,x_2)= (1+ \exp (- b_0 - b_1x_1- b_2x_2 - b_3x_3))^{-1}$

$P(Y| x_1 , \ldots, x_9)= (1 + \exp (- b_0 - b_1x_1- b_2x_2- b_3x_3- b_4x_4- b_5x_5- b_6x_6- b_7x_7- b_8x_8- b_9x_9))^{-1}$

$P(Y| x_1 , \ldots, x_6x_9)= (1 + \exp (- b_0 - b_1x_1- b_2x_2- b_3x_3- b_4x_4- b_5x_5- b_6x_6- b_7x_7- b_8x_8- b_9x_9$

$- b_{10}x_4x_7 - b_{11}x_4x_8 - b_{12}x_4x_9 - b_{13}x_5x_7 - b_{14}x_5x_8 - b_{15}x_5x_9 - b_{16}x_6x_7 - b_{17}x_6x_8 - b_{18}x_6x_9))^{-1}$

where

y = collaboration tie strength (binary variable)

$P(Y| x_1 , \ldots, x_i)$= Probability of two people collaborating (collaboration tie strength = 1)

x1~x3 = control variables

x4~x9 = independent variables

$b_0 \sim b_{18}$ are the regression coefficients

Table 7 shows the results from the logistic regression with collaboration tie strength as a binary variable, for models with control variables, main effects and interaction terms. The reduction in AIC of the three models suggests a possible improved model fit.

Table 7. Logistic Regression, Full Dataset, Collaboration Tie Strength as a Binary Variable

CollaborationTieStrength: Binary	**Control Variables**		**Main Effects**		**Interaction Terms**	
	Coeff	**Std. Error**	**Coeff**	**Std. Error**	**Coeff**	**Std. Error**
(Intercept)	-4.947(***)	.0666	-5.641(***)	.2311	-6.282(***)	.3008
KnowledgeDependency			.457	.2997	.447	1.4587
TechnicalDependency			.284(***)	.0352	.589(***)	.1635
ResourceDependency			.942(***)	.0453	1.507(***)	.1687
PowerDistance			-.005	.1221	.153	.1552
Similarity			.021	.0618	.158(**)	.0798
Familiarity			.049(*)	.0279	.107(***)	.0329
KnowledgeDependency*PowerDistance					-.11.204	283.0234
KnowledgeDependency*Similarity					.167	.3337
KnowledgeDependency*Familiarity					-.047	.1019
TechnicalDependency*PowerDistance					-.127(*)	.0733
TechnicalDependency*Similarity					-.042	.0359
TechnicalDependency*Familiarity					-.013	.0130
ResourceDependency*PowerDistance					-.078	.0752
ResourceDependency*Similarity					-.110(***)	.0394
ResourceDependency*Familiarity					-.066(***)	.0195
AcademicAdvisor	2.644(*)	1.0509	.813	1.2058	.659	1.2799
SameAcademicAdvisor	-10.830	381.1224	-11.482	349.0082	-18.786	6502.6463
PriorCollaboration	28.665	583.6166	29.702	493.122	44.845(***)	981.0512
AIC		2716		2317		2307

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Table 8 shows the results from the Firth logistic regression with collaboration tie strength as a binary variable, for models with control variables, main effects and interaction terms. The significance test of the log likelihood ratios of the three models shows an improved model fit, suggesting that the interaction model yields the best model fit.

Table 8. Firth Logistic Regression, Full Dataset, Collaboration Tie Strength as a Binary Variable

CollaborationTieStrength: Binary	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-4.945(***)	.067	-5.634(***)	.230	-6.261(***)	.298
KnowledgeDependency			.553(*)	.263	.964	1.095
TechnicalDependency			.285(***)	.035	.571(***)	.159
ResourceDependency			.940(***)	.045	1.496(***)	.167
PowerDistance			-.001	.121	.150	.154
Similarity			.022	.061	.156(**)	.079
Familiarity			.051(*)	.028	.109(***)	.032
KnowledgeDependency*PowerDistance					-.935	.680
KnowledgeDependency*Similarity					.111	.255
KnowledgeDependency*Familiarity					-.058	.075
TechnicalDependency*PowerDistance					-.108(*)	.069
TechnicalDependency*Similarity					-.041	.035
TechnicalDependency*Familiarity					-.012	.013
ResourceDependency*PowerDistance					-.075	.074
ResourceDependency*Similarity					-.108(***)	.039
ResourceDependency*Familiarity					-.065(***)	.019
AcademicAdvisor	2.998(**)	.914	1.114	1.118	.970	1.187
SameAcademicAdvisor	1.333	1.474	.444	1.597	.422	1.587
PriorCollaboration	11.949(***)	1.417	10.972(***)	1.406	10.986(***)	1.399
Likelihood Ratio Test		5694		6102		6132

*** p < 0.01; ** p < 0.05; * p < 0.10

The logistic regression and Firth logistic regression, as shown in Tables 7 and 8, yield very similar results and parameter estimates. Therefore, the results from the simpler version, logistic regression, are used as the results for this set of analyses. Based on the results, both technical dependency (β = .285, $p < .001$) and resource dependency (β = .943, $p < .001$) positively significantly predict the possibility of whether two people collaborate, i.e., as the technical dependency increases, two people are more likely to work together on a project; as the resource dependency increases, two people are more likely to work together on a project. Familiarity marginally positively and significantly predicts the possibility of whether two people collaborate (β = .051, $p = .077$). We also see significant negative interactions between resource dependency and similarity (β = -.110, p = .005), and between resource dependency and familiarity (β = -.066, $p < .001$), and marginally significant negative interaction between technical dependency and power distance (β = -.128, p = .082), suggesting that the technological factors and social factors are substitutive in predicting whether two people collaborate. In particular, with the increase in power distance, the prediction of the likelihood that two people work together on a project by technical dependency becomes weaker. Similarly, with the increase in similarity and familiarity, the prediction of the likelihood that two people work together on a project by resource dependency also becomes weaker. In all three cases, the interactions do not cancel out the main effects because the additive coefficients of the main effects and interaction terms are positive at the highest levels of the main effects, suggesting that the main effects of the technological factors remain positive despite the changes in social factors. For the control variables, both prior collaboration and academic advisor significantly positively predict whether two people collaborate. In particular, with the increase in the number of projects two people work on together in the prior spiral (spiral 3), the possibility of two people working on the same project increases. Academic advisor is a power proxy besides project role difference. The result suggests that a person is more likely to collaborate with his/her advisor than who is not.

5.2.2 Analyses using the full dataset with collaboration tie strength as a numeric count variable weighted by project size

The dependent variable, collaboration tie strength, consists of count data. In Figure 7, the histogram suggests that the dependent variable may follow a Poisson distribution. A one-sample Kolmogorov-Smirnov test is carried out to further verify the Poisson distribution. The result is non-significant, Kolmogorov-Smirnov Z = .886 ($p = .412$), suggesting that the dependent variable follows a Poisson distribution.

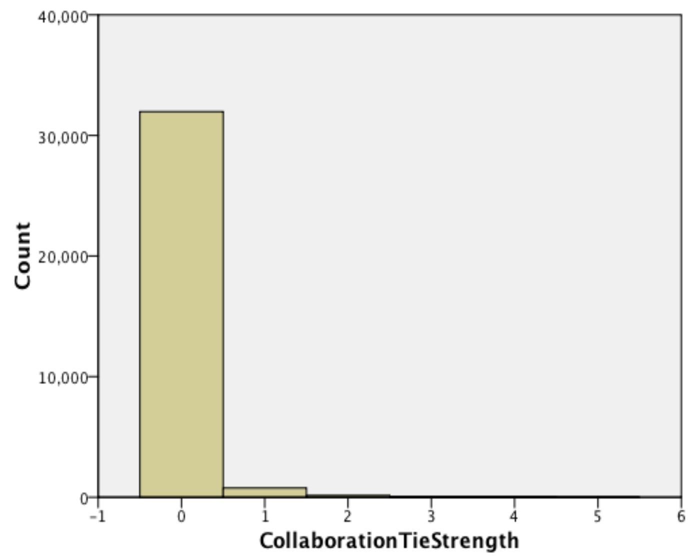

Figure 7. Histogram of Collaboration Tie Strength (Full Dataset)

The mean and variance of the dependent variable are checked to make sure they are similar. The variance slightly exceeds the mean (mean = .03, variance = .045), suggesting that negative binomial regression may be more appropriate for the analyses. Therefore, both Poisson regression and negative binomial regression are carried out for the analyses and then compared. The omnibus test yields a significant result for both models ($p < .0001$), indicating a better model fit over the null model. However, the Pearson Chi-Square of the Poisson regression, .461, a value less than 1, indicates a possible under-dispersed response variable. This under dispersion may come from a zero-inflated dataset. To address this issue, zero-inflated Poisson regression is also carried out to compare.

The Poisson regression (Table 9) and zero-inflated Poisson (Table 10) yield very similar results and parameter estimates, suggesting that the zero-inflated Poisson may not be an improvement over a standard Poisson. To further verify which method is better, a Vuong test is carried out to compare the two models and the result suggests that Poisson regression is superior to the zero-inflated Poisson for this set of analyses.

Table 9. Poisson Regression, Full Dataset, Collaboration Tie Strength as a Numeric Count Variable Weighted by Project Size

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-1.538(***)	.0517	-1.719(***)	.1068	-1.936(***)	.1606
KnowledgeDependency			.088	.1150	.154	3.1006
TechnicalDependency			.013	.0089	.100(***)	.0383
ResourceDependency			.048(***)	.0163	.109(**)	.0509
PowerDistance			.063	.0490	.114	.0788
Similarity			.025	.0261	.068(*)	.0402
Familiarity			.027(***)	.0102	.044(***)	.0155
KnowledgeDependency*PowerDistance					.001	.4949
KnowledgeDependency*Similarity					-.002	.5360
KnowledgeDependency*Familiarity					-.010	.0997
TechnicalDependency*PowerDistance					-.033(*)	.0180
TechnicalDependency*Similarity					-.008	.0085
TechnicalDependency*Familiarity					-.003(*)	.0018
ResourceDependency*PowerDistance					-.006	.0247
ResourceDependency*Similarity					-.013	.0118
ResourceDependency*Familiarity					-.005	.0050
AcademicAdvisor	.161	.2600	.178	.2614	.168	.2622
SameAcademicAdvisor	-.108	1.0005	.066	1.0018	.100	1.0024
PriorCollaboration	.677(***)	.0388	.545(***)	.0504	.536(***)	.0499
AIC		2363		2349		2357

*** p < 0.01; ** p < 0.05; * p < 0.10

Table 10. Zero-inflated Poisson Regression, Full Dataset, Collaboration Tie Strength as a Numeric Count Variable Weighted by Project Size

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-1.538(***)	.0517	-1.719(***)	.1068	-1.936(***)	.1606
KnowledgeDependency			.088	.1150	.154	3.1006
TechnicalDependency			.013	.0089	.100(***)	.0383
ResourceDependency			.048(***)	.0163	.109(**)	.0509
PowerDistance			.063	.0490	.114	.0788
Similarity			.025	.0261	.068(*)	.0402
Familiarity			.027(***)	.0102	.044(***)	.0155
KnowledgeDependency*PowerDistance					.001	.4949
KnowledgeDependency*Similarity					-.002	.5360
KnowledgeDependency*Familiarity					-.010	.0997
TechnicalDependency*PowerDistance					-.033(*)	.0180
TechnicalDependency*Similarity					-.008	.0085
TechnicalDependency*Familiarity					-.003(*)	.0018
ResourceDependency*PowerDistance					-.006	.0247
ResourceDependency*Similarity					-.013	.0117
ResourceDependency*Familiarity					-.005	.0050
AcademicAdvisor	.161	.2600	.178	.2614	.168	.2622
SameAcademicAdvisor	-.108	1.0005	.066	1.0018	.100	1.0024
PriorCollaboration	.677(***)	.0388	.545(***)	.0504	.536(***)	.0499
Log-likelihood		-1.18e+03		-1.16e+03		-1.16e+03

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

However, as was previously noted, the variance slightly exceeds the mean (mean = .03, variance = .045), suggesting that negative binomial regression may be more appropriate for the analyses. For robustness check, both negative binomial regression and the more conservative negative binomial regression (MLE) are carried for comparison. Table 11 shows the results from the negative binomial regression. Table 12 shows the results from the negative binomial regression (MLE). The comparison of the log likelihood between the two models suggests that the negative binomial regression is the better model (p < .0001). However, it not superior to the Poisson model, and the parameter estimates for Poisson, zero-inflated Poisson, negative binomial, and negative binomial (MLE) are almost identical. Therefore, the results from the Poisson regression are used for interpretation.

