CARNEGIE MELLON UNIVERSITY

TEPPER SCHOOL OF BUSINESS

DOCTORAL DISSERTATION

Knowledge Sharing and Creation through Social Media in Organizations

Submitted to the Tepper School of Business in Partial Fulfillment of the Requirements for the Degree of DOCTOR OF PHILOSOPHY in Industrial Administration

By Elina Hyeunjung Hwang

APRIL 2015

DISSERTATION COMMITTEE:

Linda Argote (co-chair)

Param Vir Singh (co-chair)

David Krackhardt

Brandy Aven

Acknowledgements

To be completed

Abstract

Organizations are becoming more creative in incorporating technologies to aid their businesses, for example, by building collaboration networks with customers for innovative ideas and by utilizing online communities to mobilize knowledge among their employees. In my dissertation, I examine how such networks among employees or customers empowered by information technology influence the way organizations learn and innovate. My dissertation consists of the following three essays.

The *first* essay empirically examines whether knowledge flows within or across boundaries and how such tendencies dynamically change as a knowledge provider gains more experiences in an internal online knowledge community. Although the previous literature has suggested that geographic and social boundaries disappear online, I hypothesize that they remain because participants prefer to share knowledge with others who share similar attributes, due to the challenges involved in knowledge sharing in an online community. Further, I propose that as participants acquire experience in exchanging knowledge, they learn to rely more on expertise similarity and less on categorical similarities such as location similarities. As a result, boundaries based on categorical attributes are expected to weaken, and boundaries based on expertise are expected to strengthen, as participants gain experience in the online community. Empirical support for this argument is obtained from analyzing a longitudinal dataset of an internal online knowledge community at a large multinational IT consulting firm.

The *second* essay investigates the complementarity of individuals' activities between two crowdsourcing communities: a customer support community and an innovation crowdsourcing community. A tie formed between a helper and a help-seeker at a customer support crowdsourcing community represents valuable information flow for new product ideation because: (a) it represents a flow of solution information from a helper to a help-seeker, and (b) it represents a flow of a help-seeker's information about his/her needs to a helper. By utilizing a natural language processing technique, I construct each individual's information network based on their helping activities, and examine how the structure of their information network, in terms of breadth and depth, affects their new product ideation outcomes at an innovation

crowdsourcing community. My analysis reveals that helping activities at a customer support community help individuals to create better quality ideas at an innovation community. Specifically, generalists, who have provided solutions on diverse problem areas, are likely to create more original ideas. Yet, those generalists who have only shallow knowledge across diverse domain areas do not perform significantly better than non-generalists in their ability to create ideas that are later implemented by a company. Only those generalists who possess expert knowledge in at least one domain area tend to outperform non-generalists.

In the *third* essay, I examine membership dynamics in online knowledge communities. This essay extends the first essay by examining whether individuals' decision of how much to contribute to an online knowledge community is based on the decisions of other participants in her/his ego-network (beyond a dyadic relation studied in the first essay). Humans have intrinsic tendency for consensus: people want to follow what others do. I propose that individuals have stronger motivation to get engaged in online community activities if their virtual neighbors, with whom they have interacted over an online community, are active. In addition, I propose that this herding tendency become stronger if their virtual neighbors are geographically proximate to them. I empirically test this conjecture, and discuss the impact of such herding behavior on the design of an online community and on the evolution of an online community population.

Table of Contents

List of Figures

List of Tables

Chapter 1.

Introduction

Organizations are becoming more creative in incorporating technologies to aid their businesses, for example, by building collaboration networks with customers for innovative ideas and by utilizing online communities to mobilize knowledge among their employees. Organizations are also utilizing online communities (e.g., Facebook, Twitter) as a marketing channel. Despite its prevalence, our understanding on the implications of such social media use in formal organizations is limited. In this dissertation, I attempt to advance our knowledge by examining how such networks among employees or customers empowered by information technology influence the way organizations learn and innovate.

This dissertation consists of three essays. The setting of the first and the third essays is an internal online knowledge community. I use proprietary data of a large multinational consulting firm's internal online knowledge community where employees exchange knowledge virtually. The setting of the second essay is an organization-hosted innovation crowdsourcing community in which customers propose new product ideas and comment on the ideas of others. I constructed a dataset from a telecommunication company's online crowdsourcing communities. These data are analyzed first by identifying networks among people and extracting information exchanged over such networks through machine-learning techniques, and then by conducting econometric analyses to identify the factors that influence the creation and sharing of knowledge. The following sections briefly overview the three essays in my dissertation.

The First Essay

The *first* essay, entitled “Knowledge Sharing in Online Communities: Learning to Cross Geographic and Hierarchical Boundaries,” empirically examines whether knowledge flows within or across boundaries and how such tendencies dynamically change as a knowledge provider gains more experiences in an internal online knowledge community.

Information technology opens opportunities for organizations to make the most of their existing knowledge base. The great opportunity comes from the *theoretical* potential of information technology to bridge boundaries and to connect otherwise unconnected people. Due to this potential, organizations are increasingly utilizing electronic networks to promote knowledge sharing among their employees. According to a survey by McKinsey, more than 50% of surveyed firms have adopted some type of social networking tool to facilitate knowledge sharing, compared to only 28% in 2009. One of the most popular social technologies adopted so far is an online knowledge-exchanging community in the form of a discussion bulletin board.

Mobilizing knowledge across boundaries has been touted as an advantage of online knowledge communities. An online knowledge community is expected to break physical knowledge silos because the Internet can eliminate spatial distance. Further, an online knowledge community is also expected to reduce social boundaries. Because participants do not interact face-to-face, less social information (e.g., about appearance or status) is available. The reduced level of social information can shrink the social distance among dissimilar people, resulting in increased interaction among them. Even though technology provides affordances that promote boundary-spanning knowledge sharing, whether this occurs is an empirical question whose answer depends on how employees use the technology. This essay aims to advance our understanding of whether online communities promote boundary-spanning knowledge flow by empirically investigating knowledge sharing dynamics in a firm-hosted online community.

Building upon the theories of experience-based learning, common ground, and diversity, I propose and empirically test a theory about how individuals choose knowledge-sharing partners in an online knowledge community. I argue that participants tend to choose similar others who are within boundaries, because interpersonal similarity increases attraction as well as common ground, which reduces the risks and challenges associated with sharing knowledge. Likewise, I argue that as participants learned about others' expertise by observing their knowledge-sharing behaviors for extended periods of time, participants will favor those who are similar in expertise over those who are similar in categorical attributes.

Empirical support for this argument is obtained from analyzing a longitudinal dataset of an internal online knowledge community at a large multinational IT consulting firm. Consistent with my predictions, I found that individuals prefer to share knowledge with similar others. Moreover, I found that as individuals learn about others' expertise information by observing their extended interactions, they increasingly favor those who are similar in expertise and decreasingly favor those who are similar in categorical attributes as knowledge-sharing partners. Consequently, categorical boundaries weaken, whereas boundaries around expertise strengthen as participants accumulate more experience in an organizational online knowledge community.

The Second Essay

The *second* essay, entitled “Jack of All, Master of Some: Knowledge Network and Innovation,” investigates the complementarity of individuals’ activities between two crowdsourcing communities: a customer support crowdsourcing community and an innovation crowdsourcing community.

With advancement of information technology, many organizations such as Starbucks, BMW, and Dell are now inviting their customers to suggest new product ideas through a practice called ‘innovation crowdsourcing.’ Through an innovation crowdsourcing community, users can propose new product or service ideas directly to a company. Most innovation crowdsourcing communities also have a separate customer support crowdsourcing community within the same platform. At a customer support crowdsourcing community, users can help each other to try to figure out solutions to the problems that they are facing.

A handful of previous studies have documented characteristics of successful individuals at innovation crowdsourcing communities. This essay extends prior work by examining the *complementarity* of individuals’ activities between two crowdsourcing communities: how individuals’ helping activities at a customer support community influence their new product ideation outcomes at an innovation crowdsourcing community. I focus on helping activities because a tie formed between a helper and a help-seeker at a customer support crowdsourcing community represents valuable information flow for new product ideation. (a) It represents a flow of solution information from a helper to a help-seeker. (b) It represents a flow of a help-seeker’s information about his/her needs to a helper. By utilizing a natural language processing technique, I construct each individual’s information network based on their helping activities, and examine how the structure of their information network, in terms of breadth and depth, affects their new product ideation outcomes at an innovation crowdsourcing community. Here, breadth refers to the scope of information one has and depth refers to the level of understanding one has in a domain area. Deep knowledge implies that the individual has expertise.

I propose that individuals who have engaged in helping others on broader problem areas (generalists) are more likely to create original ideas in an innovation crowdsourcing community because diverse information is available for them to recombine in novel ways. When there are diverse ingredients, it is more likely that the resulting recombination is new. In addition, I propose that the quality of ideas that generalists create is likely to vary: some of their ideas would be extremely high quality while other ideas would be low quality. Whereas diverse information increases the upside potential of idea quality by improving novelty aspect of ideas, I expect that it may also increase the downside potential of idea quality because individuals are less likely to effectively utilize information as the number of pieces of information grows. As a result, I finally propose that, only those generalists who possess expert knowledge in at least one domain area are likely to outperform non-generalists. In other words, without any expertise, I do not expect that generalists are superior to non-generalists in their ability to create ideas that are later implemented by a company.

At an innovation crowdsourcing community hosted by a British telecommunication company, I empirically tested the theory by evaluating 8,110 new product ideation "projects" in a real world setting. My analysis reveals that helping activities at a customer support community help individuals to create better quality ideas at an innovation community. Specifically, generalists, who have provided solutions on diverse problem areas, are likely to create more original ideas. Yet, those generalists who have only shallow knowledge across diverse domain areas do not perform significantly better than non-generalists in their ability to create ideas that are later implemented by a company. Only those generalists who possess expert knowledge in at least one domain area tend to outperform non-generalists.

The Third Essay

In the *third* essay, I examine membership dynamics in online knowledge communities. This essay extends the first essay by examining individuals' contribution decision based on the decisions of other participants in her/his ego-network (beyond a dyadic relation studied in the first essay).

Humans have intrinsic propensity for consensus. When people make decisions they tend to refer to the decisions made by previous decision makers, which results in behavior patterns of herding. It has been found that people tend to mimic others to make purchase decisions, investment decisions, product rating decisions, and an organ transplant decisions. Despite its prevalence, we have a limited understanding on how such herding tendency might affect individuals' contribution decision to online communities. In this essay, I explore whether participants of online communities also exhibit this propensity to "follow" others when they decide how much to contribute to online communities.

I propose that individuals are likely to contribute more if their virtual neighbors, with whom they have interacted over an online community, are active. I suggest that an individual's benefit from an online community is tightly linked to the activities of one's virtual neighbors. Participants are the core assets of online communities because community contents are generated by participants. One's virtual neighbors are likely to have resources that an individual need because virtual ties are formed based on common interest. In other words, if an individual's virtual neighbors are active in the community, there is higher chance that the individual can get access to the resources he or she needs.

Additionally, I propose that this herding tendency become stronger if their virtual neighbors are also geographically proximate to them. As the Internet brought possibilities to easily connect over distance, propinquity (geographic proximity) seemed to lose its role in human interaction, at least in virtual settings. Nonetheless, I argue that individuals will be motivated to contribute more if their active virtual neighbors are also geographically proximate because they regard other participants who are physically proximate more socially important. It is expected that reputation from virtual neighbors is more valuable if virtual neighbors are also geographically proximate because one can reasonably expect that his or her online reputation can spill over to offline. Further, the desirability of actively engaging in online communities is not certain to individuals. Activities of virtually and geographically proximate others are expected to send more accurate signal to individuals regarding the desirability because geographical proximity adds more commonality such as cultural and subunit similarities.

Using field data from an enterprise online community, I empirically tested this conjecture. Consistent with my predictions, the results show individuals' herding tendency to virtual neighbors in an online community. Further, the results show that this herding tendency becomes stronger if one's virtual neighbors are geographically proximate. I discuss the implications of such herding behavior on the online community design and on the evolution of online community population.

Chapter 2.

Knowledge Sharing in Online Communities: Learning to Cross Geographic and Hierarchical Boundaries[1]

2.1. Introduction

The ability to utilize existing knowledge is critical for an organization's success (Argote 2012, Grant 1996, Zander and Kogut 1995). Knowledge, however, is often insulated by boundaries that make it challenging to locate, acquire, and assimilate. By boundary, we mean a border that divides one group from another. The boundary could be geographic. For instance, many firms are organized on a geographically distributed basis, where considerable knowledge is developed and accumulated locally, which can result in knowledge silos demarcated by a physical boundary. The boundary could also be social. Individuals share information more with members of their own social group than with members of other social groups (McPherson et al. 2001).

The web plays a role in knowledge management as a solution for bounded knowledge sharing. According to a survey by McKinsey, more than 50% of surveyed firms have adopted some type of social networking tool to facilitate knowledge sharing, compared to only 28% in 2009 (McKinsey 2013). One of the most popular social technologies adopted so far is an online knowledge-exchanging community in the form of a discussion bulletin board. An online knowledge community is also known as a knowledge forum, a social question and answer

[1] This essay is a joint work with Param Vir Singh and Linda Argote.

(Q&A) forum, or a community Q&A site. An online knowledge community is a virtual space where information needs can be presented in the form of natural language (Shah 2010). Most online knowledge communities offer text-based, interactive, and asynchronous communication and rely on the voluntary participation of users to generate content.

Mobilizing knowledge across boundaries has been touted as an advantage of online knowledge communities. An online knowledge community is expected to break physical knowledge silos because the Internet can eliminate spatial distance (Friedman 2006). An online knowledge community can also reduce social boundaries. Because participants do not interact face-to-face, less social information (e.g., about appearance or status) is available. The reduced level of social information can shrink the social distance among dissimilar people, resulting in increased interaction among them (Kiesler et al. 1984, Sproull and Kiesler 1986, 1991).

The availability of electronic communication technologies, however, does not guarantee that knowledge will be shared (Alavi and Leidner 2001, Orlikowski 1996). Technology provides affordances that enable knowledge sharing, but the effects of the technology depend on how it is used (Zammuto et al. 2007). Technology makes it easier for employees to share knowledge across boundaries, but whether this occurs is an empirical question whose answer depends on how employees use the technology. Our research aims to advance understanding of the dynamics of knowledge sharing by empirically examining knowledge sharing in a firm-hosted online community.

Online knowledge communities are interesting to researchers. One stream of research on online knowledge communities studies individual motivations to contribute (Constant et al. 1996, Faraj and Johnson 2011, Jeppesen and Frederiksen 2006, Tausczik and Pennebaker 2012, and Wasko and Faraj 2005). For example, Faraj and Johnson (2011) examined how diverse individual-level motivations aggregate into network exchange patterns and found that direct reciprocity and indirect reciprocity govern network exchange patterns in online communities. Another popular stream of research focuses on developing algorithms that automatically locate experts within online knowledge communities (e.g., Pal et al. 2011, Riahi et al. 2012, Zhang et al. 2007). Researchers have also conducted studies that examined community-level constructs, such as the effect of membership size on the sustainability of an email-based online community (Butler 2001) and the impact of co-membership on online community growth (Wang et al. 2013).

Scholars have also explored the effects of various forms of similarity on online interactions. In an experimental study using video conferencing and instant messaging, Bradner and Mark (2002) found that people are more likely to deceive, be less persuaded by, and initially cooperate less with communicating partners whom they believe to be far away. Likewise, interest similarity was found to predict future interactions in an online knowledge collaboration site (e.g., Wikipedia) and an online blog portal (e.g., LiveJournal) (Crandall et al. 2008, Lauw et al. 2010). Similarities in various sociodemographic attributes (e.g., ethnicity, religion, age, nationality, and marital status) were also found to predict friendship ties in social networking sites such as MySpace and Facebook (e.g., Mislove et al. 2010, Skopek et al. 2011, Thelwall 2009). In contrast, Bisgin et al. (2012) found that interest similarity had only a marginal effect on tie formation in an online blog portal (i.e., BlogCatalog) and a social networking site (i.e., Last.fm). Ludford et al. (2004) even found a negative relationship between opinion similarities of group members and their participation rates in an online movie discussion forum. These mixed results of previous studies suggest that different types of similarities (e.g., similarities in interest, demographics, opinions) might have different effects on interpersonal interactions in online communities.

We extend this prior work by examining how different dimensions of similarity simultaneously affect knowledge sharing in a firm-hosted knowledge community and how the effects change dynamically as participants gain experience in the system. We apply theories of common ground, experiential learning, and diversity to examine whether participants share knowledge with similar or dissimilar others. Furthermore, we examine how knowledge-sharing behavior—the act of providing an answer to other participants—changes as participants acquire experience exchanging knowledge in the online community.

We propose that participants prefer to share knowledge with others with whom they have common ground, and we use joint characteristics with their communicating partners to assess the level of common ground. In particular, we examine how different kinds of dyadic similarities (categorical and expertise) affect the dyad's likelihood of sharing knowledge in an online knowledge community. By categorical similarity, we mean that participants belong to the same category. The two categories we investigate are geographical location and hierarchical status. By expertise similarity, we mean the extent to which participants' expertise about technical matters addressed in the community is similar. We propose that both categorical and expertise similarities drive knowledge sharing, because similarities reduce the risks and challenges associated with sharing knowledge. Moreover, we hypothesize that the effect of categorical similarity (location and hierarchical status) on the likelihood of knowledge sharing decreases while the effect of expertise similarity increases as a knowledge provider accumulates more experience in the system. Provider experience is expected to substitute for categorical similarity, because the familiarity and common ground obtained through experience substitute for the familiarity and common ground provided by membership in the same category. By contrast, experience enables participants to learn about others' expertise and identify those with whom they have mutual knowledge. Thus, experience complements expertise similarity. We test our hypotheses on a longitudinal data set from an online knowledge community at a global IT consulting company.

2.2. Theory and Hypotheses

2.2.1. Risks of Sharing Knowledge in an Online Knowledge Community

In an online knowledge community, geographically dispersed employees can ask for help from a broad audience without the barriers of time and space. In principle, any employee can offer solutions to questions posed in the online system. We investigate whether knowledge sharing spans geographical and social hierarchical boundaries in practice.

Enhancing one's reputation is a major individual motivation to contribute to online knowledge communities (Constant et al. 1996, Lerner and Tirole 2002, Wasko and Faraj 2005): individuals want to establish a positive reputation and gain approval from others. This motivation is particularly crucial in career-related online knowledge communities, because the quality of community participation has direct consequences for participants' careers: high quality contributions can lead to better job prospects (Lakhani and Wolf 2005, Oreg and Nov 2008). Lakhani and von Hippel (2003) found that participants, due to such motivation, strive to provide intelligent answers in open source software development projects.

Despite the desire to provide intelligent answers, it can be challenging to offer an adequate, not to mention an intelligent, answer in an online knowledge community. Further, postings to the online community are visible for all participants and are retained in the system, which increases the costs of a poor response. A major challenge to knowledge sharing in online communities is the "mutual knowledge problem." Mutual knowledge is knowledge that participants share in common and know that they share (Krauss and Fussell 1990). Increased mutual knowledge leads to higher levels of common ground, which is vital for effective communication (Cramton 2001). When a knowledge seeker and a knowledge provider share common ground, the provider can craft his or her response according to what the seeker does and does not know and thereby increase the response's effectiveness. By contrast, when a knowledge seeker and provider do not share common ground, misunderstandings and problems are likely to occur.

Establishing common ground in online communities is challenging for several reasons. First, participants in an online knowledge community might not know each other. Without knowing a knowledge seeker's background knowledge, a knowledge provider might misinterpret a question and/or inadvertently use terminologies that a knowledge seeker cannot understand. Second, in face-to-face settings, we use interactive back-channel feedback (e.g., head nods, facial expressions) to gauge the effectiveness of knowledge exchanges. However, this interactive feedback is limited in online knowledge communities. Hence, it is hard to obtain cues about whether the offered solution is understood and appropriate. Third, because of the asynchronous nature of communication (i.e., the time lapse between postings), it is difficult to resolve confusion that arises in the process of knowledge exchange.

Several mechanisms have been identified as facilitators of common ground: direct shared experience, interactional experience, and category membership (Krauss and Fussell 1990, Cramton 2001). Direct shared experience and experience interacting with individuals provide information about what individuals do and do not know and, thus, increase common ground. Membership in various social categories such as professional status or location also provides information about what individuals are likely to know. In addition, membership in social categories can increase liking because individuals like others who are similar to them (Byrne 1971).

2.2.2. Interpersonal Similarity as a Facilitator of Online Knowledge Sharing

We propose that individuals prefer to share knowledge with similar others because interpersonal similarity increases common ground and comfort and reduces the risks and challenges of knowledge sharing. Although much information about communicating partners (e.g., voice tone, appearance) evaporates online (Sproull and Kiesler 1986), some information remains. For instance, categorical information (e.g., office location, hierarchical status) is publicly disclosed as user profile information. Another type of available information is expertise information. In an online knowledge community, participants interact by posting questions and answers. By observing others' knowledge-sharing behavior, participants can learn who is good at what, because providing an answer signals that the answer poster has expertise on the question area.

We examine the interpersonal similarities of a knowledge-sharing dyad (i.e., a pair of a knowledge provider and a seeker) in two dimensions, based on information available in the online community: (1) categorical similarity and (2) expertise similarity. By categorical similarity, we mean co-membership in categories. The two categories that are part of participants' user profiles in our empirical context are geographic location and hierarchical status. Membership in these categories provides diffuse information about what individuals are likely to know (Bunderson 2003). On the other hand, expertise similarity provides fine-grained information about what individuals know. By observing participants' postings to the system, users can glean information about whether other participants have expertise on diverse technical areas that are discussed in the knowledge community.

Interpersonal similarity is a fundamental source of interpersonal connection in face-to-face settings (McPherson et al. 2001). Physical as well as social distance regulates opportunities to meet: people are more likely to interact with nearby colleagues than with distant ones, because people who are in close proximity have more chances to meet, which in turn increases the probability of forming ties. Moreover, maintaining relationships with people who are farther away requires considerable effort (Zipf 1949). Previous studies have found that factors such as the arrangement of streets (Hampton and Wellman 2000), dormitory halls (Festinger et al. 1950), gender (Moore 1990), age (Fisher 1982), school grade (Shrum et al. 1988), and status (Cohen and Zhou 1991) affect the likelihood of forming relationships. Scholars have found that in organizational settings, interpersonal similarity among managers leads to linkages across intra- as well as inter-organizational boundaries (Brass 1995, Brass et al. 2004), which enhance information flow across organizational units (Borgatti and Cross 2003, Hansen 1999, Makela et al. 2007, Markus 2001, Monteiro et al. 2008, O'Leary 2014).

Yet, once connected to the Internet, everyone has equal chances to interact (Hollingshead and McGrath 1995, Friedman 2006). For example, no extra effort should be required to provide an answer to a knowledge seeker who is located 5,000 miles away compared to one who is collocated. Hence, it has been argued that geographical distance does not constrain interactions in online knowledge communities. In addition, social dissimilarity, such as differences in status, has been suggested to carry minimal weight in electronic communication (Siegel et al. 1986, Sproull and Kiesler 1986). In face-to-face communication, people signal their social category through various cues (e.g., appearance, eye contact, voice tone, gestures, and office size), which lead to more distinctions across social attributes and less interaction across socially dissimilar people. In online knowledge communities, however, most social cues are missing or significantly reduced because communicating partners are not physically co-present. Such reduced social context cues have been found to result in more equalized participation in experimental studies comparing face-to-face and computer-mediated communication in small groups (Kiesler et al. 1984, Sproull and Kiesler 1986, 1991).

Despite the unique contextual characteristics of online communities (e.g., elimination of physical distance, reduced social context cues) that could encourage more equalized interactions across physical as well as social boundaries, the reduced contextual cues may, paradoxically, amplify individuals' preferences toward others who belong to the same category. According to Reicher (1984), individuals tend to associate more strongly with a group identity when they are de-individuated (i.e., separated) because they cannot observe differences among in-group members. Following Reicher (1984)'s line of reasoning, Lea and Spears (1991) argued and empirically demonstrated that individuals tend to overattribute a few available social cues to judge others when using electronic communication. For example, in an experimental study using computer-mediated communication, Lea and Spears (1991) found that individuals perceived a stranger more positively if the stranger belonged to the same social category.