Table 11. Negative Binomial Regression, Full Dataset, Collaboration Tie Strength as a Numeric Count Variable Weighted by Project Size

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-1.538(***)	.0517	-1.719(***)	.1068	-1.936(***)	.1606
KnowledgeDependency			.088	.1150	.154	3.1006
TechnicalDependency			.013	.0089	.100(***)	.0383
ResourceDependency			.048(***)	.0163	.109(**)	.0509
PowerDistance			.063	.0490	.114	.0788
Similarity			.025	.0261	.068(*)	.0402
Familiarity			.027(***)	.0102	.044(***)	.0155
KnowledgeDependency*PowerDistance					.001	.4949
KnowledgeDependency*Similarity					-.002	.5361
KnowledgeDependency*Familiarity					-.010	.0997
TechnicalDependency*PowerDistance					-.033(*)	.0180
TechnicalDependency*Similarity					-.008	.0085
TechnicalDependency*Familiarity					-.003(*)	.0018
ResourceDependency*PowerDistance					-.006	.0247
ResourceDependency*Similarity					-.013	.0117
ResourceDependency*Familiarity					-.005	.0050
AcademicAdvisor	.161	.2600	.178	.2614	.168	.2622
SameAcademicAdvisor	-.108	1.0005	.066	1.0018	.100	1.0024
PriorCollaboration	.677(***)	.0388	.545(***)	.0504	.536(***)	.0499
AIC		2365		2351		2359

*** p < 0.01; ** p < 0.05; * p < 0.10

Table 12. Negative Binomial Regression (MLE), Full Dataset, Collaboration Tie Strength as a Numeric Count Variable Weighted by Project Size

CollaborationTieStrength: Numeric	**Control Variables**		**Main Effects**		**Interaction Terms**	
	Coeff	**Std. Error**	**Coeff**	**Std. Error**	**Coeff**	**Std. Error**
(Intercept)	-1.538(***)	.0517	-1.719(***)	.1068	-1.936(***)	.1606
KnowledgeDependency			.088	.1150	.154	3.1006
TechnicalDependency			.013	.0089	.100(***)	.0383
ResourceDependency			.048(***)	.0163	.109(**)	.0509
PowerDistance			.063	.0490	.114	.0788
Similarity			.025	.0261	.068(*)	.0402
Familiarity			.027(***)	.0102	.044(***)	.0155
KnowledgeDependency*PowerDistance					.001	.4949
KnowledgeDependency*Similarity					-.002	.5360
KnowledgeDependency*Familiarity					-.010	.0997
TechnicalDependency*PowerDistance					-.033(*)	.0180
TechnicalDependency*Similarity					-.008	.0085
TechnicalDependency*Familiarity					-.003(*)	.0018
ResourceDependency*PowerDistance					-.006	.0247
ResourceDependency*Similarity					-.013	.0117
ResourceDependency*Familiarity					-.005	.0050
AcademicAdvisor	.161	.2600	.178	.2614	.168	.2622
SameAcademicAdvisor	-.108	1.0005	.066	1.0018	.100	1.0024
PriorCollaboration	.677(***)	.0388	.545(***)	.0504	.536(***)	.0499
AIC		2365		2350		2358

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

To account for the impact of project size on collaboration tie strength, this set of analyses uses the weighted value of collaboration tie strength by project size as the dependent variable. All of the aforementioned regression methods are carried out for models with control variables, main effects and interaction terms, with collaboration tie strength as a numeric count variable and the log of project size as the offset variable. The values of project size for all records are increased by .0001 so that all records are included in the analyses. The corresponding regression models being tested are:

$$\log(E(y)/t) = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$$\log(E(y)/t) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9$$

$$\log(E(y)/t) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9$$
$$+ b_{10}x_4x_7 + b_{11}x_4x_8 + b_{12}x_4x_9 + b_{13}x_5x_7 + b_{14}x_5x_8 + b_{15}x_5x_9 + b_{16}x_6x_7 + b_{17}x_6x_8 + b_{18}x_6x_9$$

where

y = collaboration tie strength (numeric count variable)

E(y)/t = Expected value of (collaboration tie strength/project size)

x1~x3 = control variables

x4~x9 = independent variables

b_0 ~ b_{18} are the regression coefficients

Based on the results from the Poisson regression in Table 10, models with control variables and main effects show similar results as in the prior analyses. Both resource dependency ($\beta = .048$, $p = .004$) and familiarity ($\beta = .027$, $p = .009$) positively significantly predict the number of times two people collaborate, i.e., as the technical dependency increases, two people tend to work on more projects together; as the resource dependency increases, two people tend to work on more projects together. In the interaction model, only technical dependency and two of the social factors, power distance and familiarity, are marginally significant with negative interactions ($\beta = -.033$, $p = .069$; $\beta = -.003$, $p = .080$ respectively). The negative coefficients suggest that with the increase in power distance, the positive prediction of the number of projects two people work on together by technical dependency become weaker on collaboration tie strength, and with the increase in familiarity, the positive prediction of the number of projects two people work on together by technical dependency also becomes weaker. Among all of the control variables, only prior collaboration significantly positively predicts whether two people collaborate. In particular, with the increase in the number of projects two people work on together in the prior spiral (i.e. spiral 3), the number of projects two people work on together also increases.

5.2.3 Analyses using the partial dataset (DV > 0) with collaboration tie strength as a numeric count variable weighted by project size

Table 13. Descriptive Statistics of Variables (Partial Dataset)

Variables	Data Type	N	Min	Max	Mean	Std. Dev
CollaborationTieStrength (DV)	Numeric	922	1	5	1.194	.464
ProjectSize (Weight)	Numeric	922	2	15	8.07	3.960
KnowledgeDependency (IV)	Numeric	922	0	4	.013	.192
TechnicalDependency (IV)	Numeric	922	0	36	1.600	3.260
ResourceDependency (IV)	Numeric	922	0	15	2.205	1.881
PowerDistance (IV)	Numeric	922	0	2	.626	.678
Similarity (IV)	Numeric	922	0	6	3.493	1.215
Age	Numeric	922	0	49	11.316	8.706
Major	Binary	922	0	1	.481	.500
Gender	Binary	922	0	1	.742	.438
Race	Binary	922	0	1	.531	.499
Organization	Numeric	922	0	3	.043	.215
OrganizationType	Binary	922	0	1	.806	.395
Familiarity (IV)	Numeric	922	0	14	2.100	3.100
AcademicAdvisor (control)	Binary	922	0	1	.015	.122
SameAcademicAdvisor (control)	Binary	922	0	1	.001	.033
PriorCollaboration (control)	Numeric	922	0	4	.905	.642

In this set of analyses, only the partial dataset (DV>0) is examined. The goal is to find out for those who actually collaborated, how do the various technological factors and social factors interact with each other in predicting the number of times two people collaborate. For this purpose, only records with collaboration tie strength that are non-zero are entered into the analyses. This partial dataset includes a total of 922 records. Table 10 shows the descriptive statistics of the variables for the partial dataset.

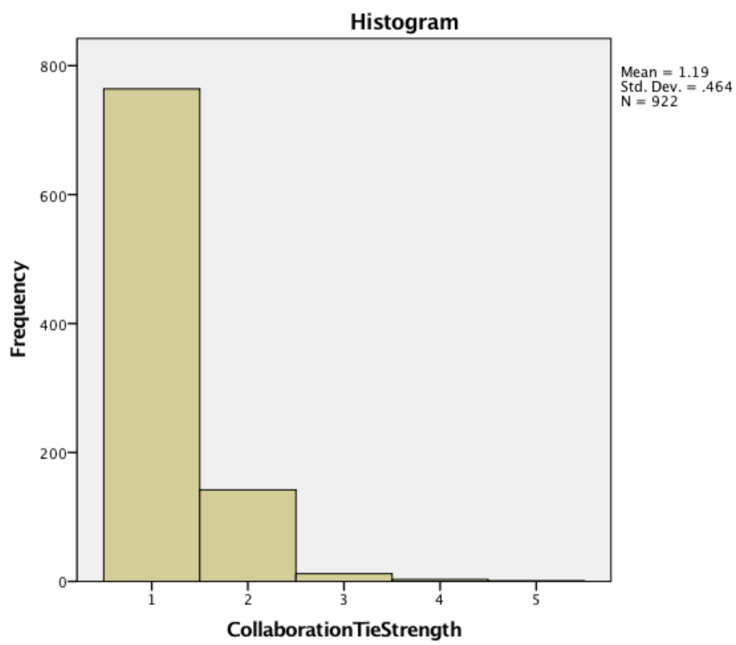

Figure 8. Histogram of Collaboration Tie Strength (Partial Dataset)

Based on the histogram shown in Figure 8, the dependent variable, collaboration tie strength, still seems to follow a Poisson distribution. However, the mean and variance are different for the dependent variable (mean = 1.19, variance = .215). Therefore, negative binomial regression might be better for the analyses. Both Poisson regression and negative binomial regression are carried out and compared. For a robustness check, negative binomial regression (MLE) is also carried out for even more conservative results. The Poisson regression (Table 14), negative binomial regression (Table 15), and negative binomial regression (MLE) (Table 16) yield almost identical results. The Vuong Non-Nested Hypothesis Test-Statistic indicates that Poisson regression is a better model ($p < .0001$), and therefore is chosen as the final results for this set of analyses.

To account for the impact of project size on collaboration tie strength, all regressions are carried out for models with control variables, main effects and interaction effects, with collaboration tie strength as a numeric count variable and the log of project size as the offset variable. All values of project size are increased by .0001 so that all records are included in the analyses.

The corresponding regression models being tested are:

$$\log(E(y)/t) = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$$\log(E(y)/t) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9$$

$$\log(E(y)/t) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9$$
$$+ b_{10}x_4x_7 + b_{11}x_4x_8 + b_{12}x_4x_9 + b_{13}x_5x_7 + b_{14}x_5x_8 + b_{15}x_5x_9 + b_{16}x_6x_7 + b_{17}x_6x_8 + b_{18}x_6x_9$$

where

y = collaboration tie strength (numeric count variable) >0

$E(y)/t$ = Expected value of (collaboration tie strength/project size)

x1~x3 = control variables

x4~x9 = independent variables

$b_0 \sim b_{18}$ are the regression coefficients

Table 14. Poisson Regression, Partial Dataset (DV>0), Collaboration Tie Strength as a Numeric Count Variable

Weighted by Project Size

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-1.088(***)	.0563	-1.202(***)	.1098	-1.231(***)	.1696
KnowledgeDependency			.102	.1180	1.006	4.8407
TechnicalDependency			.021(**)	.0089	.098(***)	.0371
ResourceDependency			.011	.0168	-.010	.0537
PowerDistance			.044	.0483	.126	.0796
Similarity			.017	.0260	.019	.0418
Familiarity			.024(**)	.0100	.017	.0159
KnowledgeDependency*PowerDistance					-.075	.5487
KnowledgeDependency*Similarity					-.151	.8486
KnowledgeDependency*Familiarity					-.032	.1241
TechnicalDependency*PowerDistance					-.045(***)	.0174
TechnicalDependency*Similarity					-.006	.0084
TechnicalDependency*Familiarity					-.001	.0018
ResourceDependency*PowerDistance					-.005	.0257
ResourceDependency*Similarity					-.003	.0125
ResourceDependency*Familiarity					-.003	.0053
AcademicAdvisor	.034	.2600	.076	.2614	.059	.2621
SameAcademicAdvisor	-.207	1.0005	-.100	1.0020	-.062	1.0026
PriorCollaboration	.392(***)	.0440	.306(***)	.0512	.297(***)	.0507
AIC		2053		2043		2053

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Table 15. Negative Binomial Regression, Partial Dataset (DV>0), Collaboration Tie Strength as a Numeric Count Variable Weighted by Project Size

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-1.088(***)	.0563	-1.202(***)	.1098	-1.231(***)	.1696
KnowledgeDependency			.102	.1180	1.007	4.8407
TechnicalDependency			.021(**)	.0089	.098(***)	.0371
ResourceDependency			.011	.0168	-.010	.0537
PowerDistance			.044	.0483	.126	.0796
Similarity			.017	.0260	.019	.0418
Familiarity			.024(**)	.0100	.017	.0159
KnowledgeDependency*PowerDistance					-.075	.5487
KnowledgeDependency*Similarity					-.151	.8486
KnowledgeDependency*Familiarity					-.032	.1241
TechnicalDependency*PowerDistance					-.045(***)	.0174
TechnicalDependency*Similarity					-.006	.0084
TechnicalDependency*Familiarity					-.001	.0018
ResourceDependency*PowerDistance					-.005	.0257
ResourceDependency*Similarity					-.003	.0125
ResourceDependency*Familiarity					-.003	.0053
AcademicAdvisor	.034	.2600	.076	.2614	.059	.2621
SameAcademicAdvisor	-.207	1.0005	-.099	1.0020	-.063	1.0026
PriorCollaboration	.392(***)	.0440	.306(***)	.0512	.297(***)	.0507
AIC		2055		2045		2055

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Table 16. Negative Binomial Regression (MLE), Partial Dataset (DV>0), Collaboration Tie Strength as a Numeric Count Variable Weighted by Project Size

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-1.087(***)	.0563	-1.212(***)	.1098	-1.239(***)	.1696
KnowledgeDependency			.102	.1180	1.007	4.8407
TechnicalDependency			.021(**)	.0089	.098(***)	.0371
ResourceDependency			.011	.0168	-.010	.0537
PowerDistance			.044	.0483	.126	.0796
Similarity			.017	.0260	.019	.0418
Familiarity			.024(**)	.0100	.017	.0159
KnowledgeDependency*PowerDistance					-.075	.5487
KnowledgeDependency*Similarity					-.151	.8486
KnowledgeDependency*Familiarity					-.032	.1241
TechnicalDependency*PowerDistance					-.045(***)	.0174
TechnicalDependency*Similarity					-.006	.0084
TechnicalDependency*Familiarity					-.001	.0018
ResourceDependency*PowerDistance					-.005	.0257
ResourceDependency*Similarity					-.003	.0125
ResourceDependency*Familiarity					-.003	.0053
AcademicAdvisor	.034	.2600	.076	.2614	.059	.2621
SameAcademicAdvisor	-.207	1.0005	-.099	1.0020	-.063	1.0026
PriorCollaboration	.392(***)	.0440	.306(***)	.0512	.297(***)	.0507
AIC		2055		2045		2055

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Based on the results from the Poisson regression shown in Table 14, the model for main effects demonstrates the best model fit based on the AIC values. The Vuong test also suggests that the main effects model is significantly better than the control model (p = .0002) and the interaction model (p = .0006).

Based on the results of the main effects model, both technical dependency (β = .021, p = .020) and familiarity (β = .024, p = .044) positively significantly predict the number of times two people collaborate for those who actually collaborated. As the technical dependency increases, two people tend to work on more projects together; as the familiarity increases, two people tend to work on more projects together. Based on the interaction model, only the interaction between technical dependency and power distance (β = -.045, p = .010) is negatively significant, suggesting that they are substitutive in predicting collaboration tie strength. The prediction of the number of projects two people work on together by technical dependency becomes weaker with the increase in power distance, but the interaction did not fully cancel out the main effect of technical dependency, because the additive value of the coefficients of the main effect and interaction remains positive. Among all the control variables, only prior collaboration positively significantly predicts the number of times two people collaborate. In particular, with the increase in the number of projects two people work on together in the prior spiral (i.e. spiral 3), the number of projects two people work on together in the current spiral also increases.

5.2.4 Summary of results (main effects)

The above analyses yield different results depending on the range and type of the dependent variable, collaboration tie strength, i.e., whether it is numeric or binary, and whether or not it includes zero value. Analysis 1 uses the full dataset with the dependent variable as a binary variable; analysis 2 uses the full dataset with the dependent variable as a numeric count variable weighted by project size; and analysis 3 uses the partial dataset with the dependent variable as a numeric count variable weighted by project size.

Table 17 summarizes the results of these analyses for the main models and Table 18 for the interaction models. The values in the tables are the significant coefficients from the corresponding chosen model tests.

Table 17. Summary of Results (Main Effects)

Hypothesis	Analysis 1 DV=(0,1) Logistic	Analysis 2 DV>=0 Weighted Poisson with Offset	Analysis 3 DV>0 Weighted Poisson with Offset
1A: Knowledge Dependency			
1B: Technical Dependency	.284 (***)		.021 (**)
1C: Resource Dependency	.942 (***)	.048 (***)	
2A: Power Distance			
2B: Similarity			
2C: Familiarity	.049 (*)	.027 (***)	.024 (**)

*** p < 0.01; ** p < 0.05; * p < 0.10

Based on the results of main effects model from analysis 1, using the logistic regression on the full dataset with the collaboration tie strength as a binary variable, both technical dependency and resource dependency positively significantly predict the possibility of whether two GENI project participants may collaborate. With one unit increase in technical dependency, the log odds of tie strength increases by .284 (p < .001), and with one unit increase in resource dependency, the log odds of tie strength increases by .942 (p < .001). Familiarity is marginally positively significant. With one unit increase in familiarity, the log odds of tie strength increases by .049 (p = .077). Figure 9 illustrates the increase in odds ratios of two people collaborating with the increase in each social or technological factor.