Similarly, we propose that participants in an online knowledge community are likely to identify themselves with similar others based on information available through the community.

Further, participants are expected to be more willing to share knowledge with similar others than with dissimilar others for several reasons. First of all, people feel more comfortable and safer when they interact with similar other people because interpersonal similarity reduces uncertainty (Brewer and Brown 1998, Hewstone et al. 2002). For example, Cross and Sproull (2004) found that people feel more comfortable asking dumb questions to explore unknown areas when their communicating partners are similar than when they are dissimilar.

Second, people are more attracted to similar others (Byrne 1971) and feel more positive about individuals who are like themselves (Byrne et al. 1966, Byrne and Nelson 1964, Singh and Tan 1992, Tajfel et al. 1971). Information is perceived as more credible if it comes from similar people (O'Reily 1983). The tendency to be more attracted to similar people than to dissimilar people leads to more interaction among similar individuals. Scholars, who often refer to this tendency as the homophily principle, have found that a tie is more likely to exist between similar people than between dissimilar people. Empirical support of the homophily principle in face-to-face interactions has been documented across various categories, such as status, gender, race, kinship, nationality, and religion (e.g., Allen 1977, Brass 1995, Cohen and Zhou 1991, Ibarra 1993, Galegher et al. 1990, McPherson et al. 2001, McPherson and Smith-Lovin 1987, Pelled 1996).

Third, interpersonal similarity increases the common ground between knowledge-sharing dyads. Having common ground with knowledge-sharing partners improves the odds of providing useful knowledge, because individuals with common ground are more likely to have a shared language and skills and an understanding of what others know. The higher the level of common ground, the more successful (understood as intended) communication is (Clark and Marshall 2002). Scholars have documented, primarily in laboratory studies, that communicating partners who have common ground are more successful in sharing knowledge than those who lack common ground (e.g., Fussell and Krauss 1992, Horton and Keysar 1996).

Knowledge-sharing partners in an online knowledge community typically do not have common ground. Intermittent interaction with the same partner and limited availability of interactive feedback (i.e., nodding, expressions such as “uhuh”) impede the establishment of common ground. Also, any confusion due to the lack of common ground can remain unresolved in online knowledge communities because of the asynchronous nature of communication. This is why it can be challenging to offer an *adequate*, not to mention an *intelligent*, answer in an online knowledge community. With shared language and skills, a knowledge provider can use prior mutual knowledge as a building block of new knowledge. Without shared language and skills, a knowledge source might use terminology that the corresponding seeker cannot understand, based on incorrect assumptions about a seeker’s background knowledge. Therefore, we propose that participants in an online knowledge community respond to others who are similar to them because mutual knowledge helps them to provide useful information: consequently, their efforts to enhance their reputations are not wasted.

In sum, we suggest that knowledge flow in online knowledge communities occurs within categorical and expertise boundaries because individuals are more willing to share knowledge with similar others in order to reduce the risks and challenges associated with sharing knowledge. Categorical and expertise similarity increase common ground and enable more effective knowledge sharing. Therefore, we expect to replicate the effect of similarity observed in face-to-face groups online and hypothesize that:

Hypothesis 1. Greater categorical and expertise similarity with a knowledge seeker increases the likelihood of knowledge sharing.

2.2.3. Shifting from Categorical to Expertise Similarity with Experience

Although we expect that participants generally prefer to share knowledge with similar others as proposed in Hypothesis 1, we additionally expect that the relative strength of different types of similarities changes as a function of a knowledge provider's experience in the online knowledge community. Specifically, we expect that the effect of categorical similarity on knowledge sharing becomes weaker and that of expertise similarity becomes stronger as a knowledge provider has more experience exchanging knowledge in the online community. A knowledge provider gains more experience in the system each time he or she posts or answers a question.

As participants become more experienced in seeking as well as sharing knowledge in an online knowledge community, they feel more comfortable in offering knowledge because of increased familiarity with the system and other participants. Previous research on the interpersonal effects of computer-mediated communication (CMC) reveals that, with extended interactions, CMC can promote positive relationships among participants. The increased familiarity, comfort, and liking toward other participants established through extended interaction in an online knowledge community are expected to substitute for participants' need to look for comfort from others who belong to similar categories.

Experience also enables participants to acquire information about other participants in an online knowledge forum. Information differs in its accessibility and interpretability: some information is easier to obtain and interpret than others. Research on diversity has proposed that individual characteristics reside at different levels: some are on the surface and others are at a deeper level (Harrison et al. 1998, Harrison et al. 2002, Jackson et al. 1995, Jehn et al. 1999, Williams and O'Reilly 1998). Characteristics of individuals that are readily detectable, generally immutable, and easily measurable are considered to be surface-level features. Categorical characteristics such as gender, location, and hierarchical status are examples of such individual characteristics. On the other hand, certain individual attributes take more effort to discern. Characteristics that are more "subject to construal and mutable (Jackson et al. 1995; 217)" than categorical attributes are considered to be deep- level features. Common examples of deeper level characteristics are a person's attitudes, values, knowledge, and skills. Because such deep information can only be learned little by little, by observing behavioral patterns, extended interactions are required to acquire deep-level information.

In online knowledge communities, expertise information is harder to obtain and interpret than categorical information and thus is an example of a deep-level feature. While only one click is enough to acquire user categorical information such as location, participants need to observe others' numerous knowledge-sharing interactions in order to gather information about who has expertise on which topics. The behavior of offering a solution signals that the solution provider possesses expertise in the area. The measure of others' expertise formed through the traces of actions (i.e., the number of answers in a topic area) has been found to reflect the actual expertise of participants reasonably well. In a large-scale public online knowledge-exchange community (Java Forum), Zhang et al. (2007) found that expertise rankings based on the number of answers were highly correlated with the expertise ratings by human raters ($\rho = 0.77$). Because participants gather information about others' expertise by observing which knowledge others are sharing, the experience of exchanging knowledge in an online knowledge community offers participants more opportunities to acquire knowledge about who knows what. Hence, more experience engaging in an online knowledge community increases information about others' expertise.

Expertise information is also not as easily measurable as categorical information. Categorical information usually has a straightforward meaning: the meanings of a job title and office location are apparent to most participants (i.e., most participants' interpretations of the information would be the same). In contrast, expertise information is more nuanced and latent and requires more effort to interpret. Participants need to observe patterns of knowledge-sharing behavior for an extended period of time in order to obtain information about the topics on which a given participant has expertise. By contrast, interpretation of categorical information is more straightforward and does not require experience.

Previous studies on diversity in face-to-face settings have shown that people tend to form their initial perceptions of others based on surface-level features and adjust those perceptions as they obtain deeper information about others through extended interaction (Amir 1969, Byrne and Wong 1962, Harrison et al. 1998, Stangor et al. 1992). For example, stereotypes formed by categorical information are replaced by more accurate knowledge as individuals obtain deeper-level information through extended interaction (Amir, 1969). Scholars have proposed that this shift from surface-level to deeper-level happens because deeper-level information tends to be more informative than surface-level information, which makes categorical information less relevant as people collect deep-level information (Jehn et al. 1999).

Similarly, in an online knowledge community, we expect that as participants have more opportunities to acquire information about others' expertise, participants place more weight on expertise similarity and less weight on categorical similarity when choosing with whom to share knowledge, because experience serves as a substitute for categorical similarity. The familiarity and common ground provided by experience takes the place of the common ground provided by categorical similarity. By contrast, experience complements expertise similarity: the more experience participants have, the more they learn about each other's expertise and identify those with whom they share common ground. The opportunity to collect expertise information increases as a knowledge provider has more experience exchanging knowledge through an online knowledge community. The key difference is that interpretation of categorical information is more straightforward and does not require experience, but experience does increase participants' knowledge of others' expertise. The familiarity and common ground provided by experience substitute for the familiarity and common ground provided by membership in the same categories. In contrast, experience enables participants to learn about others' expertise and identify those with whom they have mutual knowledge. Thus, experience complements expertise similarity. We therefore hypothesize that:

Hypothesis 2a. The effect of categorical similarity on knowledge sharing becomes weaker as a knowledge provider's experience with the online community increases.

Hypothesis 2b. The effect of expertise similarity on knowledge sharing becomes stronger as a knowledge provider's experience with the online community increases.

Furthermore, the opportunities for a knowledge provider to discern a specific seeker's expertise increase as the seeker offers more knowledge in the online forum. In order to capture the different level of opportunities offered by knowledge seekers' past knowledge-sharing activities, we incorporate seeker visibility in our model. Seeker visibility captures the amount of answers a seeker has provided to the online knowledge community. These answers provide information about the areas in which a seeker has expertise – about what he or she knows. Hence, we predict that the proposed interaction effects in hypotheses 2a and 2b become stronger as a corresponding seeker's visibility increases. That is, we predict that as seeker visibility and provider experience increase, the effect of categorical similarity decreases and the effect of expertise similarity increases. In other words, for a given dyad of a knowledge provider (*i*) and a knowledge seeker (*j*), we expect that *i* will shift his or her attention from categorical to expertise similarity more quickly if j was more active. The more active a knowledge seeker is (higher $Seeker\ Visibility_{jk}$), the more opportunities a knowledge provider has to learn about the seeker's expertise. Therefore, we hypothesize that:

> *Hypothesis 3a. The negative interaction effect of categorical similarity and a knowledge provider's experience becomes stronger as a corresponding seeker is more visible in the online knowledge community.*
>
> *Hypothesis 3b. The positive interaction effect of expertise similarity and a knowledge provider's experience becomes stronger as a corresponding seeker is more visible in the online knowledge community.*

2.3. Methods

2.3.1. The Research Context

We tested our hypotheses in an online knowledge-exchanging community in a Fortune 500 company. The organization is a leading information-technology (IT) consulting company that helps its customers to plan, develop, deploy, and manage their IT systems. Project teams of the organization reside on clients' sites across the globe and conduct tasks such as data management/migration and software development. The majority of project team members are software developers. Because project teams perform similar tasks, knowledge created in one project is useful to other project teams.

Recognizing that knowledge can be utilized across projects, the organization set up an internal online knowledge community to facilitate knowledge sharing among its dispersed employees. First launched in April 2006, the online knowledge community supports text-based, peer-to-peer, and asynchronous knowledge exchanges. The community is focused mainly on assisting the organization's technical employees (e.g., software developers) in sharing technical knowledge such as Java, .Net, and database queries. The online knowledge community is part of a project management application that the organization has developed to manage its software development projects. It is located at the top-right corner of the project management application. Because all technical employees submit their code and report task progress using the application, postings to the online knowledge community are visible to its target users.

The online knowledge community offers basic functions. It is structured as a single community without any embedded sub forums. Employees post technology-related questions and indicate the topics of their messages by choosing one of the 114 pre-specified topic tags (e.g., .Net, Java, Database, Cobol) from a drop-box menu. Nicknames are used as a main identifier of a message poster, and some personal information such as real name, job title, and office location is available in a public user profile, which can be accessed by clicking usernames. There are no personalizable features, such as automated mailing, or priority settings, which specify which messages should be displayed on top. Nor are there any functions that distinguish the quality of messages (e.g., thumbs up/down) or message posters (e.g., badges). Participation in the online knowledge community is completely voluntary, and the organization does not track how much employees contribute, nor does it provide extrinsic rewards for contributions. The participants do not take leadership roles in maintaining or managing the online knowledge community.

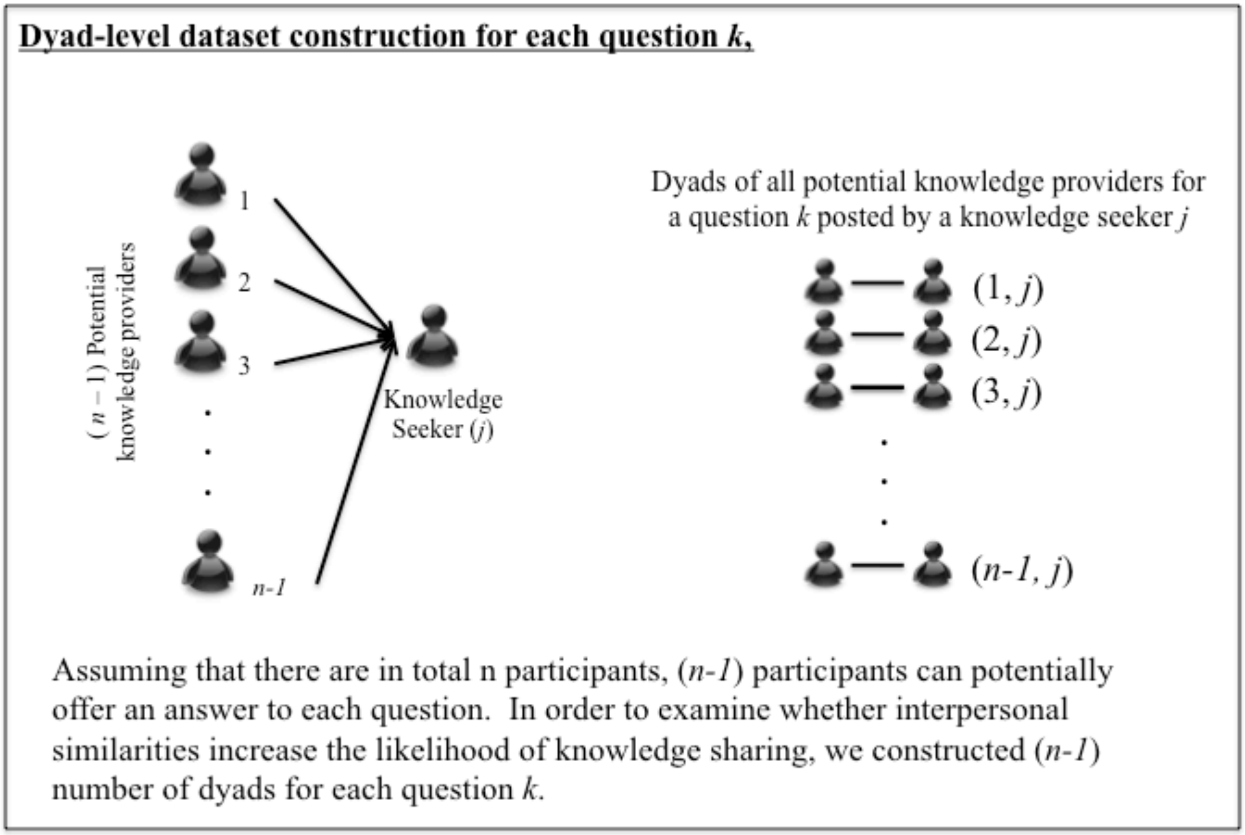
2.3.2. Data

The organization provided us two sets of data: public user profile information and knowledge-sharing data. The data cover the period from April 2006 to August 2007. Each knowledge-sharing instance provided information about who posted the message, when it was posted, and which topic the message addressed. Using a unique thread identifier, we identified who shared knowledge with whom on which question. Similar to that of other online communities, posting activity was not evenly distributed in the online knowledge community. Some active participants posted many messages, whereas others posted only a few. The distribution of number of posts per individuals was skewed to the right with an average of 95.95 total posts and a median of 62 posts during our observation window. About half of the total messages were posted by approximately 20% of participants. However, whereas previous studies find that many participants adopt roles as either askers or answerers (Wesler et al. 2007, Zhang et al. 2007), the majority of participants in our research context posted both answers and questions in a balanced way. Finally, we found no spamming or trolling behavior in the online knowledge community.

In order to examine whether a knowledge provider is more likely to offer an answer to a knowledge seeker if both parties have the same category attributes or possess similar expertise, we analyze the data at a dyad level. A dyad is a social unit of two actors. Here, a dyad consists of two individuals, a knowledge provider and a knowledge seeker. In order to systematically estimate the effects of dyadic similarities on the likelihood of knowledge sharing, we constructed a dataset that included dyads that shared knowledge as well as dyads that did not. If we only considered the dyads that shared knowledge, our model would produce biased estimators due to sample-selection issues (Greene 2011).

We used the following procedures to construct our longitudinal dataset. First, we excluded a portion of knowledge-sharing data in the early period due to little activity. As we observed from our data, significant activity started during the 16th week. Therefore, we constructed dyad-level data using the knowledge-sharing activities that took place after 16 weeks so that we have meaningful expertise profiles of participants during our observation window. The activity in the first 16 weeks of data is still incorporated to construct the expertise profiles of each participant. Second, for each question posted, we created all potential dyads in order to observe which dyads shared knowledge and which did not. For example, where *n* denotes the number of participants during our observation window, (*n-1*) dyads were constructed for a question *k* posted by employee *j* because all the other participants except for the question poster *j* could potentially provide an answer to employee *j*'s question *k*. Figure 1 illustrates how we constructed our dyad-level dataset. The unit of analysis of this study is person (*i*)-person (*j*)-question (*k*) level where *i* refers to a knowledge provider, *j* refers to a knowledge seeker and *k* refers to the question *j* posted. During our observation window (46 weeks), we identified 586 participants and 25,412 questions. Consequently, 14,866,020 dyads were constructed and used as our observations (25,412 questions x 585 potential knowledge providers for each question). Among them, 42,047 dyads actually shared knowledge.

FIGURE 2. 1. Dyad-level dataset construction

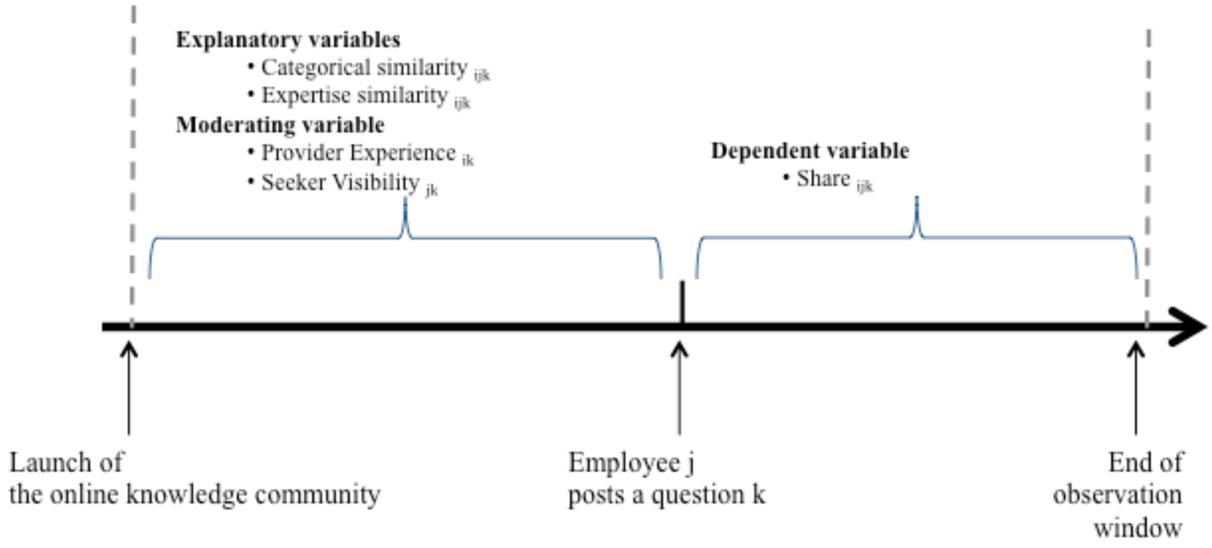

Dyad-level dataset construction for each question k,
(n – 1) Potential knowledge providers
1
2
3
n-1
Knowledge Seeker (j)
Dyads of all potential knowledge providers for a question k posted by a knowledge seeker j
(1, j)
(2, j)
(3, j)
(n-1, j)

Assuming that there are in total n participants, (n-1) participants can potentially offer an answer to each question. In order to examine whether interpersonal similarities increase the likelihood of knowledge sharing, we constructed (n-1) number of dyads for each question k.

2.3.3. Measures

Our research examines whether knowledge providers take into account dyadic similarities with a question poster when they decide whether to provide an answer in an online knowledge community. For any information to be considered, it should first be available. So, we constructed our dyadic similarity variables using the information that is available to participants through the online knowledge community. Our explanatory variables were created based on the data available *before* a focal question was posted, while our outcome variable was constructed using the data available only *after* a focal question was posted. Constructing independent and dependent variables using data in different time periods addresses the reverse causality issue. Figure 2 illustrates how we separated time periods to construct our dataset.

FIGURE 2. 2. Timeline of dataset construction



Knowledge sharing The dependent variable of this study is whether participant *i* shared knowledge with participant *j* for question *k* ($Share_{ijk}$). We consider that knowledge was shared if participant *i* posted at least one reply to participant *j*'s question *k*. In order to verify that reply messages were actually answers to the corresponding questions, we randomly sampled 200 questions and examined whether the replies to the question were actual answers. There were 367 answers to the 200 questions. Among them, only a very small number (≈ 1%) of replies were follow-up questions. We further found that those participants who posted follow-up questions also posted actual answers to the original questions later. Because we operationalized $Share_{ijk}$

as 1 when a participant *i* posted one or more reply messages to a participant *j*'s question k and 0 otherwise, $Share_{ijk}$ captures the actual knowledge-sharing actions.

Categorical similarity In the online knowledge community, a message poster's real name, job title, and office location are publicly disclosed in a user profile, which can be easily obtained through one click. According to interviews with the organization's employees, most participants in the online knowledge community seek out user profile information of question posters before posting a reply message. The main reason for checking is curiosity, but interviewees also emphasized that checking question posters' user profile information helped them to gather additional information about the context of the question. We measured the categorical similarity of a knowledge-sharing dyad using the information of job title and office location.

Location similarity ($LocationSim_{ij}$) is coded as 1 if a knowledge provider (i)'s office is located in the same city as his corresponding seeker (*j*)'s office, and 0 otherwise. *Status similarity* ($StatusSim_{ij}$) was measured based on job titles. Job titles of the organization contain information about employees' hierarchical positions but no information on functional areas. For example, chief officer, manager, and junior associate are common job titles in the organization. Because the organization's hierarchy is structured approximately in three levels, we categorize participants into three groups (high, middle, and low) based on their job titles. Participants who have a job title that contains words such as chief, manager, senior, principal, chairman, or director were categorized as high hierarchical-level participants. Job titles with assistant, junior, or trainee were categorized as low hierarchical-level. Remaining job titles were classified as middle hierarchical-level. $StatusSim_{ij}$ is coded in the same manner as $LocationSim_{ij}$: 1 if a knowledge provider (i) and a seeker (*j*) are in the same hierarchical level, 0 otherwise.

Participants in the online knowledge community were located in 73 cities in five countries. Among the dyads, about 35% were co-located. About 63% of the participants belonged to the middle, 22% to the high, and 15% to the low hierarchical level.

Expertise similarity Our measure of expertise similarity is based on publicly expressed expertise. Expertise information can be learned by observing others' knowledge-sharing behavior for an extended period of time, because the number of answers to a certain topic area signals that the knowledge provider has expertise in the area (Zhang et al. 2007). In order to measure $ExpertiseSim_{ijk}$, we first tracked a time-evolving vector representing each participant's expertise. The online knowledge community had 114 pre-specified topic areas (e.g., Java, Database, .Net). When a participant posts a question, he or she selects one topic area that best matches the question topic. Corresponding answers to the question automatically carry the same topic area. We counted the number of answers a participant offered for each topic area and constructed the expertise distribution of that participant across 114 topic areas. A participant's expertise profile is captured by a 114-dimensional vector, $E_{ik} = \left(E_{ik}^{1} \cdots E_{ik}^{S}\right)$, where E_{ik}^{S} represents the number of answers in topic area S by participant *i* up to the point question k is posted. A simple illustration of a participant's expertise profile where there are five topic areas would be (2 0 0 0 19), which indicates that the participant provided two answers on the first topic area, 19 answers on the fifth topic area, and none on the second, third, and fourth topic areas.

$ExpertiseSim_{ijk}$ of a dyad is then measured by computing the cosine similarity of two participants' expertise profiles. Cosine similarity determines whether the expertise profile vectors of two participants are pointing to the same direction by calculating the cosine angle of the two vectors. Previous studies on innovation adopted the cosine similarity measure to calculate the proximity of firms in patent class distribution (Jaffe 1986, Sampson 2007). The formula for the expertise similarity of a knowledge provider *(i)* and a seeker *(j)* for question k is defined as:

$$ExpertiseSim_{ijk} = \frac{E_{ik}E'_{jk}}{\sqrt{(E_{ik}E'_{ik})(E_{jk}E'_{jk})}}$$

where $i \neq j$. The resulting similarity ranges from 0 to 1. The greater the overlap in expertise areas, the more expertise a knowledge provider and a seeker share: the value of 0 means the two participants do not have any common expertise and 1 means the two have exactly the same expertise sets.

***Experienc*e** According to our theory, our measure of experience should capture a knowledge provider (i)'s cumulative number of opportunities to observe other participants' knowledge-sharing activities. In order to take into account the unique context of an online knowledge community, we measured participants' generic engagement in the online knowledge community to construct $Provider\ Experience_{ik}$. In an online knowledge community, all interactions leave traces, which enable participants to collect information even about others with whom they do not directly interact. As a result, we counted a knowledge provider (i)'s total number of answer-posting and question-posting activities until question k is posted to measure $Provider\ Experience_{ik}$.

In order to more closely capture a knowledge provider (i)'s opportunity to observe a specific knowledge seeker (*j*)'s knowledge-sharing activities, we included another moderating variable, $Seeker\ Visibility_{jk}$. $Seeker\ Visibility_{jk}$ captures how actively a knowledge seeker (*j*) was sharing knowledge in the online knowledge community until the knowledge seeker posted question k. It is measured by counting the number of answers provided by a knowledge seeker j until the knowledge seeker posts question k. We log-transformed both $Provider\ Experience_{ik}$ and $Seeker\ Visibility_{jk}$ to adjust for skewed distributions.

Control variables The decision to share knowledge can also be driven by other factors. In order to tease out the effects of dyadic similarity on a dyad's likelihood of knowledge sharing, we incorporated a number of control variables.

First, a participant's decision to share knowledge may be driven by the desire to reciprocate. The findings about how reciprocity affects knowledge sharing are somewhat mixed. While many studies have shown that reciprocity is an important motivation to exchange knowledge (e.g., Cross and Sproull 2004, Fulk et al. 2004, Faraj and Johnson 2011), some studies did not find an effect of reciprocity on knowledge sharing (e.g., Wasko and Faraj 2005). To control for any potential effect from reciprocity, we included $Reciprocity_{ijk}$. $Reciprocity_{\mathrm{i}jk}$ is a dummy variable and coded as 1 if a knowledge seeker j has provided at least one answer to a knowledge provider *i* up to the point question *k* is posted by j.

Second, the decision to share knowledge is also likely to be influenced by whether a knowledge provider has expertise on the question topic area. In order to control for a knowledge provider's expertise, we incorporated $Ability_{ik}$ into our model. $Ability_{ik}$ captures whether a knowledge provider (i) has expertise in the topic area of question *k* when the question *k* is posted. Ability is measured by counting the number of answers a knowledge provider *i* offered in the topic area of question *k* up to the point question *k* is posted. Similar to our measure of expertise similarity, our measure of ability is constructed based on participants' publicly expressed expertise.

Third, a participant's tendency to reciprocate may interact with a knowledge provider's ability. For example, in a technical online knowledge community, a novice may not be able to reciprocate to an expert who has helped him or her previously, if the expert asks an advanced question. In order to control for the interaction effect between reciprocity and ability, we incorporated $Reciprocity_{ijk} \times Ability_{ik}$[2].

Fourth, our dataset spans 46 weeks. To control for any unobserved effects caused by time differences (e.g., system improvement), our model includes the variable $Time_k$, which is the number of weeks since the online knowledge community was launched when question k is posted.

Fifth, a knowledge provider might be less willing to share knowledge if a question requires complicated or detailed answers. In order to control for the effect of the expected efforts required for each question, we incorporated $ExpectedEffort_k$ into our model. $ExpectedEffort_k$ is measured by calculating the average lengths of all answers to question k. We log-transformed the variable to adjust for skewed distribution.

Sixth, a novice employee may be less willing to share knowledge because he or she is not confident about expertise. To capture this effect, we included a dummy variable, $Novice_i$. Participants who have job titles that include the word "trainee" were classified as novices. Lastly, we included random effects for each participant in a dyad to control for any unobserved heterogeneity of participants, such as activeness in the online knowledge community.

[2] We thank the senior editor for suggesting that we test the interaction between reciprocity and ability.

Additionally, we incorporated random effects for each dyad to control for any unobserved dyad-specific heterogeneity.

2.3.4. Model Specification

Our observations are not independent. Because one participant can be a member of multiple dyads, error terms will be correlated across observations. If we do not account for the dependence, our model will produce artificially reduced standard errors. Among many solutions to the problem (Simpson 2001), we include random effects for each participant in each dyad: a_i for a knowledge provider and b_j for a knowledge seeker. We conducted Hausman test to ensure that a random effects model is unbiased and efficient compared to a fixed effects model (Hausman 1978). The test statistic (Chi-sq(22)=15.98, p > 0.1) indicates that the random effects model is unbiased. The random effects allow the dependent variable, $Share_{ijk}$, to vary randomly around the mean of a dyad across participants within dyads (Greene 2011). Given that our dependent variable is dichotomous, we used logistic regression to estimate our parameters where the sender (a_i)- and receiver (b_i)-specific effects of the same individual are allowed to be correlated with each other as:

$$\binom{a_i}{b_i} \sim MVN\left(\begin{bmatrix}\widehat{a}\\ \widehat{b}\end{bmatrix}, \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}\right).$$

The unobserved dyad-specific homophily is captured by using a dyad-specific unobserved random effect, d_{ij}, where $d_{ij} \sim MVN(0, \sigma_d^2)$. Furthermore, we assume that the dyad-specific unobserved effects are symmetric, i.e., $\mathrm{d_{ij}}=d_{ji}$.

A final estimation issue concerns the computational challenge. With 14,866,020 observations[3], we confronted computational and resource challenges to estimating parameters. This is a challenge often encountered in large-scale dyad-level studies of networks (e.g., Braun and Bonfrer 2009, Lu et al. 2013). To alleviate the challenge, we adopted a Bayesian inference procedure for estimation. Because the Bayesian approach does not require maximization algorithms, the estimation procedure is more efficient than the frequentist approach (Cameron and Trivedi 2005, Gelman and Hill 2007). Contrary to the frequentist approach, where the main interest is to determine the point estimate of true parameter value θ_0, the interest of the Bayesian approach lies in producing the entire distribution of the parameters of interest given the data and a prior. We estimated the parameters by using a Markov Chain Monte Carlo (MCMC) procedure, using a Gibbs sampler and the Metropolis-Hastings algorithm. To reduce autocorrelation between draws of the Metropolis-Hastings algorithm and to improve mixing of the MCMC, we used an adaptive Metropolis adjusted Langevin algorithm (Atchade 2006). In the hierarchical Bayes procedure, the first 100,000 observations were used as burn-in, and the last 50,000 were used to calculate the conditional posterior distributions. We ensured that the parameters converged by comparing the within-to between-variance for each parameter estimated across multiple chains (Gelman et al. 2003). The initial 100,000 observations were chosen for burn-in as our tests indicated that the parameters converged in approximately 50,000 observations. Models were estimated using Matlab, and the full estimation procedure is provided in the Appendix.

[3] Among 14,866,020 dyads, 42,047 dyads actually shared knowledge.

2.4. Results

The descriptive statistics for all variables are reported in Table 1. Table 2 presents the results of logistic regression analysis with dyad random effects. Effects are introduced across columns to demonstrate the stability of the results. Model 1 includes only control variables. Estimates for categorical similarity, $LocationSim_{ijk}$ and $StatusSim_{ijk}$, are added in models 2 and 3, respectively. $ExpertiseSim_{ijk}$ is further incorporated in model 4. Model 5 incorporates $Provider\ Experience_{ik}$ and $Seeker\ Visibility_{jk}$ and their interactions. Two-way interaction terms between similarity variables and the moderating variables were added in models 6 and 7. Lastly, we added three-way interactions among similarity variables, $Provider\ Experience_{ik}$ and $Seeker\ Visibility_{jk}$ in model 8. Because the direction and the significance of all coefficients are stable across models, we use the complete specification (model 8) to discuss the results.

TABLE 2. 1. Descriptive statistics and correlations

		Mean	s.d.	1	2	3	4	5	6	7	8	9	10	11
1	Share$_{ijk}$	0.00	0.06	1.00	0.05	0.00	0.03	0.02	0.08	0.02	0.03	-0.01	0.00	-0.01
2	StatusSim$_{ij}$	0.80	0.29		1.00	0.01	0.04	0.01	0.04	0.07	0.00	-0.01	0.00	-0.02
3	LocationSim$_{ij}$	0.35	0.48			1.00	0.03	0.02	0.03	0.11	0.01	-0.01	-0.01	0.02
4	ExpertiseSim$_{ijk}$	0.13	0.20				1.00	0.10	0.26	0.09	0.02	0.00	0.01	0.06
5	Provider Experience$_{ik}$	6.18	1.89					1.00	0.02	0.08	0.02	0.01	-0.01	-0.19
6	Seeker Visibility$_{jk}$	4.11	1.37						1.00	0.16	0.04	0.02	0.01	0.09
7	Reciprocity$_{ijk}$	0.02	0.23							1.00	0.09	0.01	0.01	0.17
8	Ability$_{ik}$	0.02	0.27								1.00	0.04	0.01	0.07
9	Novice$_{i}$	0.14	0.12									1.00	0.00	0.00
10	Expected Effort$_{k}$	5.89	0.93										1.00	0.01
11	Time$_{k}$	13.58	4.24											1.00

$N = 14,866,020^{+}$

Provider Experience$_{ik}$, Seeker Visibility$_{jk}$, and ExpectedEffort$_{k}$ are log-transformed

$^{+}$ *Among 14,866,020 potential knowledge-sharing dyads, 42,047 dyads actually shared knowledge.*

TABLE 2. 2. Predicting dyad's likelihood of knowledge sharing: panel logistic regression with dyad random effects

	1	2	3	4	5	6	7	8
Categorical similarity								
LocationSim$_{ij}$		0.94 ***	0.93 ***	0.89 ***	0.89 ***	1.29 ***	1.35 ***	1.83 ***
StatusSim$_{ij}$			0.59 ***	0.58 ***	0.59 ***	0.72 ***	0.79 ***	1.07 ***
Expertise similarity								
ExpertiseSim$_{ijk}$				0.34 ***	0.32 ***	0.51 ***	0.45 ***	0.42 ***
Moderating variables								
Provider Experience$_{ik}$					0.75 ***	0.89 ***	0.97 ***	1.39 ***
Seeker Visibility$_{jk}$					0.88 ***	0.82 ***	0.80 ***	0.92 ***
Interactions with a knowledge provider's experience								
LocationSim$_{ij}$ x Provider Experience$_{ik}$						-1.18 ***	-1.07 ***	-0.97 ***
StatusSim$_{ij}$ x Provider Experience$_{ik}$						-0.59 ***	-0.54 ***	-0.42 ***
ExpertiseSim$_{ijk}$ x Provider Experience$_{ik}$						1.21 ***	1.54 ***	1.75 ***
Interactions with a knowledge seeker's visibility								
LocationSim$_{ij}$ x Seeker Visibility$_{jk}$							-0.24 *	-0.15 *
StatusSim$_{ij}$ x Seeker Visibility$_{jk}$							-0.10	-0.02
ExpertiseSim$_{ijk}$ x Seeker Visibility$_{jk}$							0.29 **	0.31 **
Interactions between provider experience and seeker visibility								
Provider Experience$_{ik}$ x Seeker Visibility$_{jk}$					0.41 **	0.39 **	0.47 **	0.51 **
Three-way Interactions								
LocationSim$_{ij}$ x Provider Experience$_{ik}$ x Seeker Visibility$_{jk}$								-0.81 ***
StatusSim$_{ij}$ x Provider Experience$_{ik}$ x Seeker Visibility$_{jk}$								-0.18 **
ExpertiseSim$_{ijk}$ x Provider Experience$_{ik}$ x Seeker Visibility$_{jk}$								1.09 ***
Controls								
Reciprocity$_{ijk}$	0.41 ***	0.45 ***	0.43 ***	0.42 ***	0.23	0.24	0.22	0.21
Reciprocity$_{ijk}$ x Ability$_{ik}$					0.75 ***	0.74 ***	0.74 ***	0.74 ***
Ability$_{ik}$	0.64 ***	0.61 ***	0.62 ***	0.61 ***	0.41 ***	0.40 ***	0.39 ***	0.41 ***
Novice$_{i}$	-1.03 **	-0.99 ***	-1.01 ***	-0.98 ***	-0.95 ***	-0.94 ***	-0.96 ***	-0.93 ***
Expected Effort$_{k}$	0.22 *	0.23 *	0.20 *	0.18	0.15	0.12	0.10	0.11
Time$_{k}$	1.37 ***	1.34 ***	1.33 ***	1.35 ***	1.34 ***	1.33 ***	1.35 ***	1.34 ***
Constant	-8.44 ***	-8.11 ***	-8.21 ***	-8.19 ***	-8.09 ***	-8.01	-7.98 ***	-7.97 ***
Inverse Mills ratio	1.03 ***	0.99 ***	0.98 ***	0.98 ***	0.99 ***	0.95 ***	0.93 ***	0.94 ***
Random effects								
σ_a^2	0.47 ***	0.44 ***	0.45 ***	0.45 ***	0.44 ***	0.43 ***	0.44 ***	0.42 ***
σ_b^2	0.74 ***	0.75 ***	0.75 ***	0.74 ***	0.72 ***	0.70 ***	0.67 ***	0.67 ***
σ_{ab}	0.79 ***	0.79 ***	0.76 ***	0.76 ***	0.71 ***	0.69 ***	0.62 ***	0.62 ***
σ_d^2	0.79 ***	0.71 ***	0.64 ***	0.59 ***	0.57 ***	0.55 ***	0.54 ***	0.53 ***

Dependent variable: Share$_{ijk}$
N = 14,866,020^{+}
Provider Experience$_{ik}$, Seeker Visibility$_{jk}$, and Expected Length$_{k}$ are log-transformed
***** The 99% credible interval does not include zero.*
*** The 95% credible interval does not include zero.*
** The 90% credible interval does not include zero.*
$^{+}$ *Among 14,866,020 potential knowledge-sharing dyads, 42,047 dyads actually shared knowledge.*

The results support the first hypothesis, which predicted positive effects of all dyadic similarities on the likelihood of knowledge sharing. The coefficients of all three similarities ($LocationSim_{ij}$, $StatusSim_{ij}$, and $ExpertiseSim_{ijk}$) are positive and statistically significant, indicating that participants tend to share knowledge with similar others. Our results show that the magnitude of the main effect of expertise similarity is smaller than that of both types of categorical similarities (i.e., location and status similarities). This finding suggests that when an individual is new to an online knowledge community, he or she is more likely to offer an answer to others who are in the same location or in the same status than to others who have similar expertise.

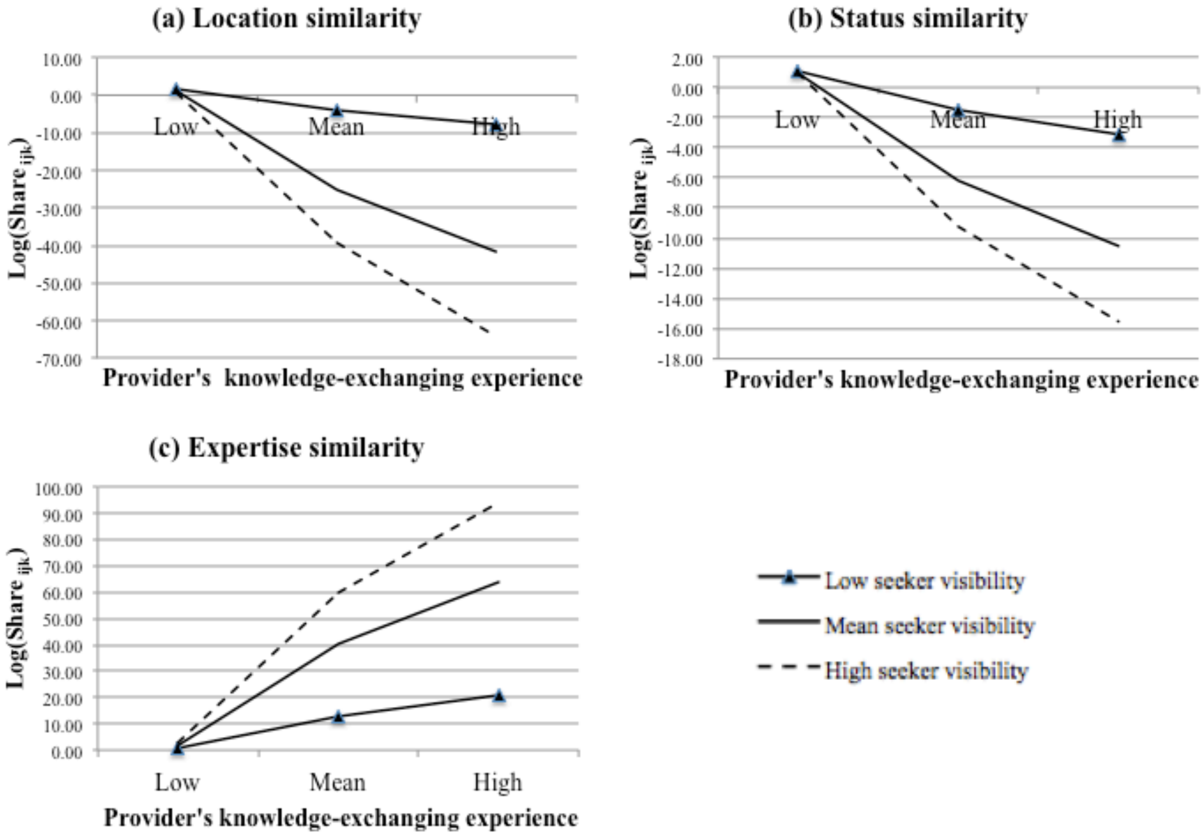
Our second set of hypotheses is also supported. The interaction terms between categorical similarities and a knowledge provider's experience ($Provider\ Experience_{ik}$) are negative and statistically significant, whereas the interaction between expertise similarity and $Provider\ Experience_{ik}$ is positive and significant. The results indicated that the effects of categorical similarities on knowledge sharing weaken and the effect of expertise similarity strengthens as provider experience increases.

Furthermore, the results from the three-way interaction terms show that the decreasing effect of categorical similarity and the increasing effect of expertise similarity on knowledge sharing are intensified when a question poster (*j*) is more active (higher $Seeker\ Visibility_{jk}$) in the online knowledge community. A higher level of $Seeker\ Visibility_{jk}$ means that there were more chances for a knowledge provider (i) to learn the corresponding knowledge seeker (*j*)'s expertise information. Thus, our results support Hypothesis 3.

Because our interaction terms are statistically significant, the main effects of similarities should be interpreted in conjunction with their interaction effects with $Provider\ Experience_{ik}$ and $Seeker\ Visibility_{jk}$ (Jaccard and Turrisi 2003). The magnitude of dyadic similarity effects on the likelihood of knowledge sharing changes according to the level of our moderating variables. Our results demonstrate that the total effect of categorical similarities on the likelihood of knowledge sharing *decreases* whereas the total effect of expertise similarity *increases* as knowledge providers accumulate more experience observing others in the online knowledge community. Compared to a knowledge provider who is new to the online knowledge community, an experienced one pays less attention to categorical similarity factors but more attention to expertise similarity factors when deciding whether to offer an answer to a knowledge seeker. The tendency of a knowledge provider (i) to shift attention from categorical to expertise similarities with experience is intensified when the provider's counterparts (j) are more actively sharing knowledge in the online knowledge community.

Plots of the relationship between dyadic similarities and the likelihood of knowledge sharing, given different levels of a knowledge provider's experience ($Provider\ Experience_{ik}$) and a knowledge seeker's visibility ($Seeker\ Visibility_{jk}$), are displayed in Figure 3. For easier interpretation, we used the log-linear form of a logistic model where the log likelihood of knowledge sharing is in a linear relationship with the estimated parameters. The downward slopes of plot (a) and (b) suggest that knowledge providers put less weight on categorical similarities as they have more chances to observe others in the online knowledge community. The steeper slopes for higher level of $Seeker\ Visibility_{jk}$ indicate that this tendency becomes stronger as a knowledge provider has more chances to observe his or her corresponding knowledge seeker. The interaction diagram for expertise similarity (plot (c)) illustrates the opposite tendency.

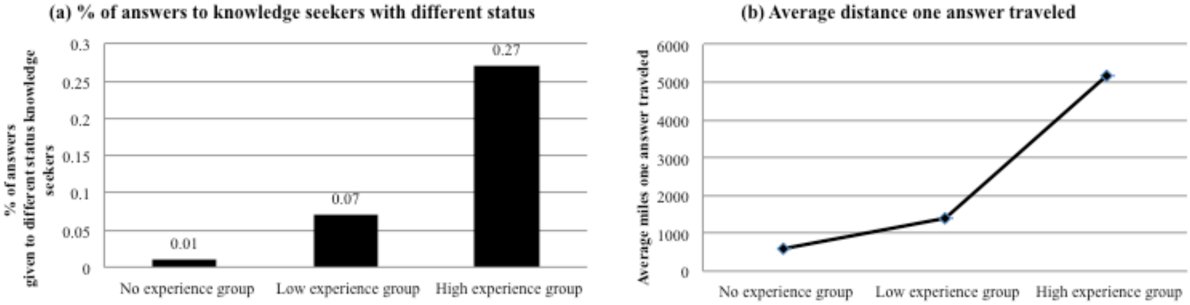
FIGURE 2. 3. Interaction effects of dyadic similarities on the likelihood of knowledge sharing



To better understand the meaning of the interaction effects, we split all the dyads that have actually shared knowledge based on the experience level of a knowledge provider and compared the degree of interpersonal similarities in each split. We found starkly different patterns in each split in terms of how far knowledge providers reached out to offer an answer (Figure 4). Compared to knowledge providers who are new to the online knowledge community (no previous experience group), high-experienced knowledge providers crossed status boundaries about 27 times more on average as depicted in Figure 4(a). More interestingly, answers from high-experience providers traveled about nine times farther away, on average, than answers from newcomers. An average answer by a knowledge provider with no previous experience travelled 581 miles (the approximate distance between Chicago and Atlanta) whereas an average answer by a high-experience knowledge provider traveled 5,182 miles (the approximate distance between San Jose in California, U.S.A. and London in United Kingdom) (Figure 4(b)).

The skewed activity distribution of participants in the online knowledge community implies variance in when a participant crosses boundaries to share knowledge. For instance, an active participant will shift his or her attention from categorical to expertise information faster because she or he can gather expertise information faster than can less active participants. On the contrary, less active participants will need significant time to acquire expertise information. As a result, it will take longer, or it may never happen, for less active participants to share knowledge with dissimilar others in an online knowledge community.