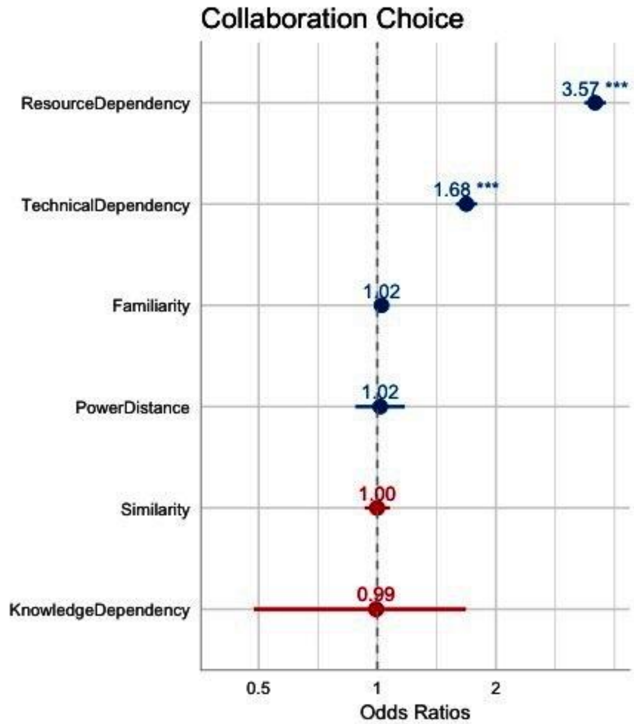

Figure 9. Main Effects from Logistic Regression for Analysis 1

In analysis 2 (on the full dataset) and analysis 3 (on the partial dataset), collaboration tie strength is kept as a numeric count variable weighted by project size. With the full dataset, as shown in analysis 2, resource dependency and familiarity both significantly positively predict the number of times two project participants collaborate. To better interpret the results, I examined the incident rate ratios rather than coefficients by calculating the exponentiation of the model coefficients and the confidence intervals, as shown in Figures 10 and 11. Each unit increase in resource dependency corresponds to 1.05% (97.5% CI, 1.015 to 1.08, p = .003) increase in collaboration tie strength controlling for other variables; each unit increase in familiarity corresponds to 1.03% (97.5% CI, 1.006 to 1.05, p = .009) increase in collaboration tie strength controlling for other variables. With the partial dataset, each unit increase in technical dependency corresponds to 1.02% (97.5% CI, 1.003 to 1.04, p = .020) increase in collaboration tie strength controlling for other variables; each unit increase in familiarity corresponds to 1.02% (97.5% CI, 1.004 to 1.04, p = .015) increase in collaboration tie strength controlling for other variables.

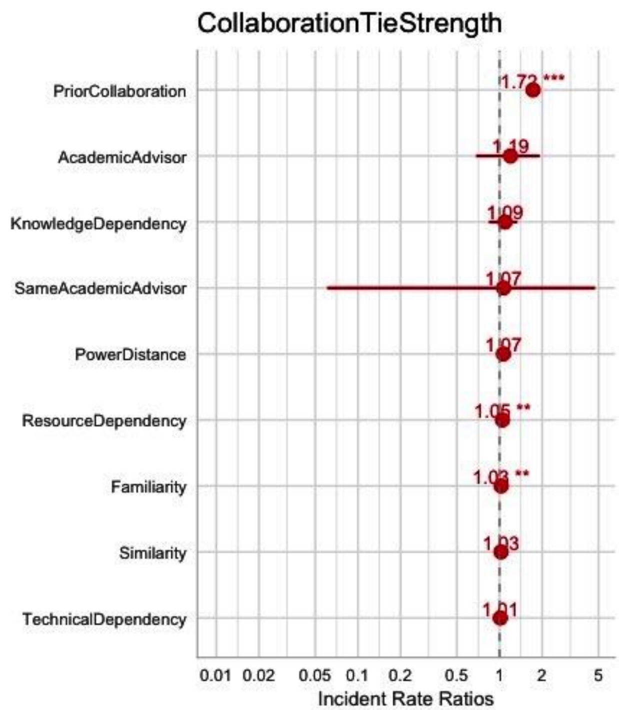

Figure 10. Main Effects from Poisson Regression for Analysis 2

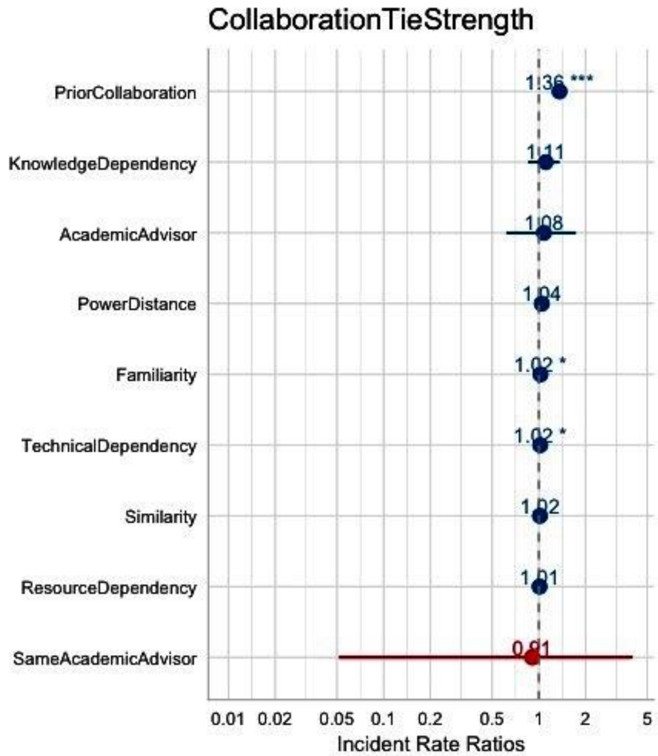

Figure 11. Main Effects from Poisson Regression for Analysis 3

To summarize, the results from the three sets of analyses suggest that both technical dependency and resource dependency positively predict whether two people may collaborate. However, once people collaborate (DV>0), resource dependency no longer significantly predicts collaboration tie strength, and technical dependency is the factor that significantly predicts repeated collaboration, suggesting that resource dependency may be an important factor for collaboration tie formation (whether two people work together on a project), but technical dependency may be important for stronger collaboration ties (how many projects two people work on together). Familiarity consistently significantly predicts whether two people may collaborate and how many times two people collaborate at any stage.

As outlined in section 1.2, CI projects are different from both distributed organizational ISD projects and open source software (OSS) projects. Unlike in the traditional organizational settings, people in CI projects are not typically assigned to projects but rather they can freely choose with whom to work. Meanwhile, such choice is not random because they are motivated to produce success in order to secure further funding. Furthermore, CI projects utilize some unique mechanisms, such as the GEC conferences in the GENI project, to foster familiarity. The results from the analysis resonate well with the practice observed in the GENI project.

In this study, familiarity is measured by the common GEC conferences two project participants both attended. The results suggest that the more common GEC conferences both project participants attended, the more projects they work on together. In these conference sessions, the GENI project stakeholders participate in a variety of working sessions, such as workshops, DEMOs and tutorials, as well as networking events, such as birds of feather dinners and round-table discussions, which become great venues to get familiar with each other. There are important reasons for the GPO to host these conferences. The GENI project, as with any other CI project, is extremely complex and grand in scale; it is even difficult to position oneself among the projects. Forming collaboration teams without a good understanding of the CI project is extremely challenging. It is important to get people to understand what others are working on and to identify the needs. A project participant commented,

"What is challenging once they [GENI participants] are in the door is to get a real understanding of what GENI is and its limitations…For Peer to Peer project with [a PI], we're trying to engage researchers who don't use networks (such as those doing simulations) and help them interpret results from the GENI testbed…We're working very hard to communicate to them…"

Another experimenter, when asked how he decided which project to get involved with, responded, “It is not that [project A] is better than [project B] – just that more familiar with [Project A] …”

The GECs provide exactly the opportunity for the project participants to get familiar with each other and gain understanding of each other’s expertise, thus fostering collaboration among the stakeholders both initially and repeatedly. Of course, there are also people who commented that the GECs may not be organized efficiently. Overall, this study suggests that the familiarity brought about by the GECs promotes collaboration.

Resources, as shown in the result, are also important for initial collaboration formation. In the GENI project, resources are mostly in the form of network bands and computing capabilities. This is different from that in open source software projects, in which these are typically not of concern, or from that in organizational ISD projects, in which resources are typically provided. Resources are so important in CI projects that it is a main reason that draws some researchers into the GENI project. As commented by some GENI experimenters who work on DDOS attacks, GENI allows them to experiment beyond simulations, and to have access to mobile networks, setting parameters and getting access to real traffic data. Others mentioned that they wanted to do real world testbed at scale, and that GENI is especially for those small universities who don’t have those resources. For collaborations decisions, resources seem to be an important factor people consider so people who could provide access to useful resources are more favorably chosen as collaborators. For example, a PI commented on a project,

“[Project] is modest in its goals but is supported by real companies. Google is deploying it in its own backbone…”

The result also suggests that technology is an important predictor for both initial and repeated collaboration, i.e., the more related technology two project participants possess, the more likely and the more projects they work on together. Inter-operability among different projects is an important objective to build the overall CI. Projects are intertwined and therefore people consider technology when choosing collaboration partners. One GENI project actually conducts research on how to marry networks. The project would like to find out "how [project A] fits with [project B] and [project C] …" As another example, a project participant commented,

"For cluster – we have our own testbed at [a university] – use the [project A] software based on [project B]. Take the software, in order for our resources to be exported to GENI community, have to write an interface between [project A] software and our physical resources…"

To summarize, in the context of CI projects, familiarity, resource and technology are important factors that influence people's decision about with whom to work. Therefore, for CI project management, it is important to keep these factors in mind when deriving policies to promote collaboration.

5.2.5 Summary of results (interactions)

Table 18 summarizes the results of the interaction models from the three sets of analyses for hypotheses testing.

Table 18. Summary of Results (Interactions)

Hypothesis	Analysis 1 DV=(0,1) Logistic	Analysis 2 DV>=0 Weighted Poisson with Offset	Analysis 3 DV>0 Weighted Poisson with Offset
1A: Knowledge Dependency			
1B: Technical Dependency	.589 (***)	.100 (***)	.098 (***)
1C: Resource Dependency	1.507 (***)	.109 (**)	
2A: Power Distance			
2B: Similarity	.158 (**)	.068 (*)	
2C: Familiarity	.107 (***)	.044 (***)	
*3A: Knowledge Dependency * Power Distance*			
*3B: Knowledge Dependency * Similarity*			
*3C: Knowledge Dependency * Familiarity*			
*3D: Technical Dependency * Power Distance*	-.127 (*)	-.033 (*)	-.045 (***)
*3E: Technical Dependency * Similarity*			
*3F: Technical Dependency * Familiarity*		-.003 (*)	
*3G: Resource Dependency * Power Distance*			
*3H: Resource Dependency * Similarity*	-.110 (***)		
*3I : Resource Dependency * Familiarity*	-.066 (***)		

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Based on the results of the interaction model from analysis 1 on the full dataset using the logistic regression with the collaboration tie strength as a binary variable, certain technical and social factors negatively interact with each other in predicting whether two people may collaborate. Specifically, the interaction between technical dependency and power distance is marginally significant (β = -.127, p = .082), so that the effect of technical dependency in predicting whether two people may collaborate becomes weaker with the increase in power distance between two people. However, the interaction does not fully cancel out the main effect of technical dependency because the main effect remains significantly positive, meaning the slope of prediction of collaboration tie strength (whether two people collaborate) by technical dependency becomes flatter but stays positive with the increase in power distance. The interaction between resource dependency and similarity is negatively significant (β = -.110, p = .005), suggesting that the effect of resource dependency in predicting whether two people may collaborate becomes weaker with the increase in similarity between two people. The interaction between resource dependency and familiarity is also negatively significant (β = -.066, p = .001), suggesting that the effect of resource dependency in predicting whether two people collaborate becomes weaker with the increase in similarity between two people. Both interactions (resource dependency*similarity and resource dependency*familiarity) do not fully cancel out the main effect of resource dependency because the main effect of resource dependency remains significantly positive at different levels of similarity and familiarity, meaning the slope of prediction of collaboration tie strength by resource dependency stays positive despite the changes in similarity or familiarity.

Based on the results of the interaction model from analysis 2, using the full dataset with collaboration tie strength as a numeric count variable weighted by project size, only technical dependency and two of the social factors, power distance and familiarity, negatively interact with each other in predicting the number of times two people collaborate. Specifically, the interaction between technical dependency and power distance is marginally significant and negative (β = -.033, p = .069), so that the effect of technical dependency in predicting whether two people may collaborate is weaker with the increase in power distance between two people. Similarly, the interaction between technical dependency and familiarity is marginally significant and negative (β = -.003, p = .080), suggesting that the increase in familiarity weakens the prediction by technical dependency on whether two people may collaborate. However, both interactions do not fully cancel out the main effect of technical dependency because the main effect remains significantly positive at different levels of power distance and familiarity, meaning the slope of prediction becomes flatter but stays positive with the increase in power distance or familiarity.

Finally, based on the results of the interaction model from analysis 3, using the partial dataset with collaboration tie strength as a numeric count variable weighted by project size, only technical dependency and power distance significantly negatively interact with each other in predicting the number of times two people collaborate (β = -.045, p = .001), so that the effect of technical dependency in predicting whether two people may collaborate becomes weaker with the increase in power distance between two people. However, the interaction does not fully cancel out the main effect of technical dependency because the main effect remains significantly positive at the highest level of power distance, meaning the slope of prediction becomes flatter but stays positive with the increase in power distance.

Figures 12, 13 and 14 demonstrate in linear graphs the significant interaction effects from the three sets of analyses using logistic regression and Poisson regression respectively. Figures 15, 16 and 17 are 3D graphs for the three interactions.

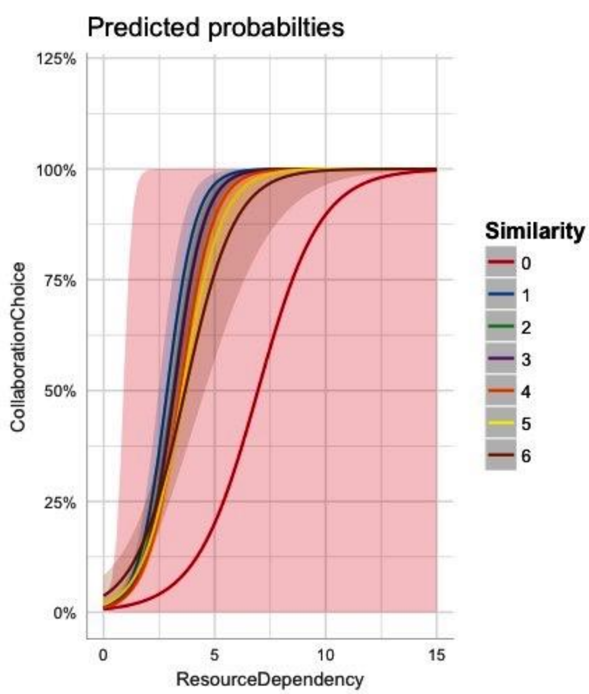

Figure 12. Interaction Effects from Logistic Regression for Analysis 1 (Resource Dependency * Similarity)

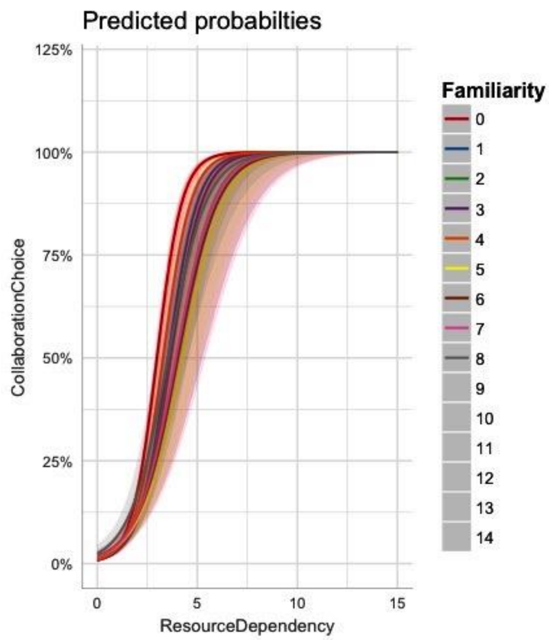

Figure 13. Interaction Effects from Logistic Regression for Analysis 1 (Resource Dependency * Familiarity)

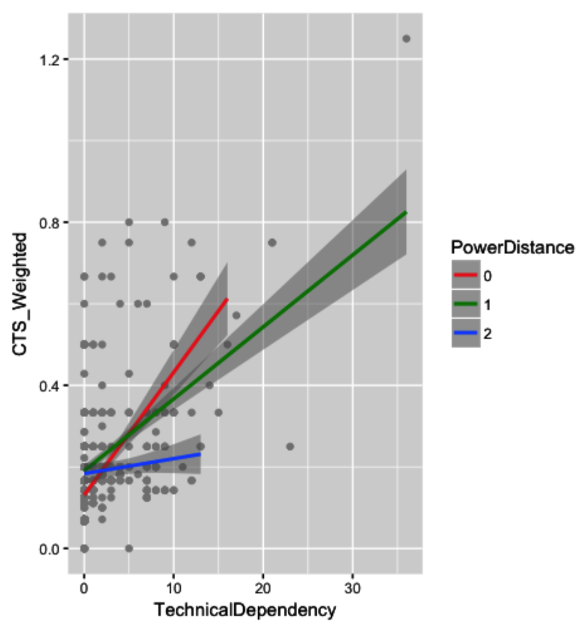

Figure 14. Interaction Effects from Poisson Regression for Analysis 3 (Technical Dependency * Power Distance)

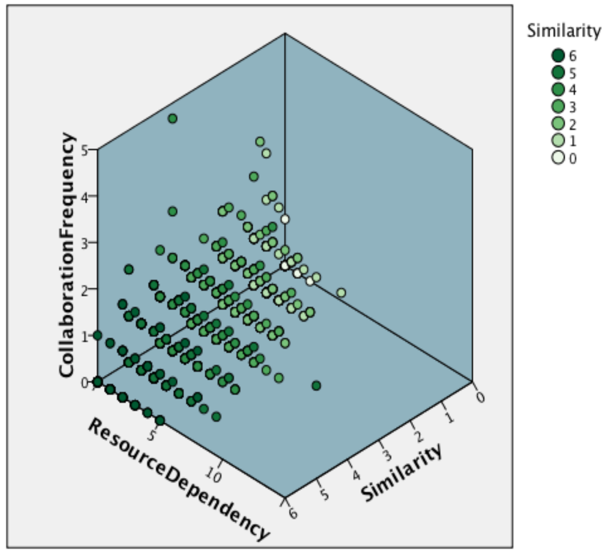

Figure 15. 3D Graph for Interaction Effects from Logistic Regression for Analysis 1 (Resource Dependency * Similarity)

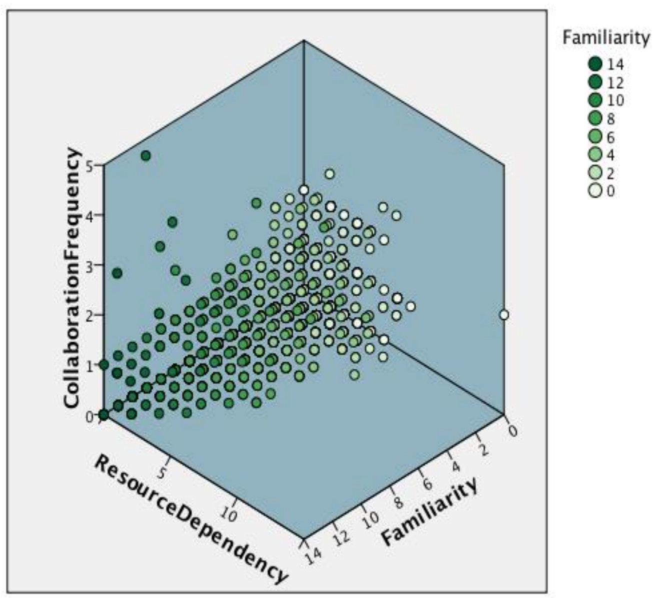

Figure 16. 3D Graph for Interaction Effects from Logistic Regression for Analysis 1 (Resource Dependency * Familiarity)

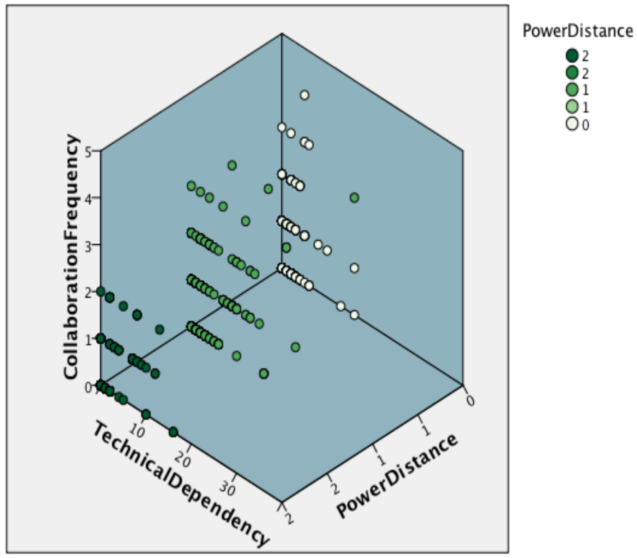

Figure 17. 3D Graph for Interaction Effects from Poisson Regression for Analysis 3 (Technical Dependency * Power Distance)

Overall, the three sets of analyses are consistent in that technological factors and social factors negatively interact with each other in predicting whether two people may collaborate or how many times two people collaborate, suggesting that social factors weaken the prediction of collaboration by technological factors. However, this substitution effect is not strong enough to negate the positive prediction by the technological factors.

In the context of CI projects, social factors such as familiarity, similarity and power distance affect the strength of influence by technological factors on collaboration choice and frequency. Although similarity is not significant on its own, it substitutes the effect of resource dependency in predicting collaboration choice, suggesting that when people are more similar, they care less about how much resource the other party can provide. For example, one participant explicitly said that she loved to collaborate with female faculty members. Similarly, with power distance, when two people hold different project roles, they care less about how much technology the other person can provide. Understanding these contingencies is important for policy derivation. For example, methods that intend to increase collaboration through increasing technical dependency and resource dependency are most effective when people are less familiar, less similar and with less power distance.

5.3 MORE EXPLORATORY ANALYSES AND RESULTS

In this section, more analyses are conducted to further explore the relationships among different variables. First, all two-way and three-way interactions among all independent variables are examined on the full dataset with collaboration tie strength as a binary variable and as a numeric count variable weighted by project size, and on the partial dataset with collaboration tie strength as a numeric count variable weighted by project size. Second, the non-weighted version of the dependent variable is used for the analysis, i.e., collaboration tie strength is measured by the number of times two people collaborate, without being weighted by the project size.

5.3.1 All two-way and three-way interactions (binary and weighted DV)

Table 19 illustrates the output from logistic regression using the whole dataset with the dependent variable, collaboration tie strength, as a binary variable for all five modes: models for control variables, main effects, two-way interactions between technological factors and social factors, two-way interactions between all factors, and three-way interactions among all factors. The reduction in AIC of the five models demonstrates possible enhanced model fit. The Vuong tests are then conducted to decide whether the changes in AIC between models reflect significant improvements in model fit. The raw statistics of the Vuong tests suggest that model 5, the three-way interaction model, is the best model, while the AIC and BIC statistics suggest that model 2, the main effects model, yields the best model fit.

Table 19. Logistic Regression, Full Dataset, Collaboration Tie Strength as a Binary Variable

CollaborationTieStrength: Binary	Control Variables		Main Effects		Interaction (tech*social)		Interaction (all 2 way)		Interaction (all 3 way)	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-4.947(***)	.0666	-5.641(***)	.2311	-6.282(***)	.3008	-6.776(***)	.4033	-7.523(***)	.4794
KnowledgeDependency			.457	.2997	.447	1.4587	.414	1.8351	-.225	2.0324
TechnicalDependency			.284(***)	.0352	.589(***)	.1635	.519(***)	.1742	.687(***)	.2189
ResourceDependency			.942(***)	.0453	1.507(***)	.1687	1.488(***)	.1780	1.853(***)	.2411
PowerDistance			-.005	.1221	.153	.1552	1.110(***)	.3740	2.144(***)	.4922
Similarity			.021	.0618	.158(**)	.0798	.339(***)	.1078	.575(***)	.1266
Familiarity			.049(*)	.0279	.107(***)	.0329	-.045	.1084	.130	.1537
KnowledgeDependency*PowerDistance					-.11.204	283.0234	-11.089	275.6535	-12.969	752.0153
KnowledgeDependency*Similarity					.167	.3337	.002	.3720	.203	.3980
KnowledgeDependency*Familiarity					-.047	.1019	.032	.1266	.030	.1280
TechnicalDependency*PowerDistance					-.127(*)	.0733	-.134(*)	.0754	-.068	.0926
TechnicalDependency*Similarity					-.042	.0359	-.035	.0370	-.099(**)	.0500
TechnicalDependency*Familiarity					-.013	.0130	-.007	.0132	-.006	.0170
ResourceDependency*PowerDistance					-.078	.0752	-.064	.0773	-.408(*)	.2340
ResourceDependency*Similarity					-.110(***)	.0394	-.115(***)	.0409	-.237(***)	.0597
ResourceDependency*Familiarity					-.066(***)	.0195	-.068(***)	.0209	-.039	.0301
KnowledgeDependency*TechnicalDependency							1.495	1.3358	565.185	17667.4194
KnowledgeDependency*ResourceDependency							.023	.3205	-.084	.3323
TechnicalDependency*ResourceDependency							.044	.0380	-.110	.1532
PowerDistance*Similarity							-.352(***)	.1031	-.728(***)	.1483
PowerDistance*Familiarity							.084(*)	.0459	-.137	.1324
Similarity*Familiarity							.026	.0254	-.034	.0409
KnowledgeDependency*TechnicalDependency*ResourceDependency									286.344	8864.7184
KnowledgeDependency*TechnicalDependency*PowerDistance									-683.109	30843.0771
KnowledgeDependency*TechnicalDependency*Similarity									-316.144	9782.3958
KnowledgeDependency*TechnicalDependency*Familiarity									86.138	2661.1702
TechnicalDependency*ResourceDependency*PowerDistance									-.080	.0666
TechnicalDependency*ResourceDependency*Similarity									-.067(*)	.0361
TechnicalDependency*ResourceDependency*Familiarity									-.004	.0166
ResourceDependency*PowerDistance*Similarity									.154(**)	.0642
ResourceDependency*PowerDistance*Familiarity									-.051	.0351
PowerDistance*Similarity*Familiarity									.086(**)	.0393
AcademicAdvisor	2.644(*)	1.0509	.813	1.2058	.659	1.2799	.668	1.2851	.683	1.217
SameAcademicAdvisor	-10.830	381.1224	-11.482	349.0082	-18.786	6502.6463	-19.605	10080.7466	-22.364	60643.7684
PriorCollaboration	28.665	583.6166	29.702	493.122	44.845	981.0512	46.657	959.1236	54.372	2609.6992
AIC		2716		2317		2307		2302		2296

*** p < 0.01; ** p < 0.05; * p < 0.10

Table 20 illustrates the output from Poisson regression using the full dataset with collaboration tie strength as a numeric count variable weighted by project size for all five modes: models for control variables, main effects, two-way interactions between technological factors and social factors, two-way interactions between all factors, and three-way interactions among all factors. The reduction in AIC from model 1 and model 2 suggests possible enhanced model fit, while the increase in AIC from model 2 through model 5 suggests possible reduced model fit. The Vuong tests are consistent with the change in AIC, suggesting that model 2, the main effects model, yields the best model fit.