FIGURE 2. 4. Weakening categorical boundaries with higher provider experience



Experience of a knowledge provider in a dyad	Split rule	% of dyads
No experience	Experience = 0	5.1%
Low experience	Lower than median	44.9%
High experience	Higher than median	50%

Although the results are consistent with our predictions, there could be alternative explanations for our findings. We took several steps to investigate alternative interpretations. First, in order to tease out the effects of similarity after controlling for other potential drivers of knowledge sharing, we incorporated a number of control variables: $Reciprocity_{ijk}$, $Ability_{ik}$, $Novice_i$, $ExpectedEffort_k$, and $Time_k$. In addition, we included random effects for each participant in a dyad as well as for each dyad in order to control for any unobserved heterogeneity. The results remained consistent after incorporating the control variables. Interestingly, $Reciprocity_{ijk}$ is not statistically significant, whereas its interaction with $Ability_{ik}$ is significant. The results indicate that an individual's propensity to reciprocate increases as the individual is capable of helping. This could be the reason for previous studies' inconsistent results concerning the effect of reciprocity on knowledge sharing. The positive coefficient of $Ability_{ik}$ indicates that as individuals are more knowledgeable on a certain topic, they are more likely to provide answers. The negative coefficient of $Novice_i$ shows that if an individual is new to the company and the field (e.g., a trainee), an individual is less likely to provide an answer. The coefficient of $ExpectedEffort_k$ is not significant, which indicates that individuals do not take into account the expected efforts required to answer a question when deciding which question to answer.

Second, multicollinearity could be a concern when incorporating interaction terms into the analysis. If substantial collinearity problems are present, parameter estimates can be unstable to very small changes in the data (Greene 2011). To examine the influence of multicollinearity, we employed a number of robustness checks. We computed variance inflation factors (VIF) for each of the independent variables and interactions among them. VIF quantifies the severity of multicollinearity by measuring how much of the variance of coefficients is increased because of correlation among the explanatory variables (Marquaridt 1970). VIF of 10 and above is suggested to indicate a multicollinearity problem (O'Brien 2007). The test revealed no diagnostic problem in our case: all of our VIF statistics are below 5 with mean VIF across all variables of 1.92. We also introduced effects hierarchically in Table 2 in order to see whether the size or signs of effects changes significantly. The directions and magnitude of effects are consistent across columns, demonstrating the robustness of results. Another factor that gives more credibility to our result is our large dataset, which produces a high level of statistical power. It has been found that big data can overcome even extremely high correlations among variables (Mason and Perreault 1991).

Third, the observed effects could be attributed to different levels of opportunities rather than selection. For example, if the participant population of the online knowledge community became more diverse in terms of categorical attributes, a knowledge provider would have more opportunities to share knowledge with dissimilar others in terms of categorical attributes. Likewise, if the number of topic areas shrank over time, a knowledge provider would have less opportunity to share diverse types of knowledge, leading to positive interaction effects between $Provider\ Experience_{ik}$ and $ExpertiseSim_{ijk}$. This explanation, however, is not viable in our research context. As a robustness check, we tracked the diversity of participants' demographic attributes and topic areas in the online knowledge community over the observation window. We employed an entropy-based index, Shannon's diversity index, which measures the richness and evenness of different categories in the population. We found that the diversity of the community participants' office location, hierarchical status, and the number of topic areas was stable during our study period.

Fourth, we controlled for potential selection bias. Not all employees of the organization participated in the online knowledge community, which might lead to concerns about selection bias: i.e., there is something systematically different about the employees who participated in the community compared to those who did not. Such a selection bias could limit the generalizability of our results. To correct for this potential self-selection bias, we used a common econometric technique: the Heckman's correction, a two-step statistical approach (Heckman 1979). In the first stage, we estimated a probit model capturing the employees' decisions to participate in the online knowledge community. To estimate the employees' decisions to participate, we used data on employee participation in an enterprise blogging platform. Six months before launching the online knowledge community, the organization had launched an enterprise blogging platform where employees could blog about topics of their interest. These blogs were accessible to all employees of the organization and the blog users include all 586 participants of the online knowledge community. So, we used the participation of employees in the enterprise blogging, such as the number of work-related blogs posted by the employee and the number of work-related blogs read by the employee, as instruments, along with their demographic characteristics (i.e., status, location, and novice status) and time, to predict whether they would participate in the online community. The number of work-related blogs posted by an employee indicates the employee's willingness to share knowledge with others and should be a good predictor of his or her participation in the online knowledge community. The number of work-related posts read by an employee indicates his or her desire to learn more from peers and should also be a good predictor of the employee's participation in the online knowledge community. We expect the likelihood of an employee participating in the online knowledge community to increase with the number of work-related posts written and read by him or her. The results indicate that both number of work blogs read and posted predicted an employee’s subsequent participation in the online community. After estimating the first stage selection model using the instruments, in the second stage, we added the inverse Mills ratio computed from the first stage model to our main analysis model (i.e., the dyadic knowledge-sharing model). By including the inverse Mills ratio, we correct the estimates from our main analysis model for potential self-selection bias, because the determinants of employees’ decisions to participate in the online knowledge community are already accounted for. The instruments were not significantly correlated with any of the other covariates in the dyadic knowledge-sharing model.

2.5. Discussion and Conclusions

Information technology opens opportunities for organizations to make the most of their existing knowledge base. The great opportunity comes from the *theoretical* potential of information technology to bridge boundaries and to connect otherwise unconnected people. This great potential, however, might be an illusion, because information technology, paradoxically, also has the potential to fragment communities. For instance, people might use advanced filtering capability to locate the most like-minded others and interact only with them, causing boundaries to intensify rather than vanish.

We proposed and empirically tested a theory about how individuals choose knowledge-sharing partners in an online knowledge community. We argued that participants tend to choose similar others who are within boundaries, because interpersonal similarity increases attraction as well as common ground, which reduces the risks and challenges associated with sharing knowledge. Likewise, we argued that as participants learned about others’ expertise by observing their knowledge-sharing behaviors for extended periods of time, participants would favor those who were similar in expertise over those who were similar in categorical attributes. Consistent with our predictions, our results show that participants share knowledge more frequently with similar others than with dissimilar others. In addition, as participants gain more experience exchanging knowledge in the online knowledge community, they pay less attention to categorical similarities (location and status) but more attention to expertise similarity.

Our study advances knowledge in several areas. First, it contributes to literature about online communities. Our dyad-level approach offers complementary insights to previous studies that examined participation motivations at the individual level (e.g., Bagozzi and Dholakia 2006, Ma and Agarwal 2007, Ren et al. 2007, von Hippel and von Krogh 2003, Wasko and Faraj 2005) and at the community level (Butler 2001, Faraj and Johnson 2011). Specifically, our study offers insight about whether dyad-specific characteristics impact individuals' willingness to offer an answer in online knowledge communities and about the consequences of those dyad-level drivers for macro-level knowledge-exchange patterns. It has been argued that the Internet will make the world "flat" (Friedman 2006), meaning that geography will not constrain interpersonal, online connections. Our findings, however, suggest that geography still shapes tie formation among individuals online, although its effect weakens as individuals learn about others' expertise information. Our finding is in line with Van Alstyne and Brynjolfsson (2005)'s theoretical argument that information technology can shift boundaries from geography to interests. Also, our findings are consistent with the view that online communities are fluid objects that constantly morph their boundaries (Faraj et al. 2011). To the best of our knowledge, ours is the first study to provide empirical evidence of the fluid nature of online communities.

Our work has implications for computer-mediated communication literature. Homophily breeds connection in face-to-face settings. For instance, even in the circumstances where categorical characteristics are not relevant cues to build relationships, individuals tend to make personal networks that are homogeneous in those aspects (McPherson et al. 2001). On the contrary, in computer-mediated communication, a dominant prediction has been the "cues-filtered-out" theory (Kiesler et al. 1984, Lancaster 1978, Linstone and Turoff 1976, Sproull and Kiesler 1986), which proposes that individuals filter out irrelevant cues and form relationships based on relevant cues in computer-mediated communication. Our findings, however, suggest that individuals do not completely disregard cues about categorical information that are less relevant than cues about expertise in online knowledge-exchanging communities. Individuals initially rely on categorical cues. However, as individuals gain experience in the online system and learn more about others' expertise, they rely less on categorical cues and more on expertise cues. As a result, we find that it takes significant time for categorical cues about location and status to be "filtered out" in online knowledge exchanges.

The study also contributes to literature on diversity. While previous studies (Harrison et al. 1998, 2002) examine how "direct" interaction experience changes the relative strength of surface- and deep-level diversity, we examine how "indirect" interaction experience changes their relative strengths. The setting for the research of Harrison and colleagues is small work groups where everyone has ample opportunity to get to know each other through direct interaction. Our setting is an online knowledge community, where a large number of unacquainted people interact through computer-mediated communication. We show that even in the setting that allows very limited and lean interaction, people discern deeper-level information and rely more on it with increasing experience. We also extend previous studies by examining different outcome variables at different units of analysis. While Harrison and his colleagues (1998, 2002) examined how different types of diversity influence team-level outcomes, such as group social integration and team task performance, we investigate how different types of interpersonal similarity affect an individual's willingness to share knowledge in an online knowledge community at a dyad-level.

Our results also advance understanding of how online knowledge communities influence knowledge sharing in organizations. In spite of the widespread use of online knowledge communities in organizations, most studies on online knowledge communities have been conducted in extra-organizational settings (e.g., Yahoo! Answers). Based on the Internet's potential to connect distant and dissimilar others, it was anticipated that employees can obtain novel solutions through online knowledge communities because dissimilar individuals are more likely to have non-redundant information than similar individuals (Burt 1992, Granovetter 1983). Thus, online knowledge communities would be able to help employees balance the classic trade-off between exploration (use of new solutions) and exploitation (use of known solutions) (March 1991). Connections to others outside of geographic or social boundaries can benefit employees because they gain access to new information and solutions not available locally. Our results reveal a nuanced view of the role of homophily online. While the effect of homophily based on categorical membership decreases with experience, the effect of homophily based on expertise similarity increases. Research is needed to determine whether this pattern of online knowledge exchange fosters exploitative or exploratory learning.

Previous studies in the knowledge-sharing literature have shown the positive influence of interpersonal similarities on knowledge sharing. However, little attention has been given to how individuals differentiate types of similarities (e.g., categorical similarity, geographic proximity) as a basis of association nor to whether the impact of similarities on knowledge sharing dynamically changes. Based on our analysis of longitudinal data from an online knowledge sharing community, the current study was able to estimate how individuals shift their attention from categorical similarity to expertise similarity when they decide with whom to share knowledge.

In recent years, organizations have made significant investments in community-based knowledge sharing platforms to enable their employees to share knowledge across geographic as well as social boundaries. Our findings can be used to develop an online knowledge-sharing system that improves knowledge flow across boundaries. First, organizations might consider what incentives would be effective to entice employees to participate continuously in online knowledge communities. This will ensure that participants gain more opportunities to observe others, which in turn will help them pick up expertise information from others. Second, organizations could make expertise information more salient so that less time and effort is required for participants to acquire it. For example, organizations can use the traces of participants' knowledge-sharing activities to induce expertise areas of participants and display the information for reference. While extended interaction would still be required to learn nuanced expertise information, providing cues about it could jump-start the process.

This study has a few limitations that suggest directions for future research. First, the current study focused only on the behaviors of active participants. However, a large number of online community participants are free riders who may read but not contribute (Nonnecke and Preece 2000). Exploring how inactive participants use community resources could provide valuable insights about how knowledge is distributed through an online knowledge community. Second, we could not incorporate previous offline relationships between a knowledge provider and a knowledge seeker in our analyses. Although we have a dyad-specific random effect in our model, which accounts for such prior offline relationships, it would be interesting to examine the effects of such relationships. Lastly, the current study does not examine the linguistic content of the conversations. Previous studies suggested that a message is more likely to be answered if it adopts rhetorical strategies, such as an introduction in the message text or a request for specific information (Burke et al. 2007). Also, better articulation of questions (e.g., being on topic and using less complicated language) can also increase the likelihood of a reply (Arguello et al. 2006, Wasko and Faraj 2005). Incorporating text-level characteristics into the analysis model would advance our understanding of additional factors that drive knowledge sharing in online knowledge communities.

In conclusion, this study applied theories of experience-based learning, common ground, and diversity to investigate knowledge sharing in online knowledge communities. We found that individuals prefer to share knowledge with similar others. Moreover, we found that as individuals learn about others' expertise information by observing their extended interactions, they increasingly favor those who are similar in expertise and decreasingly favor those who are similar in categorical attributes as knowledge-sharing partners. Consequently, categorical boundaries weaken, whereas boundaries around expertise strengthen as participants accumulate more experience in an organizational online knowledge community.

Chapter 3.

Jack of All, Mater of Some: Information Network and Innovation[4]

3.1. Introduction

With advancement of information technology, many organizations are now inviting their customers to suggest new product ideas through a practice called 'innovation crowdsourcing.' Crowdsourcing is a practice of outsourcing a function once performed by employees to an undefined large network of people through an open call (Howe 2008). Through an innovation crowdsourcing community[5], users can propose new product or service ideas directly to a company. In contrast to traditional user involvement mechanisms such as focus groups, which usually occur in a small group setting, innovation crowdsourcing involves a much larger number of users. For instance, in our empirical setting, approximately 3,000 individuals contributed over 8,000 ideas during 2.5 years. Another distinction from prior user innovation is that most recent innovation crowdsourcing communities are centered in customer product markets such as personal computers, beverages, and mobile services. While prior user innovation has been mostly active in industrial goods market setting (e.g., medical/scientific instruments), Dell,

[4] This essay is a joint work with Param Vir Singh and Linda Argote.

[5] The focus of this study is an ongoing innovation crowdsourcing community where individuals can propose any new product ideas at any time. Although this type of innovation crowdsourcing communities are among the very popular open innovation approach formal organizations take to complement their internal, "closed", innovation approach, there are other types of open innovation approaches. For example, an open tournament announce a challenge and crowds compete for better solutions for the challenge (e.g., TopCoder).

Starbucks, BMW, Nike, and BestBuy, for instance, are among the pioneering companies that have established innovation crowdsourcing initiatives.

Because users retain market needs information, they can be crucial actors in new product development efforts. Users have been actual developers of many commercially successful new products. For example, 82% of all commercialized scientific instruments are developed by actual users (von Hippel 1976, 1988). Also, it has been found that lead users play an important role in developing new products (Henkel and von Hippel 2005, Laursen and Salter 2006, Rosenberg 1982, Urban and von Hippel 1988, von Hippel 1976). First introduced by von Hippel (1986), the term *lead user* describes those users who face product needs that are still unknown to the public and who can benefit greatly a solution to these needs is developed. For instance, lead users of flashlights would be policemen and home inspectors who require bright and efficient lights as part of their day-to-day business. Given that users play an important role in new product development efforts, the recent phenomenon of innovation crowdsourcing has great promise to add value to companies.

Despite the growing popularity and potential value of innovation crowdsourcing, we have relatively limited understanding on the antecedents of successful new product ideation outcomes in innovation crowdsourcing communities. A handful of previous studies have documented characteristics of successful individuals at innovation crowdsourcing communities. It has been found that individuals who have proposed multiple ideas (Bayus 2013), who are in occupations outside of the innovation area (Jeppesen and Frederiksen 2006), and who are at technically and socially marginal positions (Jeppesen and Lakhani 2010) are most likely to generate high quality ideas. Also, individuals who are positioned at the core of user communities but also span boundaries to other communities are found to generate higher quality ideas than their counterparts who are do not span boundaries (Dahlander and Frederiksen 2012). In addition to examining characteristics of successful individuals at innovation crowdsourcing communities, research has begun to examine participation dynamics in these communities. Huang et al. (2014) found that low-ability participants tend to become inactive after they learn that they lack ability to develop high-quality ideas.

The current study extends prior work by investigating the *complementarity* of individuals' activities between innovation crowdsourcing and customer support crowdsourcing communities. Most innovation crowdsourcing communities also have a separate customer support crowdsourcing community within the same platform. At a customer support crowdsourcing community, users can help each other to try to figure out solutions to the problems that they are facing. In this study, we examine how individuals' helping activities at a customer support community influence their new product ideation outcomes at an innovation crowdsourcing community. We focus on helping activities because a helping tie between a helper and a help-seeker at a customer support community represents flow of important information. One, an act of helping in a customer support community indicates that a helper is knowledgeable on a problem area. In other words, it represents a flow of solution information from a helper to a help-seeker. Two, by helping others to find solutions, helpers can learn about the needs and problems of other customers. That is, it represents a flow of needs information from a help-seeker to a helper.

Because solution and needs information are crucial inputs to innovation processes (Baker and Freeland 1972, Baker et al. 1967), we analyze individuals' information networks based on their helping activities in a customer support community. Then, we examine how the structure of individuals' information networks, in terms of their breadth and depth, affects their new product ideation outcomes. Here, breadth refers to the scope of information one has and depth refers to the level of understanding one has in a domain area. Deep knowledge implies that the individual has expertise, so we will use the two terms interchangeably.

We propose that individuals who have engaged in helping others on broader problem areas (generalists) are more likely to create original ideas in an innovation crowdsourcing community because diverse information is available for them to recombine in novel ways. When there are diverse ingredients, it is more likely that the resulting recombination is new. In addition, we propose that the quality of ideas that generalists create is likely to vary: some of their ideas would be extremely high quality while other ideas would be low quality. Whereas diverse information increases the upside potential of idea quality by improving novelty aspect of ideas, we expect that it may also increase the downside potential of idea quality because individuals are less likely to effectively utilize information as the number of pieces of information grows (Martin and Mitchell 1998). As a result, we finally propose that, only those generalists who possess expert knowledge in at least one domain area are likely to outperform non-generalists. In other words, without any expertise, we do not expect that generalists are superior to non-generalists in their ability to create ideas that are later implemented by a company. At an innovation crowdsourcing community hosted by a British telecommunication company, we empirically test our theory by evaluating 8,110 new product ideation “projects” in a real world setting.

3.2. Theory and Hypotheses

3.2.1. Broad Information As a Source of New Ideas

"Creativity is just connecting things. … Unfortunately, a lot of people in our industry haven't had very diverse experience."

Steve Jobs in Wired, February 1995

Rather than breaking out of the old to produce the new, it has been proposed that creative thinking builds on existing knowledge (Hayes 1989, Kulkarni and Simon 1988, Weisberg 1999). Scholars have suggested that innovative solutions are generated through the process of recombining or rearranging pre-existing knowledge in new ways (Dahl and Moreau 2002, Fleming 2001, Gilfillan 1935, Nelson and Winter 1982, Schumpeter 1939, Usher 1954). That is, an established solution in one field might be used as a novel way to solve a problem in another field. In the cement industry, for instance, a common solution to keep cement in a liquid form during mass cement pours is to keep mixing cement in a revolving drum. Borrowing from this established solution, an engineer in the cement industry proposed a novel way to separate Exxon Valdez oil spill from the ocean. The engineer proposed that attaching a revolving tool to a long pole and inserting it into the oil recovery barges would keep the oil from freezing and enable it to be easily pumped from the barge (Lakhani 2009). Similarly, new product development teams at IDEO purposely invite experts from unrelated fields to encourage a fresh perspective on a problem (Hargadon and Sutton 1997). In short, the value of broad information is in furnishing raw ingredients to make new recombinations. In this paper, we use the terms new, novel, and original interchangeably.

Besides offering rich ingredients for novel recombination, broader information may also help individuals to be less constrained by pre-existing solutions. Creativity researchers have suggested that individuals tend to be fixated by prior solutions, which limits their ability to generate novel ideas (Jansson and Smith 1991, Smith et al. 1993). This tendency is called cognitive fixation. In a laboratory experiment where participants were asked to generate design solutions (e.g., spill-proof coffee cup), Jansson and Smith (1991) found that individuals tend to generate less novel solutions when they were provided with a design example than when they were not. Participants who were given an example tend to include features of the example even when the example contained flawed features. Given this tendency of cognitive fixation, scholars have explored ways to mitigate it. Sharing diverse ideas with others helps individuals to avoid cognitive fixation (Bayrus 2013, Dugosh et al. 2000, Smith 2003).

Building on this research, we expect that generalists, who have provided solutions on diverse problem areas in a customer support community, are likely to create more original ideas in an innovation crowdsourcing community than those individuals who have provided solutions on a narrower span of problem areas (non-generalists). In the course of helping others on diverse domain areas, generalists become aware of more diverse needs of others. These diverse customer needs are expected to increase the potential for generalists to create more original ideas because the combinatorial possibilities exponentially increase as the number of available resources increase. Recombining previously untried components is generally conducive to novel outcomes. Moreover, as generalists are exposed to more diverse solutions and needs than non-generalists, we expect that generalists will be mentally less constrained by existing solutions, which is also expected to increase their probability to create original ideas. Therefore, we hypothesize that:

Hypothesis 1. Generalists create original ideas than non-generalists.

3.2.2. Broad Information As a Source of Quality Variance

Even if generalists are more likely to create original ideas, we propose that they might not create high quality, "innovative" ideas. Although we tend to relate the term "innovation" only to radical or breakthrough ideas, it actually incorporates both incremental and radical ideas (Ettlie et al. 1984). In fact, new product ideas submitted to most innovation crowdsourcing communities are incremental improvements to existing product or service lines.

Original ideas are not equivalent to innovative ideas. An idea should satisfy multiple criteria in order to be innovative (Schumpeter 1939, Sternberg and Lubart 1995). Schumpeter (1939)'s definition of innovation is a commercially applicable invention. In order to be an invention, an idea should be original. In order to be commercially applicable, an original idea should also be useful and economically feasible. For example, a new product idea that no one is interested in purchasing will be of no value to a company. Similarly, a new product idea that everyone may love but is technically impossible to produce or economically unfeasible to profit from will also be of little value to a company. In short, in order to be considered an innovative new product idea, an idea should satisfy three criteria: originality, usefulness, and feasibility. Therefore, although we expect broad information to increase idea originality, its effects on idea usefulness and feasibility also must be considered in order to understand the effect of broad information on innovation.

We propose that broad information increases the variance of idea quality an individual creates. As proposed in H1, broader information contributes to novelty of ideas by offering more building blocks of information. By helping generalists to create more novel ideas, we expect that information breadth will increase the upside potential of idea quality created by individuals. Yet, at the same time, we propose that information breadth may also increase the downside potential of idea quality because individuals are less likely to process information correctly as the number of information components increases (Martin and Mitchell 1988). It is true that broad information may provide potentially infinite number of combinatorial possibilities, which is likely to produce more novel, unknown recombinations. However, the quality of unknown recombination is likely to be highly uncertain (Fleming 2001, Taylor and Greve 2006). Because humans have limited cognitive capacity to perfectly filter out low quality recombination, we expect that the quality of ideas created by generalists is likely to vary highly: some of their ideas are very high quality and others are low quality. Therefore, we hypothesize that:

> *Hypothesis 2. Generalists create ideas that are more variable in their quality than non-generalists.*

3.2.3. Contingent Effect of Information Breadth on Idea Quality

We propose that generalists can create higher quality ideas than non-generalists only when they are able to effectively utilize their diverse sets of information. We further suggest that it is deep knowledge that enhances individuals' ability to utilize diverse information. Through in-depth understanding of a knowledge domain, individuals develop more abstract representations of the knowledge, which enables them to pay more attention to relevant, and structural features across different domains (Glaser 1989, Newell and Simon 1972). In contrast, individuals lacking deep knowledge (non-experts) tend to pay more attention to less relevant and superficial features, which often distract them from attending to more meaningful connections across diverse concepts.

Deep knowledge has been found to influence an individual's ability to use analogy effectively (Casakin 2004, Collins and Burstein 1989, Vosniadou 1989). Analogical reasoning, an important psychological cognitive process, involves comparing two components in different domains in order to infer similarities and import solutions from one to the other (Dunbar 1995). When attempting to use analogical reasoning, individuals who have deep knowledge in any of the two knowledge domains were more likely to establish successful analogies (Novick 1988) because the deep knowledge enabled them to map relevant features across different knowledge domains. Conversely, novices tended to retrieve irrelevant, surface features, which deter them from making successful analogies.

Because an idea that incorporates diverse knowledge components is more likely to have unanticipated flaws (Fleming 2001), in order to create a good idea, generalists should be able filter out impractical ideas. Deep knowledge helps generalists to identify constraints of potential solutions. A thorough understanding of a problem is a crucial part of problem solving (Simon 1981). Experts have been found to dedicate substantially greater effort than novices to elaborate their understanding of a problem and add ill-defined and implicit constraints to the problem (Eckert et al. 1999). For example, in an experiment study in architecture, Casakin (2004) observed that experts added more constraints to the design problem, which decreased the total number of alternative design solutions experts produced but increased the overall quality of their ideas. On the other hand, novices generated a greater number of solutions than experts did but most of the solutions proposed by novices were impractical. Moreover, in a chess game setting, Chase and Simon (1973) found that novice players were more likely than experts to conduct an exhaustive search in order to find an appropriate solution while master players tended to successfully limit their solutions to those that would lead to promising outcomes.