Table 20. Poisson Regression, Full Dataset, Collaboration Tie Strength as a Numeric Count Variable Weighted by Project Size

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction (tech*social)		Interaction (all 2 way)		Interaction (all 3 way)	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-1.538(***)	.0517	-1.719(***)	.1068	-1.936(***)	.1606	-2.045(***)	.1856	-2.122(***)	.2185
KnowledgeDependency			.088	.1150	.154	3.1006	.486	3.5279	-.736	4.8871
TechnicalDependency			.013	.0089	.100(***)	.0383	.123(**)	.0486	.118 (*)	.0715
ResourceDependency			.048(***)	.0163	.109(**)	.0509	.116(**)	.0517	.130(**)	.0661
PowerDistance			.063	.0490	.114	.0788	.257	.1574	.361	.2242
Similarity			.025	.0261	.068(*)	.0402	.103(**)	.0488	.131(**)	.0585
Familiarity			.027(***)	.0102	.044(***)	.0155	.034	.0402	.065	.0607
KnowledgeDependency*PowerDistance					.001	.4949	-1.114	3.1102	-1.242	4.1460
KnowledgeDependency*Similarity					-.002	.5361	-.005	.6585	.214	.8663
KnowledgeDependency*Familiarity					-.010	.0997	-.054	.1590	-.031	.2149
TechnicalDependency*PowerDistance					-.033(*)	.0180	-.041(**)	.0195	-.031	.0292
TechnicalDependency*Similarity					-.008	.0085	-.005	.0092	-.001	.0159
TechnicalDependency*Familiarity					-.003(*)	.0018	-.004(**)	.0019	-.007(**)	.0033
ResourceDependency*PowerDistance					-.006	.0247	.005	.0257	.005	.0734
ResourceDependency*Similarity					-.013	.0117	-.015	.0119	-.021	.0163
ResourceDependency*Familiarity					-.005	.0050	-.002	.0056	.0004	.0093
KnowledgeDependency*TechnicalDependency							-.231	.6293	2.191	38.7863
KnowledgeDependency*ResourceDependency							.181	.5697	.216	.7556
TechnicalDependency*ResourceDependency							-.007	.0054	-.009	.0248
PowerDistance*Similarity							-.053	.0427	-.111(*)	.0661
PowerDistance*Familiarity							.008	.0173	-.016	.0485
Similarity*Familiarity							-.001	.0096	-.012	.0164
KnowledgeDependency*TechnicalDependency*ResourceDependency									-.053	2.1274
KnowledgeDependency*TechnicalDependency*PowerDistance									-10.754	1306.6666
KnowledgeDependency*TechnicalDependency*Similarity									-.746	12.2953
KnowledgeDependency*TechnicalDependency*Familiarity									.263	2.3234
TechnicalDependency*ResourceDependency*PowerDistance									-.005	.0111
TechnicalDependency*ResourceDependency*Similarity									-.001	.0053
TechnicalDependency*ResourceDependency*Familiarity									.002	.0015
ResourceDependency*PowerDistance*Similarity									.012	.0190
ResourceDependency*PowerDistance*Familiarity									-.010	.0093
PowerDsitance*Similarity*Familiarity									.013	.0144
AcademicAdvisor	.161	.2600	.178	.2614	.168	.2622	.184	.2628	.182	.2635
SameAcademicAdvisor	-.108	1.0005	.066	1.0018	.100	1.0024	.095	1.0025	.076	1.0027
PriorCollaboration	3.544(***)	.0450	2.832(***)	.0676	2.685(***)	.0908	.525(***)	.0503	.525(***)	.0506
AIC		2365		2351		2359		2367		2383

*** p < 0.01; ** p < 0.05; * p < 0.10

Table 21 illustrates the output from Poisson regression using the partial dataset (DV>0) with collaboration tie strength as a numeric count variable weighted by project size for all five modes: models for control variables, main effects, two-way interactions between technological factors and social factors, two-way interactions between all factors, and three-way interactions among all factors.

Table 21. Poisson Regression, Partial Dataset (DV>0), Collaboration Tie Strength as a Numeric Count Variable Weighted by Project Size

CollaborationTieStrength: Binary	Control Variables		Main Effects		Interaction (tech*social)		Interaction (all 2 way)		Interaction (all 3 way)	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-1.088(***)	.0563	-1.202(***)	.1098	-1.231(***)	.1696	-1.295(***)	.1931	-1.245(***)	.2259
KnowledgeDependency			.102	.1180	1.006	4.8407	1.338	6.0300	1.383	6.0314
TechnicalDependency			.021(**)	.0089	.098(***)	.0371	.074	.0467	.042	.0708
ResourceDependency			.011	.0168	-.010	.0537	-.011	.0550	-.044	.0722
PowerDistance			.044	.0483	.126	.0796	.205	.1510	.161	.2126
Similarity			.017	.0260	.019	.0418	.039	.0501	.030	.0600
Familiarity			.024(**)	.0100	.017	.0159	.040	.0402	.064	.0610
KnowledgeDependency*PowerDistance					-.075	.5487	-.018	.6385	-.031	.6403
KnowledgeDependency*Similarity					-.151	.8486	-.225	1.1438	-.231	1.1440
KnowledgeDependency*Familiarity					-.032	.1241	-.037	.1339	-.036	.1340
TechnicalDependency*PowerDistance					-.045(***)	.0174	-.039(**)	.0191	-.022	.0287
TechnicalDependency*Similarity					-.006	.0084	-.003	.0091	-.001	.0160
TechnicalDependency*Familiarity					-.001	.0018	-.001	.0020	.00004	.0034
ResourceDependency*PowerDistance					-.005	.0257	-.004	.0270	.042	.0757
ResourceDependency*Similarity					-.003	.0125	.003	.0129	.010	.0178
ResourceDependency*Familiarity					-.003	.0053	.002	.0059	.0003	.0096
KnowledgeDependency*TechnicalDependency							.013	.1134	.008	.1137
KnowledgeDependency*ResourceDependency							NA	NA	NA	NA
TechnicalDependency*ResourceDependency							.003	.0052	.018	.0244
PowerDistance*Similarity							-.022	.0420	-.018	.0643
PowerDistance*Familiarity							-.008	.0173	-.032	.0498
Similarity*Familiarity							-.005	.0097	-.011	.0164
KnowledgeDependency*TechnicalDependency*ResourceDependency									NA	NA
KnowledgeDependency*TechnicalDependency*PowerDistance									NA	NA
KnowledgeDependency*TechnicalDependency*Similarity									NA	NA
KnowledgeDependency*TechnicalDependency*Familiarity									NA	NA
TechnicalDependency*ResourceDependency*PowerDistance									-.009	.0109
TechnicalDependency*ResourceDependency*Similarity									-.001	.0052
TechnicalDependency*ResourceDependency*Familiarity									-.0003	.0015
ResourceDependency*PowerDistance*Similarity									-.009	.0194
ResourceDependency*PowerDistance*Familiarity									.001	.0095
PowerDistance*Similarity*Familiarity									. 008	.0146
AcademicAdvisor	.034	.2600	.076	.2614	.059	.2621	.056	.2628	.065	.2633
SameAcademicAdvisor	-.207	1.0005	-.100	1.0020	-.062	1.0026	-.080	1.0028	-.089	1.0030
PriorCollaboration	.392(***)	.0440	.306(***)	.0512	.297(***)	.0507	.296(***)	.0511	.293(***)	.0513
AIC		2053		2043		2053		2062		2073

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

The reduction in AIC from model 1 and model 2 suggests possible enhanced model fit, while the increase in AIC from model 2 through model 5 suggests possible reduced model fit. The Vuong tests are consistent with the change in AIC, suggesting that the model 2, the main effects model, yields the best model fit.

Based on the changes in model fit, the addition of all two-way interactions and three-way interaction models do not add to the overall model fit. Therefore, the output from the prior sections is preferred as the final results.

5.3.2 All two-way and three-way interactions (non-weighted DV)

In section 5.3.1, the dependent variable collaboration tie strength is a weighted variable (i.e. the number of collaboration ties weighted by the project size); the analyses shown in this section (5.3.2) are the corresponding analyses for the non-weighted version.

First, the full dataset is used, and the dependent variable, collaboration tie strength, is measured as a numeric count variable. The result of this set of analyses helps show how the technological and social factors predict the number of times two people collaborate. Based on the distribution of the dependent variable, collaboration tie strength, Poisson regression, zero-inflated Poisson, and negative binomial regression are carried out and compared. Negative binomial regression is found to yield the best fit.

Second, only records with collaboration tie strength that is non-zero are included. The result of this set of analyses helps show how the technological factors and social factors predict the number of times two people collaborate for those who actually collaborated. Poisson regression and negative binomial regression are carried out and compared. Poisson regression is found to be the superior model.

(1) Analyses using the full dataset with collaboration tie strength as a numeric count variable

The dependent variable, collaboration tie strength, consists of count data. In Figure 7, the histogram suggests that the dependent variable may follow a Poisson distribution. A one-sample Kolmogorov-Smirnov test is carried out to further verify the Poisson distribution. The result is non-significant, Kolmogorov-Smirnov Z = .886 (p = .412), indicating that the dependent variable follows a Poisson distribution.

The mean and variance of the dependent variable were checked to make sure they are similar. The variance slightly exceeds the mean (mean = .03, variance = .045), indicating that negative binomial regression may be more appropriate for the analyses. Therefore, both Poisson regression and negative binomial regression were carried out for the analyses and then compared. The omnibus test yields a significant result for both models ($p < .0001$), indicating a better model fit over the null model. However, the Pearson Chi-Square of the Poisson regression, .461, a value less than 1, indicates a possible under-dispersed response variable. This under dispersion may come from a zero-inflated dataset. To address this issue, zero-inflated Poisson regression was also carried out to compare.

The corresponding regression models being tested are:

$$\log(E(y))= b_0 + b_1x_1+ b_2x_2+ b_3x_3$$

$$\log(E(y))= b_0 + b_1x_1+ b_2x_2+ b_3x_3+ b_4x_4+ b_5x_5+ b_6x_6+ b_7x_7+ b_8x_8+ b_9x_9$$

$$\log(E(y))= b_0 + b_1x_1+ b_2x_2+ b_3x_3+ b_4x_4+ b_5x_5+ b_6x_6+ b_7x_7+ b_8x_8+ b_9x_9 + b_{10}x_4x_7+ b_{11}x_4x_8+ b_{12}x_4x_9 + b_{13}x_5x_7+ b_{14}x_5x_8+ b_{15}x_5x_9 + b_{16}x_6x_7+ b_{17}x_6x_8+ b_{18}x_6x_9$$

where

y = collaboration tie strength (numeric count variable)

E(y) = Expected count value

x1~x3 = control variables

x4~x9 = independent variables

b_0 ~ b_{18} are the regression coefficients

Tables 22 and 23 show the results from the Poisson regression and zero-inflated Poisson regression, with collaboration tie strength as a numeric count variable, for models with control variables, main effects and interaction terms.

The Poisson regression and zero-inflated Poisson yield very similar results and parameter estimates, suggesting that the zero-inflated Poisson may not be an improvement over a standard Poisson. To further verify which method is better, a Vuong test was carried out to compare the two models and the Vuong test suggests that Poisson regression is superior to the zero-inflated Poisson for this set of analyses.

Table 22. Poisson Regression, Full Dataset, Collaboration Tie Strength as a Numeric Count Variable

CollaborationTieStrength: Numeric	**Control Variables**		**Main Effects**		**Interaction Terms**	
	Coeff	**Std. Error**	**Coeff**	**Std. Error**	**Coeff**	**Std. Error**
(Intercept)	-3.961(***)	.0382	-4.418(***)	.1079	-5.168(***)	.1509
KnowledgeDependency			.142	.1097	-.913	1.7329
TechnicalDependency			-.059(***)	.0085	-.162(***)	.0485
ResourceDependency			.174(***)	.1057	.683(***)	.0464
PowerDistance			.242	.0538	-.008	.0750
Similarity			.046(*)	.0274	.238(***)	.0383
Familiarity			.001	.0111	.108(***)	.0146
KnowledgeDependency*PowerDistance					.179	.4542
KnowledgeDependency*Similarity					.233	.3020
KnowledgeDependency*Familiarity					-.004	.0848
TechnicalDependency*PowerDistance					.100(***)	.0199
TechnicalDependency*Similarity					.029(***)	.0102
TechnicalDependency*Familiarity					-.013(***)	.0018
ResourceDependency*PowerDistance					-.018	.0222
ResourceDependency*Similarity					-.104(***)	.0099
ResourceDependency*Familiarity					-.031(***)	.0040
AcademicAdvisor	1.407(***)	.2600	1.030(***)	.2618	.959(***)	.2624
SameAcademicAdvisor	.674	1.0006	.810	1.0012	.652	1.0014
PriorCollaboration	2.171(***)	.0215	2.179(***)	.0475	2.123(***)	.0500
AIC		6000		5636		5406

*** p < 0.01; ** p < 0.05; * p < 0.10

Table 23. Zero-inflated Poisson, Full Dataset, DV as a Numeric Count Variable

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-3.961(***)	.0382	-4.418(***)	.1079	-5.168(***)	.1509
KnowledgeDependency			.142	.1097	-.913	1.7329
TechnicalDependency			-.060(***)	.0085	-.162(***)	.0484
ResourceDependency			.174(***)	.1057	.683(***)	.0464
PowerDistance			.242	.0538	-.008	.0750
Similarity			.046(*)	.0274	.238(***)	.0384
Familiarity			.001	.0111	.108(***)	.0146
KnowledgeDependency*PowerDistance					.179	.4542
KnowledgeDependency*Similarity					.233	.3020
KnowledgeDependency*Familiarity					-.004	.0848
TechnicalDependency*PowerDistance					.100(***)	.0199
TechnicalDependency*Similarity					.029(***)	.0102
TechnicalDependency*Familiarity					-.013(***)	.0017
ResourceDependency*PowerDistance					-.018	.0222
ResourceDependency*Similarity					-.104(***)	.0099
ResourceDependency*Familiarity					-.031(***)	.0040
AcademicAdvisor	1.405(***)	.2602	1.031(***)	.2618	.959(***)	.2624
SameAcademicAdvisor	.674	1.0006	.810	1.0013	.652	1.0014
PriorCollaboration	2.171(***)	.0215	2.179(***)	.0475	2.123(***)	.0499
Log-Likelihood	-3e+3 (df = 5)		-2.81e+03 (df = 11)		-2.68e+03 (df = 20)	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

However, as was previously noted, the variance slightly exceeds the mean (mean = .03, variance =. 045), indicating that negative binomial regression may be more appropriate for the analyses. Table 24 shows the results from the negative binomial regression, and Table 25 shows the results from the negative binomial regression (MLE). The comparison of the log likelihood between the Poisson regression and negative binomial regression suggests that the negative binomial regression is not superior to the Poisson model for the analyses, and the parameter estimates are similar for all four sets of analyses. Therefore, results from the Poisson regression are used as the outcome for this set of analyses.