Because deep knowledge helps generalists to utilize their knowledge more effectively by enabling them to make more meaningful connections across diverse information and to identify constraints of potential solutions, we expect that only those generalists who have deep knowledge are capable of creating higher quality ideas than non-generalists. Therefore, we hypothesize that:

Hypothesis 3. Generalists' ability to create higher quality ideas than non-generalists is contingent on whether generalists possess deep knowledge.

3.3. Empirical Method

3.3.1. Research Context: Crowdsourcing Communities

Our empirical context is crowdsourcing communities hosted by a British telecommunication company. The company sells SIM (Subscriber Identification Module) cards, mobile phones, data plans, and bundled telecommunication services. Unlike conventional mobile telephone operators, the company crowdsources many of its operations such as customer support, new product development, marketing, and sales from its own customers. The company's two major crowdsourcing communities are customer support and innovation communities. In return for their contributions, customers are rewarded with points, which can be cashed out, credited against their monthly bills, or be donated to charitable institutions.[6] With free registration,

[6] The top earning customer earned over £13,000, who used it to pay his college tuition.

anyone can join and participate in the communities but one must be a customer of the company with an active SIM card in order to be compensated.

The customer support crowdsourcing community replaced a call center that is used by most other mobile operators. Besides confidential billing questions, which are handled by approximately thirty dedicated customer support employees, the company lets its customers to handle all other customer issues. Registered members can post their problems related to the company's products or services to the customer support community and members of the community can provide solutions to those problems. Majority of the problems are technical issues such as problems regarding exporting/importing contacts, swapping SIM cards, and how to use SIM card abroad. According to the company, the average response time for questions is three minutes, day or night, with 95% of questions being answered within an hour.

Another major crowdsourcing community of this company is an innovation crowdsourcing community. The innovation community allows its members to propose any new product or service ideas at any time. Unlike open innovation tournament settings where solutions to a specific challenge are crowd-sourced during a limited time frame[7], the innovation community does not set a specific problem to be solved. Rather, the company aims to obtain diverse ideas from its user crowd. Ideas submitted to the innovation community move along the following path. Once an idea is submitted, its status remains as *proposed* until it receives at least 20 customer votes. If a proposed idea receives 20 or more votes, it is eligible to be reviewed by the company's management team. After an idea has obtained 20 or more votes but before it is reviewed by the company, the idea is labeled as *under consideration*. The company screens all submitted ideas to filter out redundant ideas. If very similar or the same idea has been offered

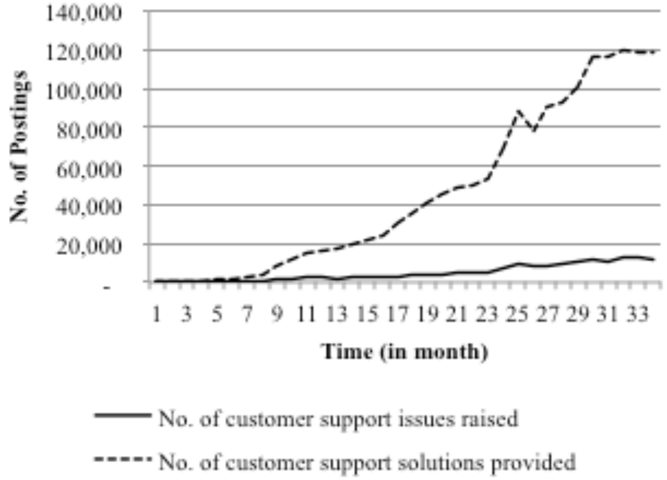
[7] Companies such as TopCoder and Innocentive hold contests to crowdsource the best solutions to specific challenges.

earlier, the idea is marked as *redirected* and linked back to the original idea. If selected for implementation by the company, the idea's status is set to be either *coming soon* or *implemented*. Until the ideas are officially launched, they are marked as *coming soon*. Once formally launched, ideas are marked as *implemented*. If not selected, the idea is marked as *not for us*. Each month, the company recognizes those individuals who created the implemented ideas by publicly announcing them in its blog. The company also financially rewards those individuals by providing points that can be either cashed out or credited against their bills.

3.3.2. Data

Our data span approximately three years (35 months) starting from the company's inception on November 2009 to October 2012. During our empirical window, total 177,560 customer support issues were posted to the customer support community and 1,692,391 solutions were offered to those issues: on average, 9 solutions were offered to each issue. We found that only a very small portion of the posted customer issues (less than 0.0003% of posted question) were not solved. Figure 1 illustrates the total posting activities in the customer support community during our empirical period. It shows that the number of postings increased significantly for the first 30 months after the launch and stabilized around 12,000 issues (118,000 solutions) per month.

FIGURE 3. 1. Customer support community activities over time

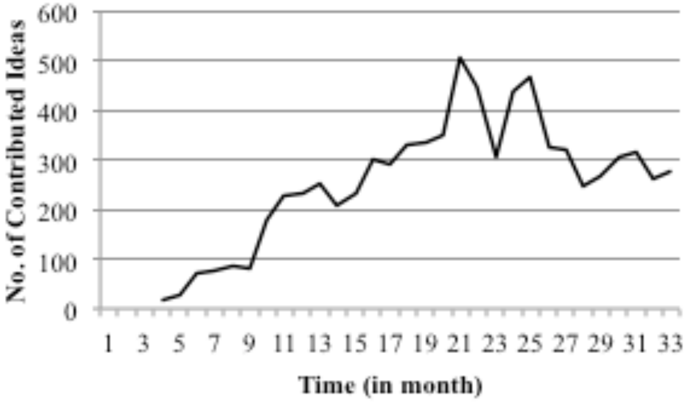


While the customer support community was launched at the company's inception, the innovation community was launched two months after the company's inception. The number of new product idea submissions was steadily increased over time and starting from 26th month, the number of idea contributions stabilized around 300 ideas per month. During our empirical period, a total of 8,396 ideas were submitted. Table 1 illustrates sample implemented and unimplemented ideas submitted to the innovation community. Similar to other innovation crowdsourcing communities, most new product ideas submitted to the innovation community are incremental improvements to existing product or service lines. Also, from our observation, higher kudos (customer votes) does not seem to guarantee idea implementation. For our empirical analysis, we dropped ideas that were still under review at the end of our empirical window. The exclusion leads us to the final dataset of 8,110 ideas that were generated by 2,705 individuals. Among the 8,110 ideas, 426 ideas (≈ 5%) were implemented by the company. The implementation ratio is a little higher than that of Dell's IdeaStorm (≈ 3%) (Huang et al. 2014). Figure 2 shows the number of contributed ideas at the innovation community during our empirical period.

TABLE 3. 1. Selected submitted ideas in the innovation crowdsourcing community

Idea Title	Idea Status	No. of Kudos Received
Direct relationship with Apple (+ resulting possibilities)	Implemented	379
Nano SIMs	Implemented	348
4G data plan for tethering	Implemented	752
Provisioning unused SIMs as replacement SIMs	Implemented	254
Cash back for unused texts, minutes and data	Not for us	1,261
Text alerts for missed calls	Not for us	130
Remove adult restriction through credit card verification	Not for us	245
500MB then 1p per MB tethering on unlimited internet data plan	Not for us	115

FIGURE 3. 2. Idea submissions over time

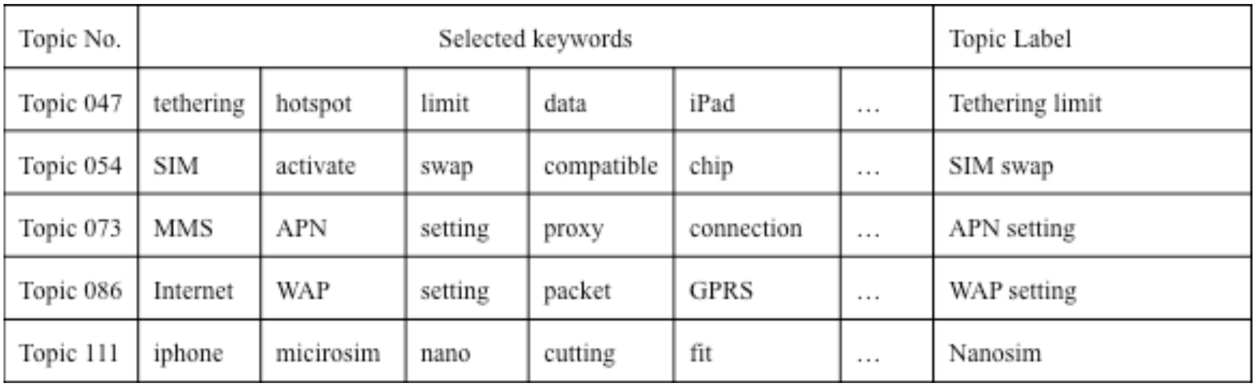


The innovation crowdsourcing setting is very appropriate for studying innovation. Unlike U.S. patent data, which is frequently used in innovation research, researchers can observe all innovation attempts including both successful and failed ones in the innovation crowdsourcing setting. This unique data opportunity enables researchers to observe a complete picture of innovation activities and examine previously unavailable characteristics of them. For instance, we were able to empirically examine quality variance and average success ratio of all ideas submitted by individuals. Innovation crowdsourcing setting also enables us to investigate innovation outcomes at the individual level, which has been difficult to investigate using U.S. patents data because most patents have multiple inventors.

Data from the innovation community were used to construct three outcomes of individuals' new product ideation efforts: originality, quality variance, and quality average of submitted ideas by each individual. Data from the customer support community were used to construct an affiliation network between individuals and information domains. As noted earlier, the act of helping in the customer support community indicates the flow of two important information pieces in new product ideation processes. First, a helping tie represents the outflow of a helper's solution knowledge to a help-seeker. Previous studies have found that the number of answers an individual offered in online knowledge communities can be used as a reliable measure of one's expertise (Zhang et al. 2007): expertise rankings based on the number of answers offered to a certain knowledge domain were highly correlated with the expertise ratings by human raters ($\rho = 0.77$). Second, a helping tie among individuals in the customer support community also represents information inflow from a help-seeker to a helper. Through helping others to find solutions, helpers learn about what the problems or needs of other customers, which is a valuable input for new product ideation. Because the innovation literature has suggested that the possession of needs and solution information is an important antecedent of innovation performance (Baker et al. 1967, Baker and Freeland 1972), we constructed individuals' information network based on their helping activities and examine how it influences their new produce ideation outcomes. Where n is the number of individuals in our dataset and m is the number of information domains, an information network is a two-mode network with size $n \times m$. The information network evolves over time.

In order to construct an information network, we extracted information contents of all helping messages utilizing a natural language processing technique, LDA (Latent Dirichlet Allocation). LDA is a topic classification technique that can automatically discover clusters of messages with similar topics. LDA is a bag-of-words model, which treats each document as a mixture of topics. LDA attempts to learn the topics of each document by backtracking from the words that appear in messages to find a set of topics that are likely to have generated the words. We used a Java-based software called Mallet (McCallum 2002) to run LDA. Based on 115 topics identified by LDA, we labeled all user-generated contents at the customer support community and constructed information networks for each individual for each time period. Figure 3 illustrates selected topics that LDA discovered from our empirical data.

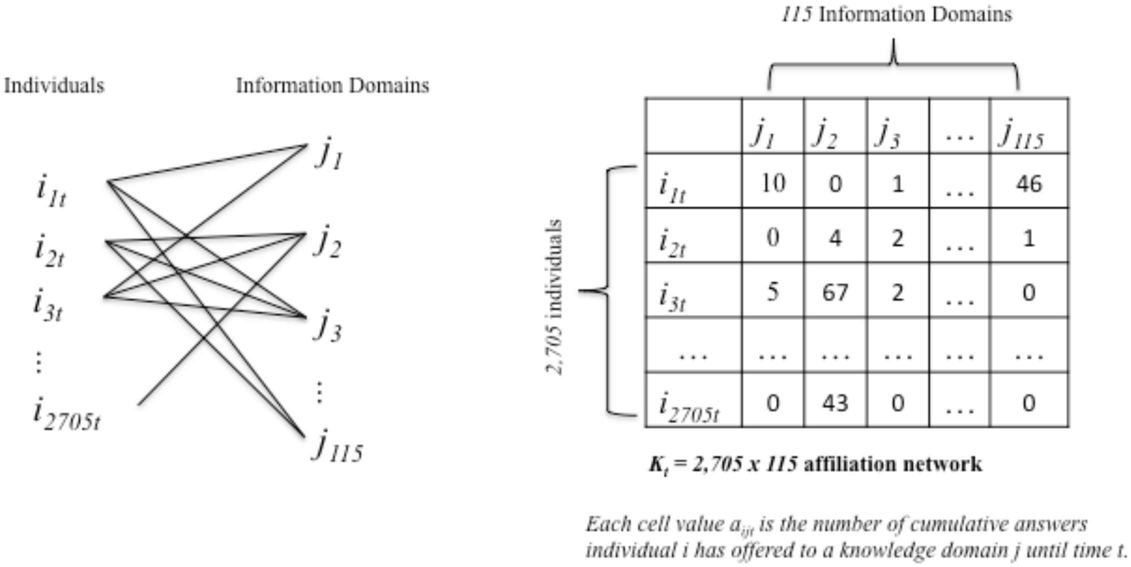
FIGURE 3. 3. Sample topics identified by Latent Dirichlet Allocation (LDA)



Topic No.	Selected keywords						Topic Label
Topic 047	tethering	hotspot	limit	data	iPad	…	Tethering limit
Topic 054	SIM	activate	swap	compatible	chip	…	SIM swap
Topic 073	MMS	APN	setting	proxy	connection	…	APN setting
Topic 086	Internet	WAP	setting	packet	GPRS	…	WAP setting
Topic 111	iphone	micirosim	nano	cutting	fit	…	Nanosim

Each individual i's information network for each time period *t* is captured as an affiliation matrix $K_t = \{a_{ijt}\}$, where the matrix K_t is a two-mode network in which rows represent individuals and columns represent information domain areas. Each cell value a_{ijt} represents the cumulative[8] number of answers individual i has offered to a domain area j up to time period t. With 2,705 individuals and 115 domain areas, K_t is a $2{,}705 \times 115$ matrix. The resulting information networks are used as base matrices to calculate each individual's level of information breadth and depth for each time period. We examine how individuals' level of breadth and depth, which is built until time *t*, affects their new product ideation outcomes at time *t+1*. Figure 4 illustrates hypothetical information network and information matrix K_t.

FIGURE 3. 4. Illustration of hypothetical information network and information matrix, K_t



	j_1	j_2	j_3	…	j_{115}
i_{1t}	10	0	1	…	46
i_{2t}	0	4	2	…	1
i_{3t}	5	67	2	…	0
…	…	…	…	…	…
i_{2705t}	0	43	0	…	0

K_t = 2,705 x 115 affiliation network

Each cell value a_{ijt} is the number of cumulative answers individual i has offered to a knowledge domain j until time t.

Because participation to the innovation community is voluntary, most individuals do not regularly contribute new product ideas. So, our dataset is unbalanced longitudinal data, which consists of different set of individuals each time period. On average, participants propose approximately two new product ideas every three months. So, to observe variance of idea

[8] Individuals may not retain information over time (Argote 2012). As a robustness check, we conducted a supplementary analysis to examine whether the results would change when we capture individual's information network based only on their recent answering activities (i.e., 3 months). The results remain consistent.

quality by each individual, we set a time interval as three months. All variables are constructed at an individual level and independent and control variables are lagged by one period.

3.3.3. Estimation Model

For estimation, we have organized below equations for each of the outcome measures of interest.

$$Originality_{it+1} = \beta_1 + \mathbf{\Gamma}_1 \boldsymbol{X}_{it} + \mathbf{\Delta}_1 \boldsymbol{Z}_{it} + \eta_1 Time_{it} + \alpha_{i1} + \varepsilon_{it+1} \qquad (1)$$

$$Quality\ Variance_{it+1} = \beta_2 + \mathbf{\Gamma}_2 \boldsymbol{X}_{it} + \mathbf{\Delta}_2 \boldsymbol{Z}_{it} + \eta_2 Time_{it} + \alpha_{i2} + \varepsilon_{it+1} \qquad (2)$$

$$Quality\ Average_{it+1} = \beta_3 + \mathbf{\Gamma}_3 \boldsymbol{X}_{it} + \mathbf{\Delta}_3 \boldsymbol{Z}_{it} + \eta_3 Time_{it} + \alpha_{i3} + \varepsilon_{it+1} \qquad (3)$$

$$Acceptance\ Ratio_{it+1} = \beta_4 + \mathbf{\Gamma}_4 \boldsymbol{X}_{it} + \mathbf{\Delta}_4 \boldsymbol{Z}_{it} + \eta_4 Time_{it} + \alpha_{i4} + \varepsilon_{it+1} \qquad (4)$$

where $\beta_1 \cdots \beta_4$ are constant terms, $\boldsymbol{X}_{it}$ includes indicator variables that distinguish individuals based on their information structure at time *t*. Based on the level of information breadth and depth at time *t*, we distinguish individuals into three groups: $Deep\ generalist_{it}, Shallow\ generalist_{it}$, and $Non\ generalist_{it}$. $\boldsymbol{Z}_{it}$ consists of individual-specific control covariates that may also influence *i*'s new product ideation outcomes. $\boldsymbol{Z}_{it}$ includes *i*'s cumulative experience of submitting ideas ($Ideation\ experience_{it}$), helping others ($Helping\ activity_{it}$), and seeking help from others ($Asking\ activity_{it}$). $Time_{it}$ is a control variable to capture any unobservable trend related to time change (e.g., a change of competition level). $\alpha_{i1} \cdots \alpha_{i4}$ are individual-specific random effects. The key parameters of our interests are $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3,$ and $\mathbf{\Gamma}_4$**.**

Observations of our panel dataset are not independent with each other because individuals can appear multiple times in different time periods. A common solution to the matter is to incorporate fixed or random effects for each individual. We have conducted a Hausman test to determine whether the random effects model is unbiased and efficient compared to the fixed effects model (Hausman 1978). The test statistic indicates that the random effects model is unbiased. Thus, we incorporated individual random effects ($\alpha_{i1} \cdots \alpha_{i4}$) into our estimation models. Besides correcting interdependence across observations, the individual random effects also control for any unobserved heterogeneity across individuals that have not been captured by other variables. For instance, some individuals might be exceptionally superior in creativity. Since our dependent measures are continuous, we used ordinary least squares (OLS) regression to estimate our parameters.

When a system of equations is to be simultaneously estimated, seemingly unrelated regression (SUR) model is often suggested (Greene 2011). Each of our four models (1) ~ (4) are a valid linear regression and the parameters can be separately estimated. Although estimates from the separate regression models are consistent, SUR model can estimate parameters more efficiently because it explicitly accounts for correlation of error terms across the equations. There are, however, two important cases when estimates using SUR model are equivalent to those estimated by equation-by-equation OLS regressions. The first case is when there are no cross-equation correlations among error terms. In this case, models are truly unrelated, so there is no need to use SUR model. The second case is when each equation contains exactly the same set of regressors. In this case, there is no gain in estimating a system of equations jointly. Our case belongs to the second special case: all of our four models contain exactly the same regressors on the right-hand-side of the equations. So, we separately estimated each model using ordinary least squares regression.

3.3.4. Measures

3.3.4.1. Dependent variables

Our three dependent variables are idea originality, idea quality variance, and idea quality average. All of the three dependent variables are constructed at an individual level and capture characteristics of ideas that are generated by individual *i* at time *t+1*.

$Originality_{it+1}$ is the outcome measure for our first hypothesis. It measures the proportion of original (non-redundant) ideas among all ideas that are submitted by individual *i* at time *t+1*. Once submitted, all ideas are initially screened for originality. If identified as a redundant idea (either identical or very similar to previously submitted ideas), the redundant idea is linked to the original idea. After being linked back to their original ideas, the redundant ideas are marked as *redirected*. We used this idea status information to construct our measure of $Originality_{it+1}$. The formula for the variable $Originality_{it+1}$ is below.

$$Originality_{it+1} = 1 - \frac{RI_{it+1}}{TI_{it+1}} \qquad (5)$$

where RI_{it+1} is the number of redirected ideas submitted by *i* at time *t+1* and TI_{it+1} is the total number of ideas submitted by *i* at time *t+1*. During our empirical period, approximately 67% of ideas submitted by each individual were original ideas.

$Quality\ Variance_{it+1}$ captures the variability of idea qualities that are submitted by individual *i* at time *t+1* and is used to test our second hypothesis. At the innovation community, customers can evaluate others' ideas by casting one vote for each submitted idea if they like the idea and would like it to be implemented. The vote is called Kudo in the innovation community. Because the number of customer votes received by an idea represents potential usefulness or popularity of a new product idea, the company utilizes the number of customer votes as an initial filter to sort out ideas. For example, only ideas that have received 20 or more customer votes are eligible for a management review. Similarly, we used the number of customer votes as a proxy of idea quality and calculated the standard deviation of customer votes of all ideas submitted by individual *i* at time *t+1* to measure idea quality variance.

$$QualityVariance_{it+1} = \sqrt{\frac{\sum_{n=1}^{TI_{it+1}}(V_{int+1} - \overline{V_{\imath t+1}})^2}{TI_{it+1}-1}} \qquad (6)$$

where TI_{it+1} is the total number of ideas submitted by *i* at time *t+1* and V_{int+1} is the number of customer votes for idea *n* that an individual *i* has submitted at time *t+1*. $\overline{V_{\imath t+1}}$ is the mean customer votes of all ideas generated by *i* at *t+1*. The average idea quality variance is 5.56.

To test how individuals' information structure influences the quality of their new product ideas, we employed two measures of idea quality: $Quality\ Average_{it+1}$ and $Acceptance\ Ratio_{it+1}$. $Quality\ Average_{it+1}$ measures average customer votes of all ideas submitted by individual *i* at time *t+1*. An idea submitted earlier would have more time to garner more customer votes. $Time_{it}$ controls for this effect. Ideas received about 5 customer votes on average and the most popular idea received 518 customer votes. $Acceptance\ Ratio_{it+1}$ captures the proportion of accepted ideas among all ideas submitted by individual *i* at time *t+1*. Because the company is likely to accept ideas that satisfy all the criteria of innovative ideas (originality, usefulness, and feasibility), $Acceptance\ Ratio_{it+1}$ can be considered as a more conservative measure of idea quality than $Quality\ Average_{it+1}$. Because receiving many customer votes does not necessarily mean that the idea will be implemented, the two measures are not highly correlated. The Pearson correlation (ρ) of $Quality\ Average_{it+1}$ and $Acceptance\ Ratio_{it+1}$ is 0.36 (Table 1). The formulas for the variables are below.

$$Quality\ Average_{it+1} = \frac{\sum_{n=1}^{TI_{it+1}} V_{int+1}}{TI_{it+1}} \tag{7}$$

$$Acceptance\ Ratio_{it+1} = \frac{AI_{it+1}}{TI_{it+1}} \tag{8}$$

where TI_{it+1} is the total number of ideas submitted by *i* at time *t+1*, V_{int+1} is the number of customer votes for idea *n* that an individual *i* has submitted at time *t+1,* and AI_{it+1} is the number of accepted ideas submitted by *i* at time *t+1*. On average, ideas generated by each individual receive about 5 customer votes and about 5% of submitted ideas are implemented. The implementation ratio is a little higher compared to that of Dell's idea storm (≈3%).

3.3.4.2. Independent variables

Our independent variables are indicator variables that distinguish individuals based on their level of information breadth and depth. The groups of our main interest are (1) one with broad information and deep knowledge and (2) the other one with broad information but no deep knowledge. For easier reference, we call the first group as *deep generalists* and the second group as *shallow generalists*. $Deep\ Generalist_{it}$ is coded as 1 if an individual *i* is a deep generalist at time *t*, 0 otherwise. $Shallow\ Generalist_{it}$ is coded in the same manner. Our control group of individuals is non-generalists. In order to identify who belongs to which group, we employed following procedure to calculate knowledge breadth and depth for each individual and for each time period.

Information breadth, which captures the scope of an individual *i*'s information represented by *i*'s helping network up to time *t*, is measured by counting the number of distinct domain areas on which an individual *i* has provided at least five answers up to time *t*. As a robustness check, we also calculated information breadth based on various thresholds. With thresholds of three answers and up, the directions and significance of our results remain consistent. On average, individuals are knowledgeable on approximately 11 domain areas with standard deviation of 29.47. Based on the breadth measure, we segregated individuals into generalists and non-generalists. We considered that an individual to be a generalist if an individual's information breadth is one standard deviation above the mean level: the threshold for generalists is 40 topic areas[9]. Further, we considered that an individual to be a non-generalist if information breadth is lower than the mean level (11 domain areas).

We further segregated generalists based on whether they have deep knowledge in any of their knowledgeable areas. Deep generalists are the ones who possess expert knowledge in at least one domain area[10] whereas shallow generalists are the ones who do not have any expert knowledge in any of the domain areas. We used the total number of answers offered to each domain area as a proxy to measure an individual's degree of knowledge depth in a domain area. We considered that an individual possesses deep knowledge in a domain area if the individual belongs to the top 10% solution providers to the domain area. As a robustness check, we experimented with various thresholds (10% ± 5%) to determine deep knowledge. The directions and significance of our results remain consistent. The threshold to be an expert varies across domains but on average an individual had to offer at least 63 solutions in order to be considered to have deep understanding on a domain area.

[9] We conducted sensitivity analyses to see how much our results are sensitive to the threshold. As we lower the threshold to the mean level (μ) from one standard deviation above the mean ($\mu + \sigma$), the effect magnitude becomes larger but effect statistical significance stayed the same.

[10] For our main analysis, we did not further differentiate deep generalists based on how many expertise areas an individual possesses.

3.3.4.3. Control variables

Individuals' new product ideation outcomes may also be influenced by other factors. In order to tease out the effect of individuals' information structure on new product ideation outcomes, we incorporated several control variables. First, individuals might learn to produce high quality ideas in the course of generating multiple ideas. To control for this learning-by-doing effect (Argote 2012), we included $Ideation\ experience_{it}$ variable, which is measured by the cumulative number of ideas submitted by each individual up to time t.

Second, $Helping\ activity_{it}$ controls for the total *quantity* of information that one's helping network represents in order to tease out the effect of information *structure* on the outcome variables. By incorporating this control variable, we can examine how information content *structure* (in terms of breadth and depth) might affect innovation outcomes independent of the information *quantity*. $Helping\ activity_{it}$ is measured as the total number of all answers contributed by individual *i* up to time *t*.

Third, we incorporated $Asking\ activity_{it}$ to control for any effect from the amount of solutions an individual *i* has obtained through the customer support community. $Asking\ activity_{it}$ is calculated by summing all answers obtained by individual *i* up to time *t*. Fourth, our dataset spans over three years. In order to control for any unobserved differences in outcome variables caused by time changes (e.g., the competition level), our model includes the variable $Time_t$, which is an index for time period.

We do not have demographic information of individuals participating to the communities. Due to this data limitation, we included an individual-specific effect into our estimation model in order to control for any unobserved individual-specific heterogeneity. Based on the result of a

Hausman specification test, we incorporated random effect for each individual instead of fixed effect. This random effect controls for any intrinsic differences (e.g., creativity) across individuals that are not captured by other variables.

3.4. Results

3.4.1. Generalists vs. Non-Generalists

The descriptive statistics and correlation among variables are reported in Table 2. Table 3 presents the results of our main estimation model, which examines the differences of new product ideation outcomes between generalists and non-generalists. The control group for all of our analyses is non-generalists. Model 1 tests our first hypothesis: whether generalists are likely to create more original ideas than non-generalists. The positive and statistically significant coefficient ($\beta = 0.18$, $p<0.001$) indicates that our first hypothesis is supported. Generalists, who have engaged in helping activities in diverse domain areas, tend to create original ideas about 18% more on average compared to non-generalists.

Model 2 tests our second hypothesis: whether ideas created by generalists are more likely to be highly variable in their quality. The results support our hypothesis. Compared to non-generalists, generalists tend to create ideas that are more variable in their quality ($\beta = 4.6$, $p<0.001$). Although we have not hypothesized the relationship, we tested whether generalists tend to create higher quality ideas than non-generalists in models 3 and 4. The results of models 3 and 4 suggest that generalists tend to create ideas that have higher levels of quality than non-generalists in both measures of idea quality (Quality Average and Acceptance Ratio). On average, generalists tend to receive approximately 4 more customer votes for their submitted ideas. More importantly, their ideas' implementation ratio is about 3.58% higher than that of non-generalists.

TABLE 3. 2. Descriptive statistics and correlation

Models 1 to 4

		mean	std.dev.	1	2	3	4	5	6	7	8	9
1	Idea Originality	0.33	0.42	1.00								
2	Idea Quality Variance	5.57	11.12	0.27	1.00							
3	Idea Quality Average	4.48	10.35	0.27	0.87	1.00						
4	Idea Acceptance Ratio	0.04	0.16	0.30	0.23	0.26	1.00					
5	Generalist	0.12	0.32	0.18	0.19	0.18	0.10	1.00				
6	Ideation experience	2.21	7.07	0.15	0.10	0.12	0.06	0.52	1.00			
7	Helping activity	119.06	612.48	0.13	0.18	0.19	0.11	0.53	0.46	1.00		
8	Asking activity	24.82	95.92	0.13	0.11	0.10	0.05	0.60	0.60	0.52	1.00	
9	Time	5.95	2.10	-0.14	0.02	-0.05	-0.13	0.11	0.13	0.12	0.13	1.00

Note. n = 3,697 for all variables except for Quality Variance (n = 1,277)

Models 5 to 8

		mean	std.dev.	1	2	3	4	5	6	7	8	9	10
1	Idea Originality	0.33	0.42	1.00									
2	Idea Quality Variance	5.57	11.12	0.27	1.00								
3	Idea Quality Average	4.48	10.35	0.27	0.87	1.00							
4	Idea Acceptance Ratio	0.04	0.16	0.30	0.23	0.26	1.00						
5	Deep generalist	0.09	0.29	0.19	0.20	0.19	0.11	1.00					
6	Shallow generalist	0.02	0.14	0.03	0.03	0.01	0.00	-0.05	1.00				
7	Ideation experience	2.21	7.07	0.15	0.10	0.12	0.06	0.49	0.17	1.00			
8	Helping activity	119.06	612.48	0.13	0.18	0.19	0.11	0.58	0.00	0.46	1.00		
9	Asking activity	24.82	95.92	0.13	0.11	0.10	0.05	0.58	0.15	0.60	0.52	1.00	
10	Time	5.95	2.10	-0.14	0.02	-0.05	-0.13	0.09	0.07	0.13	0.12	0.13	1.00

Note. n = 3,697 for all variables except for Quality Variance (n = 1,277)

TABLE 3. 3. Predicting new product ideation outcomes: Generalists vs. Non-generalists

	Model 1	Model 2	Model 3	Model 4
Dependent variables	Idea Originality$_{it+1}$	Idea Quality Variance$_{it+1}$	Idea Quality Average$_{it+1}$	Idea Acceptance Ratio$_{it+1}$
Individual type				
Generalist$_{it}$	0.1846 ***	4.6394 ***	4.4259 ***	0.0358 **
	(0.028)	(1.027)	(0.696)	(0.011)
Controls				
Ideation experience$_{it}$	0.0053 ***	-0.0116	0.0424	0.0004
	(0.001)	(0.037)	(0.031)	(0.000)
Helping activity$_{it}$	0.0000 *	0.0016 ***	0.0024 ***	0.0000 ***
	(0.000)	(0.000)	(0.000)	(0.000)
Asking activity$_{it}$	0.0000	-0.0041	-0.0067 **	0.0000
	(0.000)	(0.003)	(0.002)	(0.000)
Time$_{it}$	-0.0336 ***	-0.0904 ***	-0.3889 ***	-0.0114 ***
	(0.003)	(0.148)	(0.080)	(0.001)
Constant	0.4986 ***	5.0405 ***	6.0700 ***	0.1000 ***
	(0.020)	(0.892)	(0.499)	(0.008)
No. of observations	3,697	1,277	3,697	3,697
No. of individuals	2,643	861	2,643	2,643
Mean VIF	1.62	1.63	1.62	1.62
Control group for all models is non-generalists.				

*** *Significant at 0.001 level (two-tailed)*
** *Significant at 0.05 level (two-tailed)*
* *Significant at 0.1 level (two-tailed)*

3.4.2. Deep Generalists vs. Shallow Generalists vs. Non-Generalists

The findings that generalists tend to create higher quality ideas (refer to Models 3 and 4 in Table 2) suggest that broad information seems to be the key driver for successful innovation. In this study, we test the boundary condition of the impact of broad information on idea quality. In the hypothesis 3, we proposed that the effect of broad information on idea quality is contingent on the presence of deep knowledge. To test hypothesis 3, we further segregated generalists based on whether they possess deep knowledge or not. Table 4 presents the results of regression analyses that predict new product ideation outcomes of deep and shallow generalists. Similar to previous models, non-generalists are the control group in models 5 through 8.

TABLE 3. 4. Predicting new product ideation outcomes: Deep generalists vs. Shallow generalists vs. Non-generalists

	Model 5	Model 6	Model 7	Model 8
Dependent variables	Idea Originality$_{it+1}$	Idea Quality Variance$_{it+1}$	Idea Quality Average$_{it+1}$	Idea Acceptance Ratio$_{it+1}$
Individual type				
Deep Generalist$_{it}$	0.2131 ***	5.1160 ***	5.3880 ***	0.0428 ***
	(0.032)	(1.130)	(0.777)	(0.012)
Shallow Generalist$_{it}$	0.1055 **	3.2222 *	1.7500	0.0164
	(0.048)	(1.730)	(1.180)	(0.019)
Controls				
Ideation experience$_{it}$	0.0054 ***	-0.0095	0.0449	0.0004
	(0.001)	(0.037)	(0.031)	(0.000)
Helping activity$_{it}$	0.0000 *	0.0014 **	0.0021 ***	0.0001 ***
	(0.000)	(0.000)	(0.000)	(0.000)
Asking activity$_{it}$	0.0000	-0.0042	-0.0070 **	-0.0001
	(0.000)	(0.003)	(0.002)	(0.000)
Time$_{it}$	-0.0332 ***	-0.0785	-0.3777 ***	-0.0115 ***
	(0.003)	(0.149)	(0.079)	(0.001)
Constant	0.4967 ***	4.9729 ***	6.0049 ***	0.0995
	(0.020)	(0.895)	(0.499)	(0.008)
Number of observations	3,697	1,277	3,697	3,697 ***
Number of individuals	2,643	861	2,643	2,643
Mean VIF	1.56	1.58	1.56	1.56

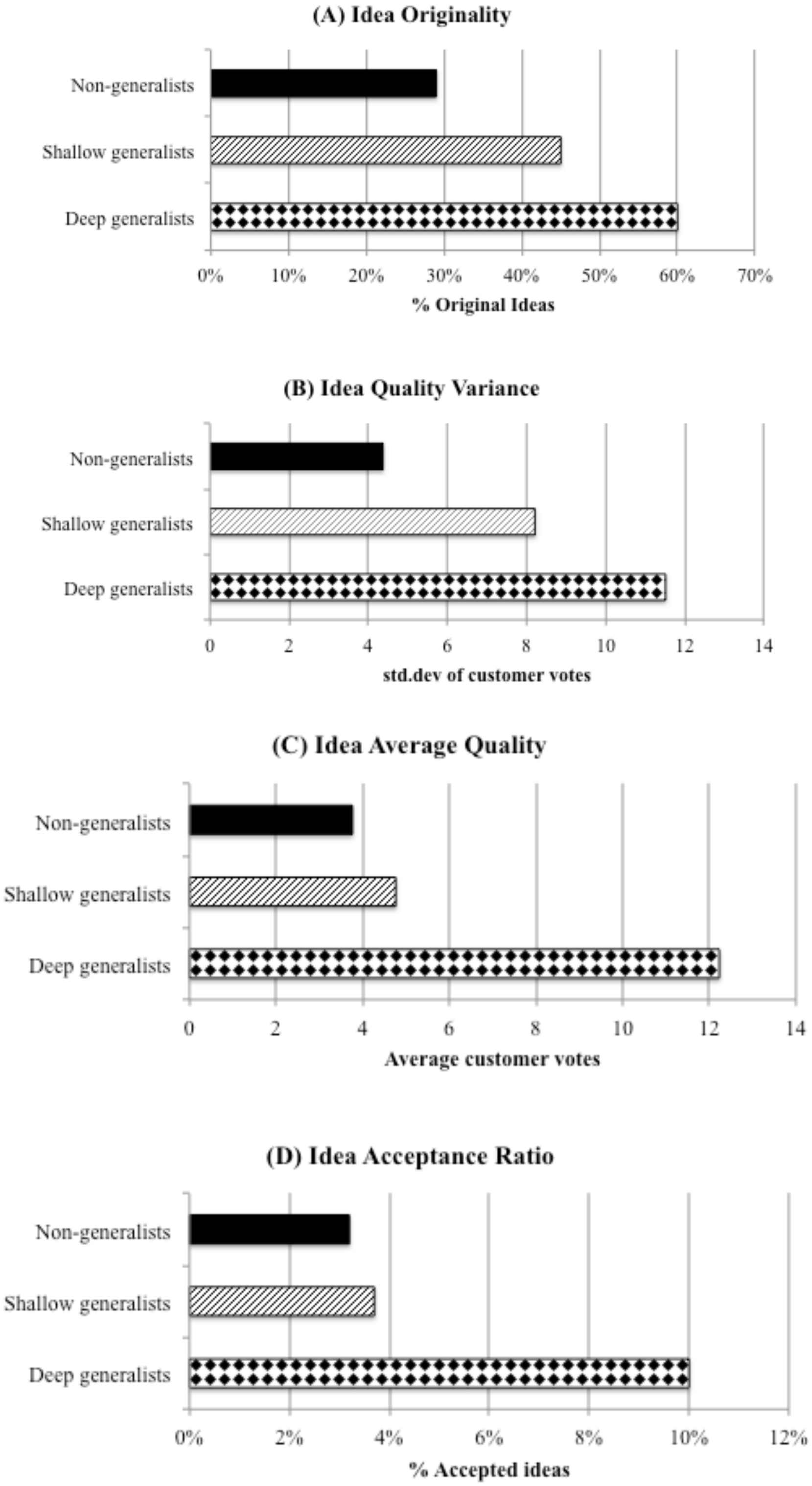
Control group for all models is non-generalists.

**** Significant at 0.001 level (two-tailed)*
*** Significant at 0.05 level (two-tailed)*
** Significant at 0.1 level (two-tailed)*

Model 5 shows that both deep and shallow generalists create more original ideas than non- generalists. On average, deep generalists tend to create 21.31% more original ideas and shallow generalists tend to create about 10.55% more original ideas compared to non-generalists. In addition, the quality of ideas created by both deep and shallow generalists tend to be more variable compared to those by non-generalists (see Model 6). Put differently, the quality of ideas created by both types of generalists tends to be more unpredictable compared to the quality of ideas created by non-generalists.

Our third hypothesis posits that only those generalists who possess deep knowledge are able to reap the benefit of broad information and create higher-quality ideas. Our third hypothesis is supported by the results of models 7 and 8. As the statistically significant positive coefficient of deep generalists and insignificant coefficient of shallow generalists indicate, it is only deep generalists who outperform non-generalists on the quality dimension. Our regression analysis shows that deep generalists tend to create higher quality ideas in both measures of idea quality, Quality Average and Acceptance Ratio. Although shallow generalists tend to be able to create more original ideas, the overall quality of their ideas did not differ from those created by non-generalists. Figure 5 depicts the average new product idea outcomes of the three different groups of individuals. To create Figure 5, we split individuals into three groups by their information structure and calculate the mean of the new product idea outcomes of each group. For all four dimensions, deep generalists' ideas outperform those by shallow generalists and non-generalists. In particular, the gap is significantly large for idea quality dimensions. Compared to 3% (non generalists) and 4% (shallow generalists) implementation ratio, 10% of ideas by deep generalists are implemented.

FIGURE 3. 5. New product idea outcomes by three different groups of individuals


(A) Idea Originality

(B) Idea Quality Variance

(C) Idea Average Quality

(D) Idea Acceptance Ratio

There could be alternative explanations for our findings. We incorporated a number of control variables in order to tease out the effect of individuals' information structure on their new product idea outcomes. The control variables include individuals' experience of submitting ideas to the innovation community ($Ideation\ experience_{it}$), total helping activities ($Helping\ activity_{it}$) and asking activities ($Asking\ activity_{it}$) in the customer support community. The estimation results $Ideation\ experience_{it}$ were surprising to us. Based the extensive research on learning-by-doing (Argote 2012), we assumed that an individual's previous experience of idea generation would help him or her to create more original and high quality ideas. However, our findings suggest that previous experience of creating ideas only helps individuals to generate original ideas but does not help to generate higher quality ideas.

$Helping\ activity_{it}$ controls for the total *quantity* of information that one's helping network represents. By incorporating this control variable, we were able to tease out how information content *structure* (in terms of breadth and depth) might affect new product ideation outcomes independent of the information *quantity*. Overall, we found that helping activities in the customer support community helps individuals to create high variance and high quality ideas (positive and significant coefficients for $Helping\ activity_{it}$ in model 6,7, and 8).
$Asking\ activity_{it}$ controls for the total asking activities in customer support community. We found that individuals who seek help more tend to generate ideas that garner smaller number of customer votes. For idea originality and idea acceptance ratio, $Asking\ activity_{it}$ had null effect. $Time_{it}$ controls for any unobservable differences of outcome measures due to time. For instance, an idea submitted earlier would have more time to garner more customer votes. $Time_{it}$ controls for this systematic difference of outcome variables. As time passes we found that it

became harder to create original ideas and to create ideas that are later implemented by the company.

We also conducted sensitivity analyses to check robustness of our results across different threshold levels of measuring individuals' information breadth and depth level. For our main analysis, we assumed that an individual is knowledgeable on a domain area if he or she has provided at least five answers up to time *t*. As a robustness check, we calculated information breadth based on various thresholds. From thresholds of three answers and up, the directions and significance of our results remain consistent. Further, we checked sensitivity of our results across different cutoff points for generalists and non-generalists. For our current analysis, the cutoff point for generalist is one standard deviation above the mean ($\mu + \sigma$), which is 40 domain areas. As we lower the cutoff point to the mean (μ) level, 11 domain areas, the effect magnitude of $Generalist_{it}$, $Deep\ generalist_{it}$, and $Shallow\ generalist_{it}$ got larger but the statistical significance stayed the same. We also took similar approach to do sensitivity analyses for information depth. For our current analysis, we considered that an individual possesses deep knowledge in a domain area if the individual belongs to the top 10% solution providers to the domain area. Thresholds between 5~15% produced consistent results.

3.5. Discussion and Conclusions

Traditionally, new product development efforts have been mostly concentrated within organizations' internal R&D departments. Since the last decade, formal organizations have increasingly started to crowdsource new product ideas from their customers. Sourcing new product development efforts from customers may add great value to companies because customers can tap important market needs information into new product development process.

This study examined the complementarity of individuals' activities between innovation crowdsourcing and customer crowdsourcing communities. Specifically, we investigated how individuals' helping activities in a customer support community influence their new product ideation outcomes in three dimensions: idea originality, quality variance, and quality average. We focused on helping activities because helping ties represent flow of two important information types: (1) pre-existing knowledge on a domain area and (2) needs information of other customers on a domain area. Because both types of information are valuable inputs for the new product development process, we constructed individuals' information networks based on their helping activities. Then, we examined how the variation of one's information *structure* (in terms of breadth and depth) influences new product ideation outcomes after controlling for the *quantity* of information one obtains through customer support crowdsourcing community.

The value of broad knowledge on innovation has been widely acknowledged by both practitioners and scholars. Because innovation process involves substantial amounts of recombining or rearranging preexisting knowledge components, with richer ingredients, it seems like we should get better output. In this study, we challenged this claim and proposed that broad knowledge by itself is not sufficient to generate high-quality ideas. We argued that the effect of broad information on idea quality is contingent on the presence of deep knowledge. Firstly, deep knowledge helps individuals to make more meaningful recombinations by enabling them to see linkages across diverse information. Secondly, deep knowledge helps individuals to generate practical ideas by helping them to identify constraints of potential solutions. Consistent with our predictions, we found that only those individuals who possess both broad and deep knowledge were able to create higher quality ideas than non-generalists. Broad information in general helped individuals to create original ideas. The quality of ideas created by generalists, however, was highly unpredictable: some ideas are very high quality but others are low quality ideas. Only generalists with deep knowledge were able to channel their original ideas into high quality contributions.

This study makes several theoretical contributions. Perhaps due to data limitations or challenges in data processing, with the exception of Aral and Van Alstyne (2011), network studies have not analyzed the information content that flows through network ties. Rather, they assumed the characteristics of information based on an actor's network position. For example, an actor occupying structural hole is presumed to obtain more novel information than an actor who is embedded in a dense network. Also, a weak tie is presumed to transmit novel information. Using a natural language processing technique, we were able to extract the actual information flowing through helping network. This content analysis confirms that even though individuals obtain the same amount of information, the content of information they obtain may differ greatly. We further found that this content difference influences individuals' new product ideation outcomes after controlling for the information quantity.

Our findings also advance understanding of how broad and deep knowledge influences innovation outcomes. Previous innovation research has examined how diverse knowledge influences innovation performance (e.g., Sampson 2007, Taylor and Greve 2006). Although considerable attention has been devoted to examining the effect of broad knowledge on innovation, little attention has been paid to the boundary conditions under which broad knowledge advances innovation. This study extends prior work by showing that the positive effect of broad knowledge is contingent on the presence of deep knowledge. Generalists with deep knowledge were able to channel their original ideas into high quality contributions while generalists with shallow knowledge were not. Further, this study advances our understanding of the effect of knowledge structure on other innovation outcomes. The innovation crowdsourcing context enabled us to observe a complete picture of new product ideation efforts, which includes both successful and failed ones. Consequently, we were able to investigate the effect of knowledge structure on diverse outcomes, including the originality, variance, and quality of ideas.

This study also has practical implications. Our findings suggest that customer support communities contain rich data on customer needs information. Companies should take advantage of the rich data in these communities to monitor the complaints and problems that customers are experiencing with their product or services. Also, in our data, we were able to trace which customer problems motivated an individual to create ideas to handle the problems. So combining customer support and innovation crowdsourcing data, companies can provide more targeted feedback to question posters in customer support communities when new products or service become available the problem.

This study has a few limitations that suggest directions for future research. First, in the current study, we captured information that individuals obtain from a customer support community based on their traces of helping activities. Yet it is also possible that individuals may collect information without leaving traces. If users' viewing log data are available, researchers would be able to capture more complete information that individuals obtain from a customer support community. Second, we constructed individuals' information network based on their helping activities because helping ties represent the flow of solution information and customer needs information, which is valuable input for innovation process. Nonetheless, individuals can also obtain information by seeking help. Through asking, help seekers reveal their current problems or needs in exchange for solution information. It would be interesting to examine how information network based on individuals' help-seeking behavior interact with information network based on individuals' helping behavior.

In conclusion, this study investigated the complementarity of individuals' activities between two crowdsourcing communities: a customer support and an innovation crowdsourcing community. Overall, we found that helping activities at a customer support community help individuals to create better quality ideas at innovation community. Specifically, we found that generalists, who have provided solutions on diverse problem areas, are likely to create more original ideas. Yet, we further found that generalists who only know shallowly across diverse domain areas are indifferent from non-generalists in their ability to create ideas that are later implemented by a company. Only those generalists who possess expert knowledge in at least one domain area tend to outperform non-generalists.