Table 24. Negative Binomial Regression, Full Dataset, Collaboration Tie Strength as a Numeric Count Variable

CollaborationTieStrength: Numeric	**Control Variables**		**Main Effects**		**Interaction Terms**	
	Coeff	**Std. Error**	**Coeff**	**Std. Error**	**Coeff**	**Std. Error**
(Intercept)	-4.462(***)	.0450	-4.758(***)	.1269	-5.268(***)	.1696
KnowledgeDependency			-.034	.1826	-.727	2.1832
TechnicalDependency			-.017	.0154	-.008	.0744
ResourceDependency			.346(***)	.0231	.753(***)	.0772
PowerDistance			.077	.0653	.065	.0869
Similarity			.028	.0333	.163(***)	.0441
Familiarity			.050(***)	.-138	.114(***)	.0171
KnowledgeDependency*PowerDistance					.260	.6139
KnowledgeDependency*Similarity					.128	.3601
KnowledgeDependency*Familiarity					.024	.1047
TechnicalDependency*PowerDistance					.096(***)	.0297
TechnicalDependency*Similarity					-.016	.0160
TechnicalDependency*Familiarity					-.012(***)	.0039
ResourceDependency*PowerDistance					-.066(**)	.0384
ResourceDependency*Similarity					-.085(***)	.0182
ResourceDependency*Familiarity					-.033(***)	.0082
AcademicAdvisor	.851(**)	.3579	.671(*)	.3509	.754(**)	.3615
SameAcademicAdvisor	.360	1.2077	.716	1.1216	.674	1.0430
PriorCollaboration	3.544(***)	.0450	2.832(***)	.0676	2.685(***)	.0908
AIC		5444		5290		5230

*** p < 0.01; ** p < 0.05; * p < 0.10

Table 25. Negative Binomial Regression (MLE), Full Dataset, Collaboration Tie Strength as a Numeric Count Variable

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-4.463(***)	.0565	-4.758(***)	.1325	-5.268(***)	.1696
KnowledgeDependency			-.031	.2566	-.726	2.1832
TechnicalDependency			-.017	.0203	-.007	.0744
ResourceDependency			.346(***)	.0289	.754(***)	.0772
PowerDistance			.077	.0694	.065	.0869
Similarity			.028	.0346	.163(***)	.0441
Familiarity			.050(***)	.0144	.114(***)	.0171
KnowledgeDependency*PowerDistance					.260	.6139
KnowledgeDependency*Similarity					.127	.3601
KnowledgeDependency*Familiarity					.024	.1047
TechnicalDependency*PowerDistance					.096(***)	.0297
TechnicalDependency*Similarity					-.016	.0160
TechnicalDependency*Familiarity					-.012(***)	.0039
ResourceDependency*PowerDistance					-.066(*)	.0384
ResourceDependency*Similarity					-.085(***)	.0182
ResourceDependency*Familiarity					-.033(***)	.0082
AcademicAdvisor	.851(**)	.3822	.671(*)	.3808	.754(**)	.3615
SameAcademicAdvisor	.359	1.1378	.716	1.0624	.674	1.0430
PriorCollaboration	3.546(***)	.0981	2.833(***)	.0952	2.687(***)	.0908
AIC		5444		5290		5230

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

The reduction in AIC among the three models indicates an improved model fit. Therefore, the interaction model among the three models is considered the best fit in this set of analyses.

Based on the results of main effects model, resource dependency ($\beta = .346$, $p = .000$) and familiarity ($\beta = .050$, $p = .001$) both positively significantly predict the number of times two people collaborate. However, the interactions get more complicated, with some being positive and some negative. The interaction between technical dependency and power distance is positively significant ($\beta = .096$, $p = .001$), indicating they are complementary in predicting collaboration tie strength, i.e., with the increase in power distance, technical dependency has a stronger positive prediction on collaboration tie strength. However, most other interactions, including interactions between technical dependency and familiarity (β = -.012, p = .002), resource dependency and power distance (β = -.066, p = .087), resource dependency and similarity (β = -.085, p = .000), resource dependency and familiarity (β = -.033, p = .000), are all negatively significant, indicating they are substitutive in predicting collaboration tie strength. In looking at the coefficients of the main effects and interaction terms, the increase in social factors yields different changes in the prediction of the technological factors. However, overall the increase in power distance weakens the prediction of resource dependency. The increase in similarity and familiarity also weakens the prediction of resource dependency.

(2) Analyses using the partial dataset (DV > 0) with collaboration tie strength as a numeric count variable

In this set of analyses, only the partial dataset (DV>0) is examined. The goal is to find out, for those who actually collaborated, how do the various technological factors and social factors interact with each other in predicting the number of times two people collaborate. For this purpose, only records with collaboration tie strength that are non-zero are entered into the analyses. This partial dataset includes a total of 922 records.

Based on the histogram shown in Figure 8, the dependent variable, collaboration tie strength, still seems to follow a Poisson distribution. However, the mean and variance are different for the dependent variable (mean = 1.19, variance = .215). Therefore, negative binomial regression (MLE) might be better for the analyses. Poisson regression, negative binomial regression and negative binomial regression (MLE) are all carried out and compared.

The corresponding regression models being tested are:

$\log(E(y)) = b_0 + b_1x_1 + b_2x_2 + b_3x_3$

$\log(E(y)) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9$

$\log(E(y)) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9$

$+ b_{10}x_4x_7 + b_{11}x_4x_8 + b_{12}x_4x_9 + b_{13}x_5x_7 + b_{14}x_5x_8 + b_{15}x_5x_9 + b_{16}x_6x_7 + b_{17}x_6x_8 + b_{18}x_6x_9$

where

y = collaboration tie strength (numeric count variable) > 0

E(y) = Expected count value

x1~x3 = control variables

x4~x9 = independent variables

$b_0 \sim b_{18}$ are the regression coefficients

Table 26 shows the results from Poisson regression using the reduced dataset with collaboration tie strength as a numeric count variable, for models with control variables, main effects and interaction terms. Table 27 shows the results from negative binomial regression using the reduced dataset with collaboration tie strength as a numeric count variable, for models with control variables, main effects and interaction terms. Table 28 shows the results from Negative Binomial regression (MLE) using the reduced dataset with collaboration tie strength as a numeric count variable, for models with control variables, main effects and interaction terms.

Table 26. Poisson Regression, Partial Dataset (DV>0), Collaboration Tie Strength as a Numeric Count Variable

CollaborationTieStrength: Numeric	**Control Variables**		**Main Effects**		**Interaction Terms**	
	Coeff	**Std. Error**	**Coeff**	**Std. Error**	**Coeff**	**Std. Error**
(Intercept)	-.218(***)	.0567	-.237(**)	.1106	-.243	.1696
KnowledgeDependency			.118	.1169	.632	4.7720
TechnicalDependency			.011	.0089	.050	.0371
ResourceDependency			.016	.0169	.000	.0546
PowerDistance			-.018	.0490	.024	.0811
Similarity			.000	.0259	.004	.0415
Familiarity			.008	.0101	-.003	.0162
KnowledgeDependency*PowerDistance					.022	.5488
KnowledgeDependency*Similarity					-.094	.8359
KnowledgeDependency*Familiarity					-.014	.1236
TechnicalDependency*PowerDistance					-.025	.0173
TechnicalDependency*Similarity					-.004	.0083
TechnicalDependency*Familiarity					.000	.0018
ResourceDependency*PowerDistance					.002	.0261
ResourceDependency*Similarity					.001	.0126
ResourceDependency*Familiarity					.004	.0053
AcademicAdvisor	-.126	.2600	-.102	.2612	-.113	.2618
SameAcademicAdvisor	-.183	1.0005	-.125	1.0020	-.120	1.0026
PriorCollaboration	.402(***)	.0444	.354(***)	.0516	.347(***)	.517
AIC		2009		2016		2031

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Table 27. Negative Binomial Regression, Partial Dataset (DV>0), Collaboration Tie Strength as a Numeric Count Variable

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-.218(***)	.0567	-.237(**)	.1106	-.243	.1699
KnowledgeDependency			.118	.1169	.632	4.7720
TechnicalDependency			.011	.0089	.050	.0371
ResourceDependency			.016	.0169	.0001	.0546
PowerDistance			-.018	.0490	.024	.0811
Similarity			.0001	.0259	.004	.0415
Familiarity			.008	.0101	-.003	.0162
KnowledgeDependency*PowerDistance					.022	.5488
KnowledgeDependency*Similarity					-.094	.8359
KnowledgeDependency*Familiarity					-.014	.1236
TechnicalDependency*PowerDistance					-.025	.0173
TechnicalDependency*Similarity					-.004	.0083
TechnicalDependency*Familiarity					.000	.0018
ResourceDependency*PowerDistance					.002	.0261
ResourceDependency*Similarity					.001	.0126
ResourceDependency*Familiarity					.004	.0053
AcademicAdvisor	-.126	.2600	-.102	.2612	-.113	.2618
SameAcademicAdvisor	-.138	1.0005	-.125	1.0020	-.120	1.0026
PriorCollaboration	.402(***)	.0444	.354(***)	.0516	.347(***)	.517
AIC		2011		2018		2033

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Table 28. Negative Binomial Regression (MLE), Partial Dataset (DV>0), Collaboration Tie Strength as a Numeric Count Variable

CollaborationTieStrength: Numeric	Control Variables		Main Effects		Interaction Terms	
	Coeff	Std. Error	Coeff	Std. Error	Coeff	Std. Error
(Intercept)	-.218(***)	.0567	-.237(**)	.1106	-.243	.1696
KnowledgeDependency			.118	.1169	.632	4.7720
TechnicalDependency			.011	.0089	.050	.0371
ResourceDependency			.016	.0169	6.706E-5	.0546
PowerDistance			-.018	.0490	.024	.0811
Similarity			.000	.0259	.004	.0415
Familiarity			.008	.0101	-.003	.0162
KnowledgeDependency*PowerDistance					.022	.5488
KnowledgeDependency*Similarity					-.094	.8359
KnowledgeDependency*Familiarity					-.014	.1236
TechnicalDependency*PowerDistance					-.025	.0173
TechnicalDependency*Similarity					-.004	.0083
TechnicalDependency*Familiarity					.000	.0018
ResourceDependency*PowerDistance					.002	.0261
ResourceDependency*Similarity					.001	.0126
ResourceDependency*Familiarity					.004	.0053
AcademicAdvisor	-.126	.2600	-.102	.2612	-.113	.2618
SameAcademicAdvisor	-.138	1.0005	-.125	1.0020	-.120	1.0026
PriorCollaboration	.402(***)	.0444	.354(***)	.0516	.347(***)	.517
AIC		1980		2018		2033

*** p < 0.01; ** p < 0.05; * p < 0.10

The Poisson regression, negative binomial regression, and negative binomial regression (MLE) yield almost identical results. The Vuong Non-Nested Hypothesis Test-Statistic indicates that Poisson regression is the best model ($p < .0001$), and therefore is chosen for this set of analyses.

Based on the results of the Poisson regression, the increase in AIC values among the three models indicates a reduced model fit. No significant results are found on any independent variables. Only the control variable prior collaboration positively significantly predicts the number of times two people collaborate for those who actually collaborated ($\beta = .402$, $p < .001$).

6.0 DISCUSSION AND CONCLUSION

6.1 DISCUSSION OF FINDINGS

This study examines the antecedents of collaboration tie strength in cyberinfrastructure projects from both the social and technological perspective. Specifically, using the GENI project data, this study focuses on two sets of antecedents, technological factors that involve knowledge dependency, technical dependency and resource dependency, and social factors that involve power distance, social similarity and familiarity.

The study explores the antecedents of collaboration tie strength in a variety of ways. In particular, the current research not only examines the individual effects of each of these factors, but also explores the interactions among these factors in predicting collaboration tie strength. In addition, the dependent variable, collaboration tie strength, is examined from two different angles, i.e., as a binary variable representing whether two people collaborate, and as a numeric count variable representing the number of times two people collaborate. Furthermore, both the full dataset that involves all records and the partial dataset that involves only the records with non-zero collaboration tie strength values are examined for hypotheses testing. Table 29 summarizes the results of hypotheses testing from the three sets of analyses.

In the following sections, the results of the different types of analyses are discussed in detail. Specifically, I discuss the result of hypotheses testing with collaboration tie strength operationalized as a binary variable, as a numeric count variable and as a positive numeric count variable.

Table 29. Summary of Hypotheses Testing

Hypothesis	Analysis 1 Full Dataset Binary	Analysis 2 Full Dataset Numeric	Analysis 3 DV>0 Numeric
1A: Knowledge Dependency			
1B: Technical Dependency	supported		supported
1C: Resource Dependency	supported	supported	
2A: Power Distance			
2B: Similarity			
2C: Familiarity	supported	supported	supported
*3A: Knowledge Dependency * Power Distance*			
*3B: Knowledge Dependency * Similarity*			
*3C: Knowledge Dependency * Familiarity*			
*3D: Technical Dependency * Power Distance*	supported	supported	supported
*3E: Technical Dependency * Similarity*			
*3F: Technical Dependency * Familiarity*		supported	
*3G: Resource Dependency * Power Distance*			
*3H: Resource Dependency * Similarity*	supported		
*3I : Resource Dependency * Familiarity*	supported		

6.1.1 Factors that predict whether two people collaborate

When the dependent variable, collaboration tie strength, is measured as a binary variable, the results demonstrate how different technological and social factors predict whether two people collaborate or how likely two people are to work on the same project.

As hypothesized, all technological factors positively predict collaboration tie strength. Consistent with the hypotheses, two of the technological factors, technical dependency and resource dependency, positively significantly predict whether two people collaborate, with resource dependency having stronger prediction based on the coefficients of the two variables. Knowledge dependency, however, does not significantly predict whether two people collaborate. In this study, knowledge dependency is measured by the number of listserv threads that involve a dyad of collaborators. I was aware of the comparatively large standard error of this variable. However, after a further look at the data, I found the large standard error is caused by the variations in people's listserv participation. Some people intensively engage in the listserv discussions while others do not. After a thorough examination of the data, I decided to keep the original data intact because removing extreme values may not reflect the true state of the data. More importantly, the extreme values do not distort the analysis result because the extreme values cancel out each other's effects on the outcome, although the standard error seems large because it is square rooted. The study tried methods of both mean values and removal of extreme values, and neither method alters the current result.

It is possible that we would see significant prediction by knowledge dependency had everyone engaged in the listserv discussions. However, it doesn't seem to be the case in the current state of the GENI project. It is reasoned before that people who participate in the listserv threads demonstrate knowledge dependency. However, people who do not participate in the listserv discussion may not be able to publicly demonstrate their knowledge. People may figure out knowledge dependency in other ways, e.g., through conference presentations or conversations, or through referrals, and then they may form collaborations. Future research could try to capture knowledge dependency from a variety of angles. If available, email exchanges between two people could be used to measure knowledge dependency. Other ways of communication, such as meetings, phone calls, discussions on social media and blogs, and the frequency of the communication could also be considered as part of the measure. Furthermore, surveys could be a possible venue to find out whether two people possess the knowledge of each other.

Also as hypothesized, all social factors positively predict collaboration tie strength. However, only familiarity is found to be marginally significant in positively predicting whether two people collaborate, controlling for all other factors. Both power distance and similarity do not significantly predict whether two people collaborate. Overall, the social factors have much less effect in predicting whether two people collaborate compared to the technological factors.

6.1.2 Factors that predict how many times two people collaborate

When the dependent variable is measured as a numeric count variable, the results demonstrate what factors significantly predict the overall number of times that two people collaborate or how many projects two people work on together.

The major difference in the results from this set of analyses compared to the prior analyses is that technical dependency is no longer a significant predictor for collaboration tie strength. Resource dependency and familiarity, as are in the first set of analyses, still significantly positively predict collaboration tie strength. Based on the values of the regression coefficients, resource dependency has slightly stronger prediction for collaboration tie strength than does familiarity.

6.1.3 Factors that predict how many times two people collaborate for those who actually collaborated

Finally, through the analysis on the partial dataset which only includes records that have non-zero collaboration tie strength, the results show what factors predict the number of times that two people collaborate for those who actually collaborated.

Interestingly, in this set of analyses, technical dependency becomes a significant predictor for collaboration tie strength while resource dependency is no longer a significant predictor. This suggests that resource dependency could be a major factor that promotes an initial collaboration tie, as demonstrated in the first set of analyses. However, once people form a collaboration tie, strong technical dependency becomes a more important factor in predicting multiple collaboration ties. Such result suggests to practitioners that people put weight on different factors when choosing collaboration partners at different stages of the CI projects.

Consistent with both prior analyses, familiarity is still a significant predictor for collaboration tie strength. It actually yields a relatively stronger influence than technical dependency in predicting collaboration tie strength.

6.1.4 Interaction effects between social factors and technological factors

Another important goal of this study is to investigate how the technological factors and social factors interact with each other in predicting collaboration tie strength. The results from all three sets of analyses suggest that social factors and technological factors substitute for each other in predicting whether two people collaborate.

In the first set of analyses, where collaboration tie strength is a binary variable, three sets of interactions are found to be significantly negative, i.e., technical dependency and power distance (marginal significant), resource dependency and similarity, and resource dependency and familiarity. When examining the coefficients of the interaction terms together with those of the main effects, the results show that with the increase in power distance, the prediction of technical dependency becomes weaker, and with the increase in similarity and familiarity, the prediction of resource dependency also becomes weaker. Despite all the negative interactions, the substitution effects do not negate the positive prediction of both technical dependency and resource dependency, because even at the highest value of power distance, similarity and familiarity, the coefficients of technical dependency and resource dependency remain positive.

In the rest of the analyses where the collaboration tie strength is measured as a numeric count variable, resource dependency no longer significantly interacts with any social factors. Technical dependency still significantly negatively interacts with power distance, and marginally significantly negatively interacts with familiarity.

Overall, the coefficients of the interaction terms are all very small negative values compared to the main effects of the technological factors, suggesting that the interaction effects are significant, but only slightly change the prediction of the technological factors on how many times two people collaborate. The interaction, in a way, demonstrates how people weigh the importance of technological factors and social factors in making collaboration choices. One way to interpret the result is that when people become more socially attracted to each other based on power distance, social similarity and familiarity, they consider less of the technological factors in making collaboration decisions, although technological factors still matter to them.

6.2 LIMITATIONS AND FUTURE WORK

The findings of study should be interpreted in light of several limitations. In this study, the dependent variable, collaboration tie strength, is conceptualized as the number of projects two people both participate in during a project time period weighted by the project size. As mentioned in prior sections, a collaborative relationship could be examined from many facets, such as duration, frequency and intensity. In this study, collaboration is mainly examined from the perspective of frequency, although the study makes an effort to capture certain facets of collaboration intensity through weighting the collaboration frequency by project size. A major motivation of the study is to find out how the dyadic level of collaboration, as the building block of the community level of collaboration, forms. This study focuses on collaboration frequency rather than other aspects of collaboration, such as duration, intensity and reciprocity.

Second, this study considers only one form of collaboration: collaboration through forming project teams. There may be other ways that people could collaborate without working on the same project team in CI projects. People may engage in knowledge exchange behaviors through phone calls, emails or meetings. However, in the specific context of the GENI project, the foundational units of the overall CI project are project teams. People participate in the GENI project mainly through forming project teams and working on projects. Therefore, in the specific context of a CI project, this study focuses on examining the collaborative relationship formed through projects.

The conceptualization and measurement of collaboration tie strength in this study strike a great similarity to many network research studies in which tie strength is examined through a co-authorship network (Cockburn and Henderson 1998; McFadyen 2009; Newman 2001, 2004; Stephan 1996; Stephan and Levin 1991; Zucker et al. 1996, 1998). The difference in conceptualization of tie strength between those studies and this one is that the former uses joint ownership of papers and the latter uses joint ownership of projects. Just as co-authorship reflects genuine professional interaction between scientists (McFadyen et al. 2009; Newman 2004), so does joint ownership of research projects. Future research could examine other facets of collaboration such as intensity, duration and reciprocity, and also attempt to include more forms of collaboration to garner more insights about collaboration in CI projects.

As a third limitation, the results of this study could not be used to establish causality because the study uses cross-sectional data. Although the theories that motivate the hypotheses imply directions to the relationships between the antecedents and the dependent variable, this directionality cannot be confirmed through the analyses of the data in this study. This study does attempt to control for collaborations that occurred in the prior GENI project spiral and also utilized measurements from prior spirals for certain constructs, such as knowledge dependency, to address the concern of potential tautology; however, to provide evidence of causality, we need longitudinal studies to garner further insights on how collaboration evolves over time. Future research can leverage the data collection and measurement techniques in this study to include data from other project spirals, and utilize other state-of-the-art data analysis techniques, such as time-series analysis, to examine the changes and evolution of collaboration in CI projects over time.

Finally, the study only utilized data from one CI project; therefore, we need to take caution in generalizing the findings to other settings. Future research could extend this study to other CI projects with similar and different project structures.

6.3 IMPLICATIONS FOR PRACTICE

Collaboration is extremely important for the success of a CI project. Taking the GENI project as an example, collaboration between individuals is the building block for community formation. Individuals form project teams, and project teams form the GENI community to build a functioning CI. At the later stage of the GENI project when individual project findings need to roll out to the meso-scale, collaboration becomes even more important. The results of the study provide several implications that may be helpful to the management team and funding providers of CI projects in crafting better strategies to promote collaboration.

First, the main effects of technical dependency, resource dependency and familiarity in different stages positively predict collaboration tie strength to different degrees. In predicting whether two people collaborate, all three have positive effects. In predicting the number of times two people collaborate, resource dependency and familiarity have significant positive predictions. In predicting the number of times two people collaborate for those who actually collaborated, technical dependency and familiarity have significant positive predictions.

In crafting strategies to promote collaboration, project managers may consider increasing the technical dependency, resource dependency and familiarity between project participants. In designing the projects, they could consider funding more interdependent projects rather than independent projects. They could also consider involving more people with access to organizations that could provide important resources for a variety of projects. In addition, letting project participants realize each other's resource access and project dependencies is also very important. Project managers may consider different advertising strategies to increase people's understanding of the project dependencies. Furthermore, the result suggests that familiarity established through common attendance to conferences may help promote collaboration. For the GENI project, although holding the quarterly GECs costs a lot of money and resources, the results of the study suggest that the effort is not made in vain. GPOs should consider continuing the practice of sponsoring these conferences. Future research could further this study by examining what mechanisms utilized in these conferences could most effectively promote collaboration.

The results of the study also increase our understanding toward how technological factors and social factors interact with each other in predicting collaboration tie strength in CI projects. People may take into account a variety of factors when choosing collaboration partners, but how these factors work together in predicting collaboration stays a myth. This study demonstrates that at different levels of social attractiveness, the effects of technological factors are different. For example, at a higher level of familiarity, i.e., when people become more familiar with each other, the increase in technical dependency and resource dependency still corresponds to more collaboration ties, but the increase in the number of collaboration ties is less compared to when they are less familiar. In other words, the increase in familiarity weakens the strength of prediction by technological factors. This finding has an important practical implication for the GENI project managers. As the GENI project progresses, familiarity among project participants increases, and the results show that at a higher level of familiarity, resource dependency has a weaker prediction on collaboration formation (whether two people collaborate) and technical dependency has a weaker prediction on collaboration tie strength (how many times two people collaborate). Therefore, to promote collaboration, increasing the technical dependency and resource dependency helps, but the effects are stronger when people are less familiar. At the earlier stage of the GENI project, when people are not very familiar with each other, they put a lot of weight on technological factors in choosing their collaboration partners. GPOs could intentionally fund more related projects, especially at the earlier stages of the GENI project, to promote collaboration.

Power distance itself does not predict collaboration; however, higher power distance suppresses the prediction of collaboration by technological factors. The practical implication could be that the GENI project managers should be careful about adding hierarchies to the GENI project. Adding more levels of management, especially at the later stages of the project, may potentially negate the effort put into increasing the technical dependency and resource dependency among project participants.

Similarity positively predicts whether two people collaborate and does not predict how many times two people collaborate. The increase in similarity level also suppresses the prediction of technological factors. The finding on similarity may not provide specific guidance to CI project management, but it increases our understanding of how similarity plays out in different respects of collaboration, and also how similarity interacts with technological factors in predicting collaboration. In this study, similarity is measured by an aggregated similarity index. Future research could look further into how different aspects of similarity, whether in age, gender or ethnicity, may affect people's collaboration choice.

In terms of generalizability, the findings of this study could be applied to CI projects with similar structure as that of the GENI project, in which project participants could freely choose their collaborators with a limited amount of organizational administration imparted. The findings could also be generalized to business projects with similar structures. For example, large consulting firms may have employees across the nation or even globally. The consultants may collaborate on a case-by-case basis while they need to conform to the overall organizational rules. For multinational R&D projects, to encourage innovation, an agile method that is used in the GENI project may be desirable. In both of the examples above, mechanisms that help increase the familiarity among project participants, such as conferences and workshops, may be very important to get the stakeholders to understand each other better and thus help them make wise and effective collaboration choices.

6.4 IMPLICATIONS FOR RESEARCH

In addition to the implication for practice, this study contributes to IS research on IT collaboration, especially in the context of CI projects.

First, this study contributes to the literature on collaboration in general and on a collaboration network in IT projects, in particular through the study of the antecedents of collaboration tie strength. Existing research on collaboration provides a variety of theoretical arguments about the antecedents of collaboration with different foci. The conceptualization of collaboration tie strength in this study marries the traditional definition of collaboration and the concept in network research. Such conceptualization adds to our understanding of collaboration network formation in IT projects. In addition, the study integrates prior studies on collaboration antecedents by including both technological factors and social factors and extends existing research through examining the interactions among these factors, thus unpacking the dynamics and complexities involved in the way people forming collaboration ties.

Second, the work contributes to the emerging body of research on CI projects in IS research. There have been increasing calls for IS research on CI projects (Kirsch and Slaughter 2013a) and this study is an effort to respond to the calls. Based on an extensive literature review, most work done so far on CI projects tends to be conceptual, and this study attempts to expand our understanding of CI projects in the empirical domain. Through studying the collaboration of project participants in the GENI project, the research demonstrates that, in general, technological and social factors both positively predict collaboration tie strength. Technological factors have stronger prediction than the social factors; however, the strength of prediction becomes weaker with an increase in social factors, although the overall prediction by all factors remains positive.

Third, the study examines both the relational and accumulative perspectives of CI project collaboration. In particular, the study provides insights on what factors predict whether two people are on the same project team, as well as what factors predict how many projects two people work on together. The current study could not speak to which perspective is more important, which could be a direction for future research. However, based on the observation of the project and interaction with the project participants, we could reason that the relational perspective is important to achieve at-scale collaboration for the overall CI project, while the cumulative perspective speaks to how effective and successful a collaboration tie turns out. This is because repeated collaboration often is a decision made based on a person's subjective evaluation of prior collaboration success. Future study could look into the different perspectives of collaboration and examine how each perspective contributes to the individual project success, as well as the overall CI project success, and what perspective of collaboration matters more at different stages of a CI project.

6.5 CONCLUSION

This study examines the technological antecedents (knowledge dependency, resource dependency and technical dependency) and the social antecedents (power distance, familiarity and similarity) for collaboration tie strength among CI project participants. It not only studies the individual effects of these antecedents, but also investigates the interactions among them. The findings demonstrate that the technological factors, resource dependency and technical dependency, and the social factor, familiarity, all significantly positively predict collaboration tie strength, with the technological factors yielding stronger prediction. The findings also suggest that the increase in similarity and familiarity both suppresses the positive prediction of resource dependency on whether two people collaborate, and that the increase in power distance suppresses the positive prediction of technical dependency on how many times two people collaborate for those who actually collaborated.

The study contributes to collaboration research by integrating social antecedents and technological antecedents for collaboration tie strength. It answers the calls for more research on CI projects by conducting an empirical study in the specific context of the GENI project. Furthermore, the research contributes to practice by providing insights on what people may consider when choosing collaboration partners and how people may weigh the importance of social factors and technological factors when making collaboration decisions. These insights could potentially benefit CI project managers in crafting better strategies to promote collaboration, depending on the project stages and project needs. Future research could explore more CI project phases in more CI project settings.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation under Grant Nos. 909611 and 909833. Any opinions, findings, and conclusions or recommendations expressed in the material are those of the author and do not necessarily reflect the views of the National Science Foundation. The participation of the GENI community in this research is acknowledged and appreciated. Dr. Laurie Kirsch and Sandra Slaughter shared with me the interview data, making the research in this article possible.

BIBLIOGRAPHY

Allport, G. W. (1985). *The historical background of social psychology. In G. Lindzey, and E. Aronson, (Eds.), Handbook of Social Psychology, 1, (3), 1-46.*

Aksulu A., & Wade, M. R. (2010). A Comprehensive Review and Synthesis of Open Source Research. *Journal of the Association for Information Systems. 11*(11), 576-656.

Aral, S., & Walker D. (2014). Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment. *Management Science*. 60(6),1352-70.

Aram, J.D., & Morgan, C.P. (1976). The Role of Project Team Collaboration in R & D Performance. *Management Science*. 22(10), 1127-37.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., Messina, P., Ostriker, J. P., & Wright, M. H. (2003). *Final report of the NSF blue ribbon advisory panel on cyberinfrastructure: Revolutionizing science and engineering through cyberinfrastructure*. From http://hdl.handle.net/10150/106224

Banker, R. D., Bardhan, I., & Asdemir, O. (2006). Understanding the Impact of Collaboration Software on Product Design and Development. *Information Systems Research, 17*(4), 352-373.

Berman, F. (2008). Got data?: A guide to data preservation in the information age. *Commun. ACM, 51*(12), 50-56.

Bietz, M. J., Baumer, E. P. S., & Lee, C. P. (2010). Synergizing in cyberinfrastructure development. *Computer Supported Cooperative Work (CSCW), 19*(3-4), 245-281.

Bird, C., Nagappan, N., Gall, H., Murphy, B., Devanbu, P. (2009), *Putting It All Together: Using Socio-technical Networks to Predict Failures*, 20th International Symposium on Software Reliability Engineering, Mysuru, Karnataka, 109-119.

Boehm, B. (1986). A spiral model of software development and enhancement. *SIGSOFT Softw. Eng. Notes, 11*(4), 14-24.

Boh, W. F., Ren, Y. Q., Kiesler, S., & Bussjaeger, R. (2007). Expertise and collaboration in the geographically dispersed organization. *Organization Science, 18*(4), 595-612.

Boh, Fong W., Slaughter, S. A., & Espinosa, J. A. (2007) Learning from experience in software development: A multilevel analysis. *Management Science*. 53(8), 1315-31.

Boland, R. J. (1979). Control, causality and information system requirements. *Accounting, Organizations and Society. 4*(4), 259-72.