Chapter 4.

Online Community with Propinquity: The Effect of Physical Distance on Membership Herding[11]

4.1. Introduction

Increasingly, organizations are utilizing electronic networks to promote knowledge sharing among their employees. An online knowledge community, where members of an organization can exchange information over a virtual space, is a popular example of such electronic networks. According to McKinsey survey (2013), more than a half of surveyed firms have adopted online communities to manage their knowledge in 2013, which is a significant jump compared to only 28% in 2009. Despite its significant growth in terms of the number of adoptions, it is well documented that not all members in an online community contribute equally. Only a minority of participants is actively contributing to online communities: about 1~2% of users account for almost all the action, 9% of users contribute a little, and the remaining are free riders (Whittaker et al. 1998). With such a low engagement by users, an online community will have difficulty to be sustainable over time. In fact studies have revealed that it is not uncommon for online communities, even the once largest social networking community such as Friendster, to suddenly lose significant portion of its members and to eventually fail (Garcia et al. 2013, Oh and Jeon 2007).

[11] This essay is a joint work with David Krackhardt and Param Vir Singh.

In an effort to encourage users' contribution to online communities, scholars investigated what motivates individuals to contribute contents. Since participation is voluntary in online communities, it is found that individuals who value pro-social behavior tend to contribute more (Wasko and Faraj 2005). Additionally, other factors such as reputation and sense of self-worth are documented to increase individuals' motivation to contribute to online communities (Bock et al. 2005, Faraj and Johnson 2011, Wasko and Faraj 2005). Among others, reputation is a crucial individual motivation to contribute, especially when an online community is a career-related one such as an open source software community and an intra-organizational community (Constant et al. 1996, Lerner and Tirole 2002, Lakhani and Wolf 2005, Oreg and Nov 2008, Stewart 2005). In fact, it is now known that reputable individuals in open source software communities are often recruited through the communities.

The insights gained from the previous studies guided further research on how to design an online community to promote more active engagements. For example, based upon the finding that reputation and sense of self-worth are important drivers to contribute, Ma and Agarwal (2007) have proposed that adding features such as virtual co-presence and deep profiling to an online community would advance community members' perceived identity verification by others, which in turn increases individuals' level of contribution. Also, scholars in Human-Computer Interaction (HCI) are actively conducting research on how other design features such as gamification[12] improve members' active engagement to online communities (e.g., Bista et al. 2012, Deterding et al. 2011). Gamification features such as points or badges are found to significantly increase user engagement. For example, a New York based food ordering website,

[12] *Gamification* is an informal umbrella term that refers to the use of video game elements in non-gaming systems to improve user engagement. Examples of gamification features are badge, points, leaderboard, and time constraint (Deterding et al. 2011).

Campus food.com, experienced up to 20% increase in the return of new users after adding gamification features (MacMillan 2011).

In this paper, we extend these prior studies by exploring social factors that drive individuals' contribution to online communities. Humans have intrinsic propensity for consensus: people want to follow what others do. Since the seminal papers by Banerjee (1992) and Bikhchandani et al. (1998), scholars have documented extensive evidence of human herding behavior in diverse settings. To name a few, it has been found that people tend to mimic others to make purchase decisions (Bikhchandani et al. 1998, Salganik et al. 2006), investment decisions (Agarwal et al. 2011, Scharfstein and Stein 1990, Zhang and Liu 2012), product rating decisions (Lee et al. forthcoming, Scholosser 2005), and an organ transplant decision (Zhang 2010).

Despite its prevalence, we have a limited understanding on how such herding tendency might affect individuals' contribution decision to online communities. Although previous studies have examined how other social factors such as direct- and general-reciprocity drive user engagement in online communities, with an exception of Oh and Jeon (2007), no studies focused on how herding tendency affects user engagement in online communities. In their conceptual paper, Oh and Jeon (2007) offered a valuable insight on how members' exit decision in an open source software (OSS) community might be explained by herding tendency. Specifically, they modeled participants' interaction dynamics using the Ising theory. The Ising theory is a physics theory on magnetic interactions among microscopic magnets that exhibit herding tendency. Drawing upon the theory they explain how once successful online communities may suddenly fail.

In this study, we theorize that participants will be motivated to contribute more if their virtual neighbors, with whom they have interacted over an online community, are active. In addition, we propose that this herding tendency will become stronger if their virtual neighbors are also geographically proximate to them. Using field data from an enterprise online community, we empirically test our propositions. Finally, we discuss the implications of our findings on community design as well as on the evolution of online community population.

4.2. Theory and Hypotheses

4.2.1. Herd Behavior in Online Communities

Humans have intrinsic propensity for consensus. When people make decisions they tend to refer to the decisions made by previous decision makers, which results in behavior patterns of herding (Banerjee 1992, Bikhchandani et al. 1998). We propose that participants of online communities will also exhibit this propensity to "follow" others when they decide how actively they engage in online communities for the following reasons.

Any social groups, both traditional and virtual, have to offer net benefits in order to attract and/or to retain members (Moreland and Levine 1982). For example, social groups offer benefits such as opportunities to affiliate with others and access to new information. Because groups should possess available resources (e.g., knowledge, time, and financial assets) (Butler 2001, Rice 1982) in order to offer benefits to their members, previous studies proposed that the amount of available resource is correlated with the size of a social group. For example, there is higher probability that a larger voluntary association has more economic resources (McPherson 1983) and a larger group has access to more information to solve the problems at hand

(Wittenbaum and Stasser 1996). As a result, it is argued that the entire size of a group leads to further growth of the group. For example, in a virtual setting of a Listserv, Butler (2001) found that the membership size of a listserv predicts new membership gains.

Although the entire size of a social group is proposed to increase the level of benefit of a group, we suggest that the entire group size is only a crude indicator of an individual's potential benefit from a group. Instead of the entire size of a community, we suggest that an individual's benefit is more tightly tied to the activities of other participants with whom an individual have interacted in an online community. We will refer to those participants as *virtual neighbors*. Because virtual ties are more likely to be formed based on participants' common interest (e.g., knowledge domain), virtual neighbors are more likely to share common interest. As a result, if an individual's virtual neighbors are actively engaging in the community, there is higher chance that there exist someone who can potentially solve his or her problems. Also, there is higher chance that there exist someone who can potentially seek help for which the focal individual can provide solutions.

Herding tendency may also be driven by the uncertainty of decision outcome. When there exists uncertainty of a decision outcome (as it is in most cases), we often defer our decision until other people make their decisions. Bikhchandani et al. (1998) states that this tendency will become stronger as one's private information on the outcome is more uncertain: when one has only a little amount of worthwhile private information, one tends to put more weight on public information that is accumulated through others. For example, when an innovative product such as Apple's iWatch or Google's self-driving car rolls out, majority of us prefer to wait until other people experience the products and more public opinion is formed about the products. Similar to such situation, the benefit of actively engaging in online communities is uncertain to participants.

Because an individual incurs costs to create contents for online communities (e.g., time and effort to formulate and to physically type posts), participants need to evaluate the net benefit of contributing contents. Even if an individual has some private information through one's own experience of participation, we expect that this evaluation process will occur constantly because the benefit of online community keeps changing as new members join and existing members exit the community.

Besides the uncertainty related to benefit from contributing to online communities, it might also be highly uncertain to individuals whether active engagement to an online community is a desirable behavior. We expect that individuals will care and worry more about this uncertainty, as it is easier to locate real identity of online community participants. Let us take an example of an online community within a company. Contributing contents to organizational online communities requires time and efforts, which naturally takes away the time and effort that an employee can put on one's current job responsibility. Although an employer company initiated an online community and encourages using it, employees are likely to need further psychological assurance from their peers. If many others are participating, an individual will feel safer to increase one's engagement level because one can share the blame with others just in case there are any adverse consequences of participating. This "sharing-the-blame" motivation is also found to drive money managers' tendency to herd (Scharfstein and Stein 1990).

Given the uncertainty of net benefit as well as desirability of active engagement, we expect that individuals will look to others, especially their virtual neighbors, when they decide how actively they will engage in online communities. Besides the fact that virtual neighbors share common interest, which is expected to lower uncertainty of contribution benefit, the activities of virtual neighbors are likely to be more salient to a focal individual than the activities of other participants with whom a focal individual has never interacted. Therefore, we hypothesize that:

H1: Participants tend to herd to their virtual neighbors.

4.2.2. Geographical Proximity and Herding in Online Communities

Scholars have emphasized the fundamental role of geographical proximity (propinquity) on human interaction. Evidence shows that people are more likely to form relationship with nearby others because propinquity increases chances to meet, which in turn increases the probability of forming ties (e.g., Festinger et al. 1950, Hampton and Wellman 2000, Kraut et al. 1990, McPherson et al. 2001).

Yet as the Internet brought possibilities to easily connect over distance, propinquity seemed to lose its role in human interaction, at least in virtual settings. In fact, it is argued that the world becomes flat (Friedman 2007) and that physical distance is dead (Cairncross 1997) with the arrival of the Internet. Nevertheless, empirical studies showed mixed evidence for this claim. It seems that the impact of propinquity on virtual relationships depends on contexts. In one setting where the goal of an electronic network is to *complement* face-to-face relationships and to maintain existing relationships (e.g., Facebook), it is expected that geographical proximity predict virtual relationships (Mislove et al. 2010, Skopek et al. 2011, Takhteyev et al. 2012, Thelwall 2009). Yet in another setting where the goal is to *substitute* face-to-face relationships and to facilitate meetings among strangers based on their common interest (e.g., Stackoverflow), it has been argued that geography is not expected to influence virtual connections (Van Alstyne and Brynjolfsson 2005).

However, in this study, we argue that even in a setting where an electronic network is to *substitute* face-to-face relationships, geographical proximity may still paly an important role if geographical proximity is positively correlated with social importance. For example, the goal of an organizational knowledge sharing community, the empirical setting of this study, is to facilitate connections among unacquainted employees based on their common interest (e.g., knowledge domains such as .Net, Java). The main motivation to contribute contents in such a setting is to gain reputation from one's peers (Constant et al. 1996, Lerner and Tirole 2002, Wasko and Faraj 2005). As proposed by previous studies, reputation from purely virtual neighbors is valuable and thereby is expected to motivate people to actively engage. However, we expect that reputation will be even more valuable if it comes from virtually 'and' geographically proximate neighbors because in this case, one can reasonably expect that his or her online reputation can spill over to offline. So, we expect that participants will be more motivated if their active virtual neighbors are also geographically proximate.

Further, as noted earlier, the benefit and the desirability of active engagement in online communities are not certain to participants. As a result, participants are expected to look for a social norm that can guide them which behavior is appropriate and desirable. In that regard, what virtually and geographically proximate others are doing is expected to send more accurate signal to individuals because geographical proximity will add more commonality among participants such as cultural and subunit similarities. So, we expect that individuals will be more motivated to actively engage in online communities if their active virtual neighbors are also geographically proximate. Therefore, we hypothesize that:

> *H2*: Participants' tendency to herd to virtual neighbors becomes stronger if their virtual neighbors are also geographically proximate.

4.3. Empirical Method

4.3.1. Research Context and Data

The empirical context of this study is an organizational online knowledge community of a fortune 500 Information Technology consulting company. The online knowledge community is an electronic bulletin board where employees can voluntarily exchange technical knowledge in topics such as .Net, Java, Database, and Cobol. Employees exchange knowledge by posting questions and answers. All interactions are text-based and at the time of the study there were no gamification features (e.g., badges, thumbs up/down) embedded to the online community. Also, the company did not provide extrinsic rewards for contributions.

The company provided us demographic information of all individuals. The demographic information includes geographic location (at city level), job title, gender, age, job tenure, performance, and real name of each participant. Geographically, participants of the online community were distributed across 146 cities and 11 countries. The countries include United States, India, United Kingdom, France, Belgium, Switzerland, Netherland, Japan, China, Hong Kong, and Australia. Participants in United States were most geographically distributed (111 cities), followed by United Kingdom (18 cities) and India (8 cities).

Our data include detailed information of all messages posted to the internal online knowledge community during 1.5 years starting from April 2006. The data shows when each message was posted, by whom, and on which topic. Similar to other online communities in extra-organizational setting, there exist active users who post disproportionately large number of messages to the online community. We found that about half of the total messages were posted by approximately 20% of participants. Yet, one difference from extra-organizational setting is that active participants tend to post both questions and answers in a balanced portion. According to previous empirical studies (Wesler et al. 2007, Zhang et al. 2007), many participants in extra-organizational setting tend to take roles as either askers or answers. Further, we found no spam messages in our empirical setting. During our empirical period, 2,688 individuals contributed 76,279 messages to the online community.

4.3.2. Measures

4.3.2.1. Dependent variable

The dependent variable of this study is i's level of active engagement to the online community, $Contribution_i$. We operationalized the level of active engagement by the number of messages i contributes to the online community. During our empirical period, individuals contributed average 28.37 messages to the online community. The distribution was highly skewed because small number of very active participants contributed large number of posts. Due to the high skewness, we log-transformed our dependent variable using the following formula.

$$Contribution_i = \log\,(number\ of\ messages\ of\ i + 1) \qquad (1)$$

4.3.2.2. Independent variables

The independent variables of this study consist of two parts, variables for individual effects and variables for social effects.

<u>Variables for Individual Effect</u>

We assume that participant’s contribution decision will be in large part determined by their baseline propensity. To account for heterogeneity across individuals in terms of their baseline propensity, we incorporated several individual-level covariates into our model. Based on the data provided by the company, partial sociodemographic characteristics of each participant are included in our estimation model, which includes gender, job tenure, performance, innovativeness, and hierarchical status.

$Gender_i$ is 1 if i is female and 0 if male. Approximately 25% of all participants are female and the remaining participants are male. $Job\ Tenure_i$ is the number of years passed since i joined the company at the start time of the empirical period. Most of the participants had short job tenure with the company: less than 2 years at the company. The average job tenure is 1.49 years with the maximum job tenure of 13 years. The company provided us performance data of participants at the start of the empirical testing period. $Performance_i$ is i's performance at the start time of the empirical period. $Performance_i$ is measured as an ordered performance rank (1~4) of each participant, which is evaluated by i's supervisor. 4 is the highest performance level. The average performance of all participants is 3.1278. $Innovativeness_i$ is based on the first month i contributed to the online community. The earlier i enter the online community, we consider that i is more innovative. We reverse coded each participant's first online community entrance month in order to get $Innovativeness_i$. The value for $Innovativeness_i$ ranges from 1 to 17 with 17 indicates the highest level of innovativeness. The average value for $Innovativeness_i$ is 5.75.

$Status_i$ is hierarchical status of i. Hierarchical status of participants is measured based on their job titles. Job titles of the organization contain information about employees' hierarchical positions but no information on functional areas. For example, chief officer, manager, and junior associate are common job titles in the organization. Because the organization's hierarchy is structured approximately in three levels, we categorize participants into three groups (high, middle, and low) based on their job titles. Participants who have a job title that contains words such as chief, manager, senior, principal, chairman, or director were categorized as high hierarchical-level participants. Job titles with assistant, junior, or trainee were categorized as low hierarchical-level. Remaining job titles were classified as middle hierarchical-level. High hierarchical status participants are coded as 3, middle as 2, and low as 1. The mean value of $Status_i$ is 2.05.

Variables for Social Effect

Our central interest of this study is to model how their reference groups influence the contribution decision of individuals. In order to identify virtual and geographic neighbors of each participant, we created two adjacency matrices that contain proximity information among individuals in a virtual world and in a face-to-face world respectively.

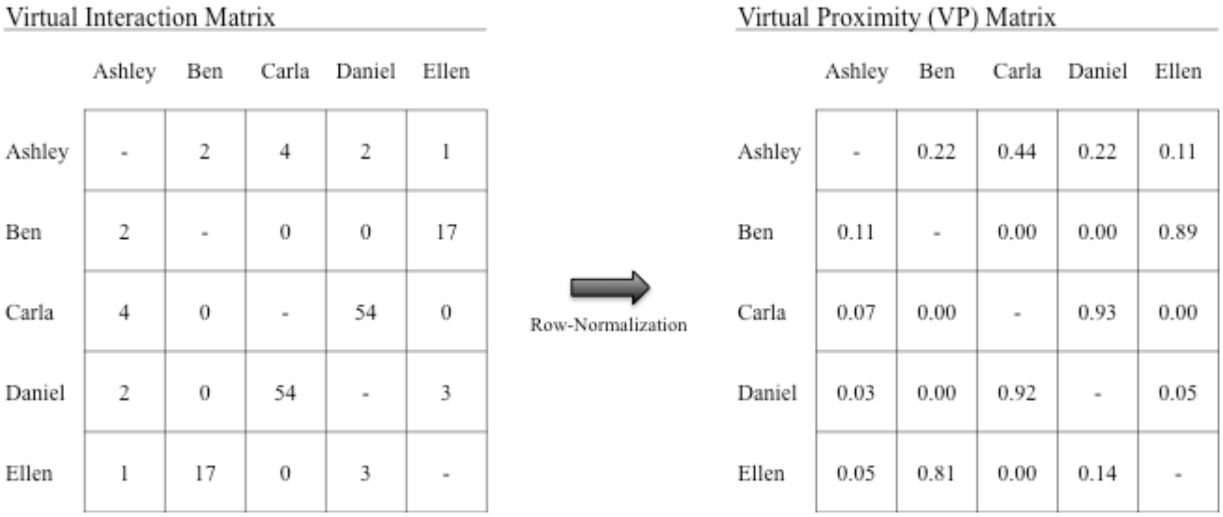
Virtual proximity (VP) The first matrix contains the degree of virtual proximity among individuals. We refer to the matrix as a Virtual Proximity matrix, VP for abbreviation. Individuals are considered to be virtually proximate if they have interacted in the online community. Suppose that Ashley posted a question to the online community and Ben responded to Ashley by posting an answer to Ashley's question. We consider that Ashley and Ben are virtual neighbors. We used unique thread ID to identify who answered to whose question at the online community. Our VP matrix is a person-to-person matrix with a size of 2,668 by 2,668. Each cell v_{ij} represents the extent to which *i* and *j* have interacted at the online community: the total number of interactions between *i* and *j*. The ties are non-directional and diagonal values are not an interest of this study. Because v_{ij} represents the number of interactions between *i* and *j,* it takes integer values. 0 means *i* and *j* have not interacted at the online community. The higher the value of $v_{ij},$ the more virtually proximate are *i* and *j*. After we constructed VP matrix, we normalized the matrix by row. Row normalization adjusts the influence from each alter to an ego based on how many interactions an ego has. Simply put, row normalization decreases the strength of influence each interaction exerts by the total number of interactions (Leenders 2002). Virtual proximity matrix, VP, is row normalized using the following formula:

$$vn_{ij} = \frac{v_{ij}}{v_{i.}} \qquad (2)$$

where $v_{i.} = \sum_j v_{ij}$, the i th row sum of VP. Simply put, $v_{i.}$ denotes the total number of i's interactions with all alters (i's virtual neighbors). With row normalization, the same weight is attached to every interaction of i, proportional to the total number of interactions by i. For example, if i has interacted 30 times, each interaction's influence to i will be weighted by $\frac{1}{30}$.

Figure 4.1. illustrates a hypothetical Virtual Proximity (VP) matrix with a size of 5 × 5.

FIGURE 4. 1. Hypothetical Virtual Proximity (VP) adjacency matrix



Virtual Interaction Matrix

	Ashley	Ben	Carla	Daniel	Ellen
Ashley	-	2	4	2	1
Ben	2	-	0	0	17
Carla	4	0	-	54	0
Daniel	2	0	54	-	3
Ellen	1	17	0	3	-

Row-Normalization

Virtual Proximity (VP) Matrix

	Ashley	Ben	Carla	Daniel	Ellen
Ashley	-	0.22	0.44	0.22	0.11
Ben	0.11	-	0.00	0.00	0.89
Carla	0.07	0.00	-	0.93	0.00
Daniel	0.03	0.00	0.92	-	0.05
Ellen	0.05	0.81	0.00	0.14	-

Geographic proximity (GP) The second adjacency matrix contains the degree of geographic proximity among individuals. It is constructed based on geographic distance among participants.

We refer to this matrix as Geographic Proximity matrix, GP for short. Our GP matrix is also a person-to-person matrix with a size of 2,668 by 2,668. To construct this matrix, we used geography information of each participant. Each participant's geographic location information is publicly available to all participants of the online community through each user's profile. Geographically, the online community participants were distributed across 146 cities in 11 countries[13] during our empirical testing period. Geographic distance is calculated at a city level location. In our dataset, a participant pair that is farthest apart is 11,632 miles away (Boston in US and Perth in Australia) and a participant pair that is located in the same city is considered to be the closest pair geographically (their distance is 0). Each cell g_{ij} in Geographical Distance (GD) matrix represents the extent to which *i* and *j* are distant geographically (in 1,000 miles). Similar to virtual ties (v_{ij}), geographical ties are non-directional and diagonal cells are not the interest of this study. Because we need geographical *proximity* matrix rather than geographical *distance* matrix, we reverse coded Geographical Distance (GD) matrix by using the following formula:

$$gp_{ij} = \max_{i,j \in \{1.2.\cdots,2668\}} g_{ij} + \min_{i,j \in \{1.2.\cdots,2668\}} g_{ij} - g_{ij} \qquad (3)$$

$$Let \quad GP = [gp_{ij}]$$

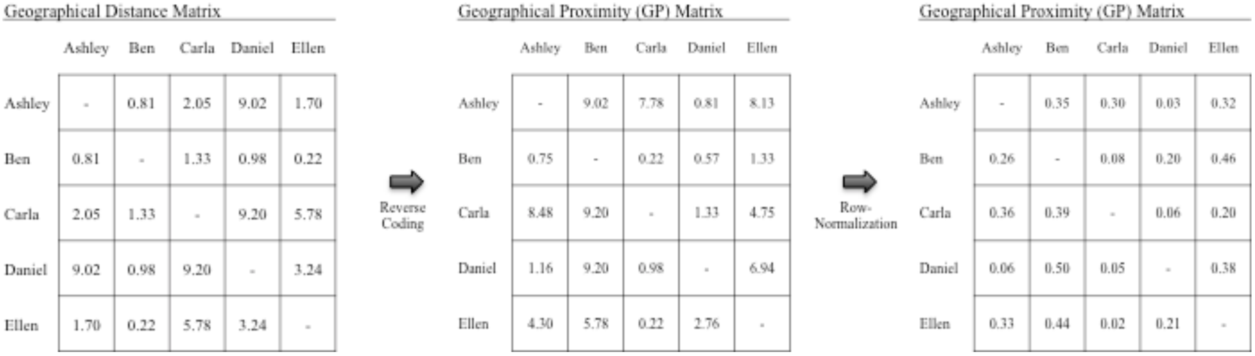
where gp_{ij} is the reverse-coded cell value. Then, we row-normalized the GP matrix in the same manner as we did with VP matrix. The cell value of row-normalized GP matrix is gn_{ij}.

$$gn_{ij} = \frac{gp_{ij}}{g_{i.}} \qquad (4)$$

[13] The countries include United States, India, United Kingdom, France, Belgium, Switzerland, Netherland, Japan, China, Hong Kong, and Australia.

where $g_{i.} = \sum_j gp_{ij}$, the i^{th} row sum of GP. Figure 4.2. illustrates a hypothetical Geographical Proximity (GP) matrix with a size of 5 × 5.