Bostrom, R. P., & Heinen, J. S. (1977). MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes. *MIS Quarterly*, *1*(3), 17-32.

Brass, D. J. (1985). Men's and women's networks: A study of interaction patterns and influence in an organization. *The Academy of Management Journal*, *28*(2), 327-43.

Brooks, F. P. (1995). *The mythical man-month: essays on software engineering:* Addison-Wesley Pub. Co.

Brown, J. S. & Duguid, P. (1998). Organizing Knowledge, *California Management Review*, *40*(3).

Brown, H. G., Poole, M. S., & Rodgers, T. L. (2004). Interpersonal Traits, Complementarity, and Trust in Virtual Collaboration. *Journal of Management Information Systems*, *20*(4), 115-137.

Burt, R. S. (1991). Measuring age a structural concept. *Social Networks*, *13*(1), 1-34.

Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA.

Byrne, D. (1971). *The Attraction Paradigm*. Academic Press, New York.

Cataldo, M., Herbsleb, J. D., & Carley, K. M. (2008). Socio-technical congruence: A framwork for assessing the impact of technical and work dependencies on software development productivity. In *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '08, 2-11, New York, NY, USA, ACM.

Cecez-Kecmanovic, D., Galliers, R. D., Henfridsson, O., Newell, S., & Vidgen, R. (2014). The sociomateriality of information systems: current status, future directions. *MIS Quarterly*, 38(3), 809-830.

Cockburn, I. & Henderson, R. (1998). Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. *The Journal of Industrial Economics*, *46*(2), 157-182.

Cramton, C. D. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science*, *12*(3), 346-371.

Cummings T. G. (1978). Self-Regulating Work Groups: A Socio-Technical Synthesis. *Academy of Management Review*, 3(3), 625-34.

de la Flor, G., Ojaghi, M., Martínez, I. L., Jirotka, M., Williams, M. S., & Blakeborough, A. (2010). Reconfiguring practice: The interdependence of experimental procedure and computing infrastructure in distributed earthquake engineering. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4073-4088.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51*(3), 629-636.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Williams, R. (2009). Introduction: An Agenda for Infrastructure Studies. [Article]. *Journal of the Association for Information Systems, 10*(5), 364-374.

Espinosa A, Kraut R, Slaughter S, Lerch J, Herbsleb J, & Mockus A. (2002). Shared mental models, familiarity, and coordination: A multi-method study of distributed software teams. *ICIS 2002 Proceedings*. 39.

Espinosa J. A., Slaughter S. A., Kraut R. E., & Herbsleb J. D. (2007b). Team knowledge and coordination in geographically distributed software development. *Journal of Management Information Systems*, 24(1), 135-69.

Espinosa J. A., Slaughter S. A., Kraut R. E., & Herbsleb J. D. (2007a). Familiarity, complexity, and team performance in geographically distributed software development. *Organization Science*. 18(4), 613-30.

Fairhurst, G.T., Green, S., Courtright, J. (1995). Inertial Forces and the Implementation of a Socio-Technical Systems Approach: A Communication Study. *Organization Science*, 6(2),168-85.

Finholt, T. A., & Birnholtz, J. P. (2006). If we build it, will they come? The cultural challenges of cyberinfrastructure development. In W. S. Bainbridge & M. C. Roco (Eds.), *Managing Nano-Bio-Info-Cogno Innovations* (pp. 89-101). Netherlands: Springer.

Fulmer, J. (2009). "What in the world is infrastructure?". *PEI Infrastructure Investor* (July/August), 30–32.

Granovetter, M. S. (1973). The Strength of Week Ties. American Journal of Sociology. 78(6), 1360-80.

Gregory, R. W., Beck, R., & Keil M. (2013). Control Balancing in Information Systems Development Offshoring Projects *MIS Quarterly*. *37*(4),1211-1232.

Hahn J., Moon, J. Y., Zhang, C. (2008). Emergence of New Project Teams from Open Source Software Developer Networks: Impact of Prior Collaboration Ties. *Information Systems Research*. *19*(3), 369-91.

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.

Hardy, C., Lawrence, T. B., & Grant, D. (2005). Discourse and Collaboration: The Role of Conversations and Collective Identity. *Academy of Management Review, 30*(1), 58-77.

Hart, D. (2014). A Grand Convergence. http://www.nsf.gov/news/special_reports/cyber/agrand.jsp

Henderson J. C., & Lee S. (1992). Managing I/S Design Teams: A Control Theories Perspective. *Management Science. 38*(6): 757-77.

Herbsleb, J. D. (2007, May). Global software engineering: The future of socio-technical coordination. In *2007 Future of Software Engineering* (pp. 188-198). IEEE Computer Society.

Hofstede, G. (1983). The cultural relativity of organizational practices and theories. *Journal of international business studies*, *14*(2), 75-89.

Hossain, L., & Wigand, R. T. (2004). ICT Enabled Virtual Collaboration through Trust. *Journal of Computer-Mediated Communication, 10*(1), 00-00.Ibarra, H. (1992). Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative Science Quarterly, 37*(3),422-47.

Ibarra, H. (1995). Race, opportunity, and diversity of social circles in managerial networks. *The Academy of Management Journal, 38*(3), 673-703.

Ibarra, H. (1997). Paving an alternative route: Gender differences in managerial networks. *Social Psychology Quarterly, 60*(1), 91-102.

Katz, A., & Te'eni, D. (2007). The contingent impact of contextualization on computer-mediated collaboration. *Organization Science, 18*(2), 261-279.

Kelman, H. C. (1958). Compliance, identification, and internalization: Three processes of attitude change. *Journal of conflict resolution*, 51-60.

Kirkman, B. L., Chen, G., Farh, J-L, Chen, Z. X., Lowe, K.B. (2009). Individual Power Distance Orientation and Follower Reactions to Transformational Leaders: A Cross-Level, Cross-Cultural Examination. *Academy of Management Journal*, 52(4), 744-64.

Kirsch, L. J. (1997). Portfolios of control modes and IS project management. *Information Systems Research. 8*(3), 215-39.

Kirsch, L. J., & Slaughter, S. A. (2013a). Managing the unmanageable: How IS research can contribute to the scholarship of cyber projects. *Journal of the Association for Information Systems*, *14*(4), 2.

Kirsch, L. J., & Slaughter, S. A. (2013b). Beneath the surface of research projects. *International Innovation, North America, May*.

Kotlarsky, J., & Oshri, I. (2005). Social ties, knowledge sharing and successful collaboration in globally distributed system development projects. *Eur J Inf Syst*, *14*(1), 37-48.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Universtiy of Chicago Press, Chicago.

Lawrence, B. S. (2006). Organizational reference groups: A missing perspective on social context. *Organization Science*, 17(1), 80-100.

Laumann, E. O. (1966). *Prestige and Association in an Urban Community*. Bobbs-Merrill, Indianapolis.

Lee, C. P., Dourish, P., & Mark, G. (2006). The human infrastructure of cyberinfrastructure.

Leonardi, P.M. (2012). Materiality, Sociomaterialtiy, and Socio-Technical Systems: What Do These Mean? How Are They Related? Do We Need Them? In P. M. Leonardi, B.A. Nardi, & J. Kallinikos (Eds.) *Materiality and Organizing: Social Interaction in a Technological World* (pp. 25-48). Oxford University Press.

Levina, N., & Vaast, E. (2005). The Emergence of Boundary Spanning Competence in Practice: Implications for Implementation and Use of Information Systems. *MIS Quarterly, 29*(2), 335-363.

Levina, N., & Vaast, E. (2008). Innovating or doing as told? status differences and overlapping boundaries in offshore collaboration. *MIS Quarterly, 32*(2), 307-332.

Malhotra, Y., & Galletta, G. F. (1999). Extending the technology acceptance model to account for social influence: Theoretical bases and empirical validation. *Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*. IEEE, 1999.

Manz, C. C., Stewart, G. L. (1997). Attaining Flexible Stability by Integrating Total Quality Management and Socio-Technical Systems Theory. *Organization Science*, 8(1), 59-70.

Markus, L. (1984). *Systems in Organizations: Bugs and Features*: Ballinger.

Markus, L. (2007). The governance of free/open source software projects: monolithic, multidimensional, or configurational? *Journal of Management & Governance. 11*(2),151-63.

Marsden P. V. (1987). Core discussion networks of Americans. *American Sociological Review*. 52(1), 122-31.

Marsden P. V. (1988). Homogeniety in confiding relationships. *Social Networks*, 10(1), 57-76.

Marsden P. V., & Campbell K. E. (1984). Measuring Tie Strength. *Social Forces*. 63(2), 482-501.

McFadyen M. A., Semadeni, M., & Cannella, A. A. (2008). Value of Strong Ties to Disconnected Others: Examining Knowledge Creation in Biomedicine. *Organization Science, 20*(3), 552-64.

McPherson, J. M., & Smith-Lovin, L. (1986). Sex segregation in voluntary associations. *American Sociological Review, 51*(1), 61-79.

McPherson, J. M., & Smith-Lovin, L. (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review. 52*(3), 370-9.

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

Mintzberg, H. (1979) The Structuring of Organisations B2 - The Structuring of Organisations. Englewood Cliffs, NJ: Prentice-Hall.

Moody, G. D., Kirsch, L. J., Slaughter, S. A., Dunn, B. K., & Weng, Q. (2016). Facilitating the Transformational: An Exploration of Control in Cyberinfrastructure Projects and the Discovery of Field Control. *Information Systems Research 27*(2), 324-346.

Moody, J. (2001). Race, school integration, and friendship segregation in America. *American Journal of Sociology*, *107*(3), 679-716.

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

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

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

Oliveira, N., Lumineau, F. (2017). How Coordination Trajectories Influence the Performance of Interorganizational Project Networks. *Organization Science*,

Orlikowski, W. J. (1991) Integrated information environment or matrix of control? The contradictory implications of information technology. *Accounting, Management and Information Technology, 1*(1), 9-42.

Paul, D. L., & McDaniel, R. R., Jr. (2004). A Field Study of the Effect of Interpersonal Trust on Virtual Collaborative Relationship Performance. *MIS Quarterly, 28*(2), 183-227.

Peters, L. M., & Manz, C. C. (2007). Identifying antecedents of virtual team collaboration. *Team Performance Management, 13*(3/4), 117-129.

Pfeffer J, Salancik GR. (1978). *The external control of organizations: A resource dependence perspective*: Stanford University Press.

Polanyi, M. (1966). *The tacit dimension*. Garden City, N.Y: Doubleday.

Reagans, R. (2011). Close Encounters: Analyzing How Social Similarity and Propinquity Contribute to Strong Network Connections. *Organization Science, 22*(4), 835-49.

Reagans, R., & McEvily, B. (2003). Network Structure and Knowledge Transfer: The Effects of Cohesion and Range. *Administrative Science Quarterly, 48*(2), 240-267.

Santoro, F. M., Borges, M. R. S., & Rezende, E. A. (2006). Collaboration and knowledge sharing in network organizations. *Expert Systems with Applications, 31*(4), 715-727.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*(3), 437-53.

Schwab, A., & Miner, A. S. (2008). Learning in Hybrid-project Systems: The Effects of Project Performance on Repeated Collaboration. [Article]. *Academy of Management Journal, 51*(6), 1117-1149.

Schrage, M. (1990). *Shared minds: The new technologies of collaboration*. New York: Random House.

Shaw, M., & Garlan, D. (1996) *Software Architecture: Perspectives on an Emerging Discipline*. Prentice Hall.

Sosa, M. E. (2010). Where Do Creative Interactions Come From? The Role of Tie Content and Social Networks. *Organization Science*. 22(1):1-21.

Stephan, P. E. (1996). The economics of science. *Journal of Economic Literature. 34*(3), 1199-235.

Stephan, P. E. & Levin, S. G. (1991). Inequality in scientific performance: Adjustment for attribution and journal impact.

Stewart, K. J., & Gosain, S. (2006). The impact of ideology on effectiveness in open source software development teams. *MIS Quarterly. 30*(2), 291-314.

Tillquist, J., King, J. L., & Carson W. (2002). A Representational Scheme for Analyzing Information Technology and Organizational Dependency. *MIS Quarterly*. 26(2), 91-118.

Tortoriello, M., Reagans, R., & McEvily, B. (2012). Bridging the Knowledge Gap: The Influence of Strong Ties, Network Cohesion, and Network Range on the Transfer of Knowledge Between Organizational Units. *Organization Science. 23*(4),1024-39.

Trist, E. (1981). The evolution of socio-technical systems. Occasional paper.

Turner, J. C. (1991). *Social influence*: Open University Press Milton Keynes.

Ulrich, D., & Barney, J. B. (1984). Perspectives in Organizations: Resource Dependence, Efficiency, and Population. *The Academy of Management Review. 9*(3), 471-81.

Wagner, C. (2004). Wiki: A technology for conversational knowledge management and group collaboration. *Communications of the Association for Information Systems, 13*(19), 265-289.

Wainfan, L., & Davis, P. K. (2004). *Challenges in virtual collaboration: Videoconferencing, audioconferencing, and computer-mediated communications (Vol. 273)*: Rand Corporation.

Wei, C., Sun, X., Liu, J., Zhou, C., & Xue, G. (2016). High Power Distance Enhances Employees' Preference for Likable Managers: A Resource Dependency Perspective. *Frontiers in Psychology*, *7*, 2066. http://doi.org/10.3389/fpsyg.2016.02066

Wellman, J. L. (2009). *Organizational Learning: How Companies and Institutions Manage and Apply Knowledge.* Palgrave Macmillian.

Winter, S., Berente, N., Howison, J., Butler, B. (2014). Beyond the organizational 'container': Conceptualizing 21st century sociotechnical work. *Information and Organization*, 24(4), 250-69.

Wilkinson, D. M., & Huberman, B. A. (2007). *Cooperation and quality in wikipedia*.

Wood, D. J., & Gray, B. (1991). Toward a Comprehensive Theory of Collaboration. *The Journal of Applied Behavioral Science*, *27*(2), 139-162.

Yamaguchi, K. (1990). Homophily and social distance in the choice of multiple friends: An analysis based on conditionally symmetric log-bilinear association model. *Journal of the American Statistical Association, 85*(410), 356-66.

Zenger, T. R., & Lawrence, B. S. (1989). Organizational demography: The differential effects of age and tenure distributions on technical communication. *The Academy of Management Journal, 32*(2), 353-376.

Zucker, L. G., Darby, M. R., & Brewer, M. B. (1998). Intellectual human capital and the birth of U.S. biotechnology enterprises. *The American Economic Review. 88*(1), 290-306.

Zucker, L. G., Darby, M. R., Brewer, M. B., & Peng, Y. (1996). Collaboration structure and information dilemmas in biotechnology: organizational boundaries as trust production. In R. Kramer & T. Tyler *Trust in organizations: Frontiers of theory and research* (pp. 90-113). Thousand Oaks, CA: SAGE Publications Ltd.