FIGURE 4. 2. Hypothetical Geographic Proximity (GP) adjacency matrix



Geographical Distance Matrix

	Ashley	Ben	Carla	Daniel	Ellen
Ashley	-	0.81	2.05	9.02	1.70
Ben	0.81	-	1.33	0.98	0.22
Carla	2.05	1.33	-	9.20	5.78
Daniel	9.02	0.98	9.20	-	3.24
Ellen	1.70	0.22	5.78	3.24	-

Reverse Coding

Geographical Proximity (GP) Matrix

	Ashley	Ben	Carla	Daniel	Ellen
Ashley	-	9.02	7.78	0.81	8.13
Ben	0.75	-	0.22	0.57	1.33
Carla	8.48	9.20	-	1.33	4.75
Daniel	1.16	9.20	0.98	-	6.94
Ellen	4.30	5.78	0.22	2.76	-

Row-Normalization

Geographical Proximity (GP) Matrix

	Ashley	Ben	Carla	Daniel	Ellen
Ashley	-	0.35	0.30	0.03	0.32
Ben	0.26	-	0.08	0.20	0.46
Carla	0.36	0.39	-	0.06	0.20
Daniel	0.06	0.50	0.05	-	0.38
Ellen	0.33	0.44	0.02	0.21	-

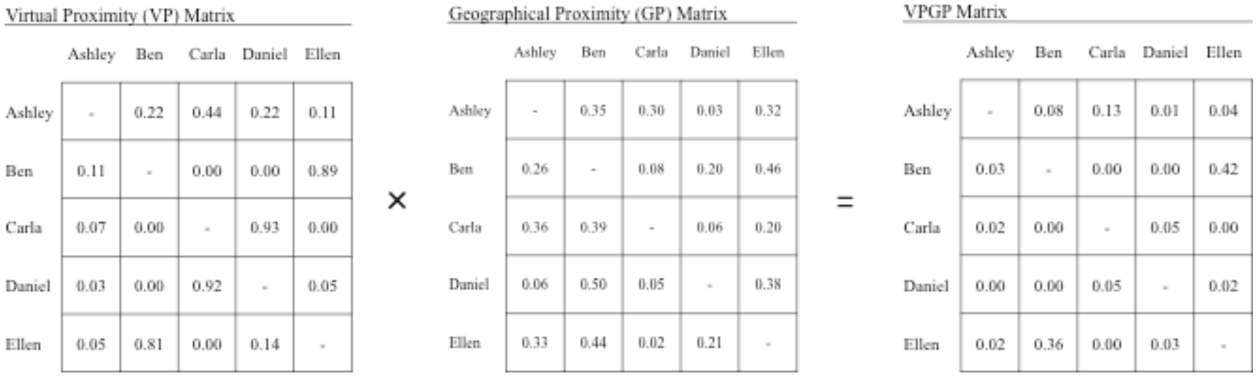
Virtual and Geographic proximity (VPGP) The second hypothesis posits that if an individual's virtual neighbor is also geographically proximate, the herding tendency will become stronger. In order to test the second hypothesis, we created another adjacency matrix that shows the interaction between virtual and geographic proximity among participants. We refer to the matrix as VPGP. We constructed VPGP matrix by multiplying VP and GP element-by-element:

$$Define\ vg_{ij} = vn_{ij} \times gn_{ij} \qquad i,j \in \{1,2,\cdots,2668\}$$

$$Let \quad VPGP = [vg_{ij}] \tag{5}$$

Figure 4.3. illustrates a hypothetical VPGP adjacency matrices with a size of 5x5.

FIGURE 4. 3. Hypothetical VP, GP, and VPGP adjacency matrices



Virtual Proximity (VP) Matrix

	Ashley	Ben	Carla	Daniel	Ellen
Ashley	-	0.22	0.44	0.22	0.11
Ben	0.11	-	0.00	0.00	0.89
Carla	0.07	0.00	-	0.93	0.00
Daniel	0.03	0.00	0.92	-	0.05
Ellen	0.05	0.81	0.00	0.14	-

×

Geographical Proximity (GP) Matrix

	Ashley	Ben	Carla	Daniel	Ellen
Ashley	-	0.35	0.30	0.03	0.32
Ben	0.26	-	0.08	0.20	0.46
Carla	0.36	0.39	-	0.06	0.20
Daniel	0.06	0.50	0.05	-	0.38
Ellen	0.33	0.44	0.02	0.21	-

=

VPGP Matrix

	Ashley	Ben	Carla	Daniel	Ellen
Ashley	-	0.08	0.13	0.01	0.04
Ben	0.03	-	0.00	0.00	0.42
Carla	0.02	0.00	-	0.05	0.00
Daniel	0.00	0.00	0.05	-	0.02
Ellen	0.02	0.36	0.00	0.03	-

4.3.3. Model Development: Network Autocorrelation Model

Our key objective is to examine how individuals' decision of contribution level is influenced by activities of their reference groups: virtual neighbors and geographic neighbors. To develop our estimation model, we use a network autocorrelation model. A network autocorrelation is an approach that network scholars typically take to model social influence (Doreian 1989, Leenders 2002). In the model, actors are assumed to be responsive to the contextual signals provided by their significant others' actions. We chose the model because it allows us to explicitly account for both individual effect (exogenous part) and social effect (autocorrelation part) on an individual's decision. Our basic network autocorrelation model for an individual *i*'s active engagement level decision is:

$$y_i \;= \boldsymbol{X_i}\mathbf{B} + \;\mathbf{P}\boldsymbol{W}y_j + \varepsilon_i \qquad (6)$$

where y_i is our dependent variable, the number of messages an individual *i* contributes to the online community. We assume that y_i is largely driven by *i*'s baseline propensity. $\boldsymbol{X_i}$ is a matrix that contains an intercept term and individual-specific covariates that determines *i*'s baseline propensity. Based on the data provided by the company, sociodemographic characteristics of each participant such as gender, job tenure, status, and performance are incorporated into the model. Although we had age data of participants, due to its high correlation with job tenure and status, we omit Age_i. Also, we incorporate innovativeness of each participant based on when a participant first joined the community. The earlier an employee joined the online community; we assume that an employee is more innovative. In sum, $\boldsymbol{X_i}$ is a (2668 × 6) matrix of values for the 2,668 individuals on 6 covariates. **B** is a (6 × 2668) matrix of parameters that we are estimating.

The central interest of this study is whether individuals adjust their baseline contribution propensity by activities of their reference groups. A network autocorrelation model incorporates this social effect through an autocorrelation term, $\mathbf{P}\boldsymbol{W}y_j$. Because every alter j of *i* is not expected to have the same influence to *i*, a set of weight matrices $\boldsymbol{W}$ specify the extent of influence each j exerts on *i*. By including the network autocorrelation term $\mathbf{P}\boldsymbol{W}y_j$, we are explicitly modeling that *i*'s decision, y_i, is related to a weighted combination of *i*'s significant others' decision, y_j. In this study, we are examining the influence of an individual's two different reference groups' (virtual neighbors and geographic neighbors) activities on his or her online community contribution. **P** is a matrix of estimated coefficients for the social effects. In order to examine social effects from the two reference groups and their interaction effects, we created three network autocorrelation terms by multiplying each weigh matrix with y. The resulting autocorrelation terms are $VPy, GPy,$ and $VPGPy$. VPy is the weighted combination of individual *i*'s virtually proximate others' contributions to the online community. For easier reference, we call VPy as Virtual Neighbors, GPy as Geographic Neighbors, and $GPy \times VPy$ as Virtual Neighbors × Geographic Neighbors. Herding tendency exists if **P** is statistically significant and positive. If $\mathbf{P}$ turns out to be 0, it suggests that there is no herding tendency of online community engagement: individuals do not take into account others' decision in determining one's participation level. The other extreme case is, $\mathbf{B} = 0$, where an individual's engagement decision is purely formed by others' decision.

The key issue in identifying social effect is the reflection problem (Manski 1993). Reflection problem refers to a difficulty in identifying a social effect of an actor's decision. According to the homophily principle, people tend to be attracted to similar others and be connected to similar others (McPherson et al. 2001). Because of this tendency, members of one's reference group are likely to be similar to the focal actor. As a result, it is problematic to claim that a focal actor changed one's behavior because his or her reference group changed their behavior, as it might be also true that they ended up with the same decision because of their similarity even if they made their decision independently. The estimation problem lies in that the factors that determine their similarities are not observable to researchers. For example, in our context, a focal individual and her virtual neighbor may stop contributing to the online community together not because a virtual neighbor's dormancy discouraged a focal actor (social effect) but because other unobservable factors (e.g., launch of an extra-organizational online community) discouraged both of them to contribute to the community.

Following Bramoullé et al. (2009), we attempted to mitigate this reflection problem by incorporating a fixed effect for each component of a virtual proximity matrix. A component of a graph is a sub-graph that is connected within but disconnected between sub-graphs (Hanneman and Riddle 2005). As suggested by the homophily principle (McPherson et al. 2001), people's connection is not random. By incorporating a fixed effect for a virtual matrix component, we control for unobservable similarities that may drive the same behavior of virtually connected participants independent of social effect. After incorporating the fixed effect, we eliminated this fixed effect by using differencing approach suggested by Bramoullé et al. (2009). We estimated the model using LNAM (Linear Network Autocorrelation Model) function in SNA (Social Network Analysis) package in R (Butts 2010). Estimation details are described in Butts (2008).

4.4. RESULTS

The descriptive statistics of the individual effect variables in our estimation model are reported in Table 1 and graph correlation of two weight matrices (Virtual Proximity matrix and Geographic Proximity matrix) is reported in Table 2.

TABLE 4. 1. Descriptive statistics and correlations of individual effect variables

	min.	max.	mean.	std.dev.	1	2	4	5	6	7
1 Contribution	1	1177	28.59	67.47	1.00					
2 Gender	0	1	0.25	0.43	0.01	1.00				
4 Job tenure	0	13	1.49	1.66	-0.13	-0.07	1.00			
5 Status	1	3	2.05	0.39	-0.03	-0.02	0.10	1.00		
6 Innovativeness	1	17	5.75	3.28	-0.09	-0.13	0.29	0.07	1.00	
7 Performance	1	4	3.12	0.68	-0.01	-0.05	0.05	0.02	0.06	1.00

Notes.

- No. of observations: 2,688
- Contribution is log-transformed after adding 1 to reduce heavy skewness

TABLE 4. 2. Correlation matrix of weight matrices

Weight matrices	1	2
1 Virtual Proximity (VP) matrix	1.0000	
2 Geographical Proximity (GP) matrix	0.0040	1.0000

Table 3 presents our main results: the parameter estimates of our network autocorrelation model. Effects are introduced across columns to demonstrate the stability of the results. Model

1 includes only individual effect variables. Two network autocorrelation terms, Virtual Neighbors (VPy) and Geographic Neighbors (GPy), are incorporated into Model 2. The two network autocorrelation terms are the key variables of our interest, which models social effects on individuals' contribution decision. Model 3 further incorporates a two-way interaction term between Virtual Neighbors and Geographic Neighbors ($VPy \times GPy$). Because the effects are consistent across columns, we will use Model 3 to discuss the results.

TABLE 4. 3. Main results: Parameter estimates of individual and social effects on contribution

		(1)	(2)	(3)
Social Effects (PWy)				
Virtual Neighbors (VPy)	ρ_1		0.520 ***	0.594 ***
			(0.031)	(0.034)
Geographic Neighbors (GPy)	ρ_2		0.840 ***	0.812 ***
			(0.062)	(0.062)
Virtual Neighbors (VPy) × Geographic Neighbors (GPy)	ρ_3			1.094 ***
				(0.195)
Individual Effects (XB)				
Gender	β_1	-0.042	-0.045	-0.044
		(0.045)	(0.044)	(0.044)
Job tenure	β_2	-0.072 ***	-0.067 ***	-0.066 ***
		(0.011)	(0.011)	(0.011)
Status	β_3	-0.064	-0.064	-0.065
		(0.047)	(0.044)	(0.044)
Innovativeness	β_4	-0.034 ***	-0.032 ***	-0.033 ***
		(0.007)	(0.007)	(0.007)
Performance	β_5	-0.014	-0.015	-0.014
		(0.028)	(0.027)	(0.027)
Intercept				
	β_0	0.461 ***	0.461 ***	0.467 ***
		(0.023)	(0.022)	(0.022)

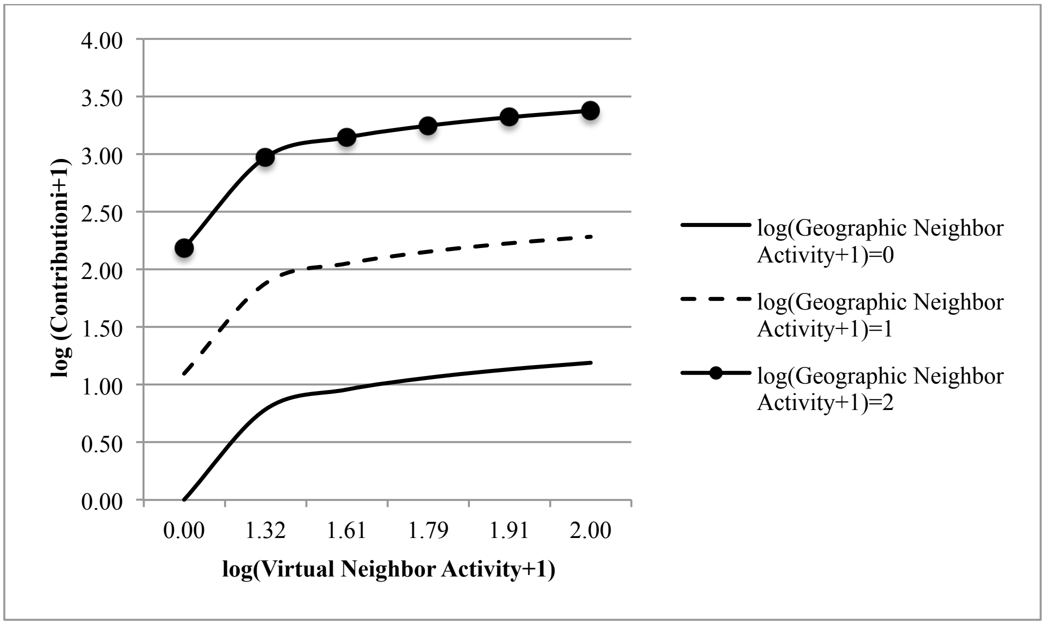
Notes.

- Dependent variable is *Contribution*$_i$
- *Contribution*$_i$ is log-transformed after adding 1 in order to reduce high skewness.
- No. of observations: 2,668
- Weight matrices VP and GP are mean-centered to reduce correlation with interaction terms.
- Job tenure, Status, Innovativeness, and Performance are mean-centered.
- Statistical significance *** $p < .001$
** $p < .05$
* $p < .1$

The first hypothesis predicts that individuals herd to their virtual neighbors. Our results support the first hypothesis: the parameter estimate of our first network autocorrelation term, Virtual Neighbors is positive and statistically significant (ρ_1 = 0.52, p < 0.001). It means that individuals are likely to contribute more if their virtual neighbors are active. Although we did not hypothesize the main effect of Geographic Neighbors on individuals' contribution, our results indicate that individuals are also likely to contribute more if their geographically proximate others are active. Further, our results show that the magnitude of the main effect of Geographic Neighbors is even larger than that of Virtual Neighbors. This suggests that individuals tend to herd more to geographic neighbors than virtual neighbors.

In the second hypothesis, we propose that participants' tendency to herd to virtual neighbors will become stronger if their virtual neighbors are also geographically proximate. The positive and significant coefficient of the two-way interaction term Virtual Neighbors × Geographic Neighbors ($\rho_3 = 1.094, p < 0.001$) supports our hypothesis. The positive interaction effect indicates that the social effect of virtual neighbors strengthens if virtual neighbors are also physically proximate. Figure 4.4 illustrates the interaction plot of this effect. It shows that individuals' tendency to herd to virtual neighbors strengthens if their virtual neighbors are also geographically proximate.

FIGURE 4. 4. Interaction plot: Virtual Neighbors x Geographic Neighbors



As noted earlier, we believe that an individual's contribution decision is in large part determined by one's baseline propensity. Based on the data provided by the company, we incorporated five individual-specific covariates as determinants of one's baseline propensity. Our results suggest that gender, hierarchical-status, and performance level do not affect the level of active engagement. The null effect of $Performance_i$ is disappointing in that companies would want high-performing employees to share quality knowledge in internal online communities. We further found that $Job\ tenure_i$ $(\beta_2 = -0.072, p < 0.001)$ negatively affects an individual's active engagement level. It suggests that employees who joined the company earlier (longer job tenure) are less likely to contribute to the online community. We speculate that those employees who have longer job tenure are likely to have higher level of face-to-face social capital to get resources, which lowers their motivation to go online. Also, our data show that those employees who have longer job tenure are likely to be old. So, the negative impact might be partially due to the generation gap. For younger generation employees, it is more natural to interact over the electronic network. Lastly, individuals who started using the online community earlier (early adopters) are found to contribute less to online community ($\beta_4 = -0.034, p < 0.001$).

As supplementary analyses we examined whether individuals' heterogeneity affects their herding tendency. We checked individual heterogeneity across five dimensions (i.e., gender, job tenure, status, innovativeness, and performance) by incorporating two-way interaction terms with network autocorrelation terms (Virtual Neighbors and Geographic Neighbors). The results are presented in Table 4. The results indicate that gender, job tenure, status, and performance heterogeneity of individuals do not affect their herding tendency. Yet, innovativeness of individuals is found to influence their herding tendency. Model 4 shows that individuals' tendency to herd to their virtual neighbors weakens as an individual is more innovative ($\rho_{10} = -0.053, p < 0.001$). On the other hand, it also shows that more innovative individuals tend to herd to geographic neighbors more ($\rho_{11} = 0.097, p < 0.001$). In sum, early adoptors have stronger herding tendency to their geographic neighbors but weaker herding tendency to virtual neighbors.

TABLE 4. 4. Supplemental results: Individual heterogeneity in herding tendency

		(4)	(5)	(6)	(7)	(8)
Social Effects (PWy)						
Virtual Neighbors (VPy)	ρ_1	0.575 ***	0.589 ***	0.595 ***	0.566 ***	0.595 ***
		(0.037)	(0.035)	(0.034)	(0.034)	(0.034)
Geographic Neighbors (GPy)	ρ_2	0.851 ***	0.817 ***	0.813 ***	0.763 ***	0.812 ***
		(0.070)	(0.063)	(0.062)	(0.067)	(0.062)
Virtual Neighbors (VPy) × Geographic Neighbors (GPy)	ρ_3	1.136 ***	1.131 ***	1.110 ***	1.784 ***	1.099 ***
		(0.197)	(0.202)	(0.196)	(0.255)	(0.195)
Gender Heterogeneity						
Virtual Neighbors (VPy) × Gender	ρ_4	0.084				
		(0.077)				
Geographic Neighbors (GPy) × Gender	ρ_5	-0.180				
		(0.157)				
Job tenure heterogeneity						
Virtual Neighbors (VPy) × Job tenure	ρ_6		-0.021			
			(0.022)			
Geographic Neighbors (GPy) × Job tenure	ρ_7		-0.006			
			(0.032)			
Status heterogeneity						
Virtual Neighbors (VPy) × Status	ρ_8			-0.063		
				(0.076)		
Geographic Neighbors (GPy) × Status	ρ_9			0.059		
				(0.131)		
Innovativeness heterogeneity						
Virtual Neighbors (VPy) × Innovativeness	ρ_{10}				-0.053 ***	
					(0.017)	
Geographic Neighbors (GPy) × Innovativeness	ρ_{11}				0.097 ***	
					(0.024)	
Performance heterogeneity						
Virtual Neighbors (VPy) × Performance	ρ_{12}					-0.047
						(0.046)
Geographic Neighbors (GPy) × Performance	ρ_{13}					0.080
						(0.092)
Individual Effects (XB)						
Gender	β_1	-0.048	-0.043	-0.043	-0.044	-0.043
		(0.044)	(0.044)	(0.044)	(0.044)	(0.044)
Job tenure	β_2	-0.067 ***	-0.067 ***	-0.066 ***	-0.065 ***	-0.066 ***
		(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Status	β_3	-0.065	-0.064	-0.068	-0.066	-0.066
		(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Innovativeness	β_4	-0.033 ***	-0.033 ***	-0.033 ***	-0.037 ***	-0.033 ***
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Performance	β_5	-0.014	-0.014	-0.014	-0.013	-0.013
		(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Intercept						
	β_0	0.468 ***	0.466 ***	0.047 ***	0.458 ***	0.467 ***
		(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Mean VIF		1.22	1.14	1.11	1.31	1.10

Notes.

- Dependent variable is *Contribution*$_i$
- *Contribution*$_i$ is log-transformed after adding 1 in order to reduce high skewness.
- No. of observations: 2,668
- Weight matrices VP and GP are mean-centered to reduce correlation with interaction terms.
- Job tenure, Status, Innovativeness, and Performance are mean-centered.
- Statistical significance *** p < .001

** p < .05
* p < .1

As a robustness check, we controlled for unobservable heterogeneity across individuals. In order to control for any time-invariant individual heterogeneity, we incorporated individual fixed effect, α_i, and conducted first differencing to eliminate it. The resulting equation is below:

$$y_{it} = \boldsymbol{X_{it}}\mathbf{B} + \mathbf{P}\boldsymbol{W}y_{jt} + \varepsilon_{it} + \alpha_i \qquad (7)$$

$$y_{it-1} = \boldsymbol{X_{it-1}}\mathbf{B} + \mathbf{P}\boldsymbol{W}y_{jt-1} + \varepsilon_{it-1} + \alpha_i \qquad (8)$$

Differencing both equations removes unobserved α_i and gives equation (9).

$$\Delta y_{it} = \Delta\boldsymbol{X_{it}}\mathbf{B} + \mathbf{P}\Delta\boldsymbol{W}y_{jt} + \Delta\varepsilon_{it} \qquad (9)$$

After eliminating unobserved individual heterogeneity, the results were consistent. We did not use this procedure for our main analysis because first differencing also eliminates all of our time-invariant independent variables (i.e., gender, status, job tenure, performance, innovativeness).

4.5. Discussions and Conclusion

This paper examined membership dynamics in online knowledge communities. Specifically, we explored whether individuals tend to herd to other participants when they decide how much to contribute to an online knowledge community. We proposed that participants would be motivated to contribute more if their virtual neighbors, with whom they have interacted over an online community, are active. In addition, we proposed that this herding tendency would become stronger if their virtual neighbors are also geographically proximate to them. Consistent with our predictions, in the context of an internal online knowledge community of a large IT consulting company, we found that participants tend to herd to their virtual neighbors. Also we found that this herding tendency becomes stronger if their virtual neighbors are also geographically proximate.

Our study advances knowledge in several areas. First, it contributes to literature on online communities. Our ego-network level approach offers complementary insights to previous studies that examined participation motivations at the individual level (e.g., Bagozzi and Dholakia 2006, Ma and Agarwal 2007, Ren et al. 2007, von Hippel and von Krogh 2003, Wasko and Faraj 2005), at a dyad-level (Hwang et al. forthcoming), and at a community level (Butler 2001, Faraj and Johnson 2011). Humans have intrinsic propensity for consensus: people want to follow what others do. Scholars have documented extensive evidence of human herding behavior in diverse settings (Bikhchandani et al. 1998, Scharfstein and Stein 1990, Zhang 2010). Despite its prevalence, we had limited understanding on how this herding tendency affects individuals' contribution decision to online communities. This study extends prior works by exploring herding tendencies that drive individuals' contribution to online communities.

Second, it has been argued that the Internet will make the world "flat" (Friedman 2006) because the Internet brought possibilities to easily connect people over distance. Propinquity, geographical proximity, seemed to lose its fundamental role in determining human interactions, at least in virtual setting. This is particularly expected in the case where the goal of an electronic network is to facilitate meetings among strangers based on their common interest other than geography. The empirical setting of this study is an internal online knowledge community where sharing knowledge among unacquainted others is the main objective. In this setting, propinquity is not expected to and also not desirable to drive connections. However, our findings suggest that propinquity is still influencing membership dynamics because participants regard other participants who are physically proximate more importantly.

The findings of our study can be used to improve design features of an online community. The insights from the previous studies on individual-level motivation guided design improvements of an online community. For example, various gamification functions (i.e., badges and reputation points) are employed based on the findings that reputation enhancing is an important driver for contributions. Badges and points make an individual's reputation more salient and long lasting in online communities. We expect that community features focusing on this herding tendency, which make other participants' activities more salient, would increase an individual's motivation to contribute.

We suggest directions for future research. In this study, we measured virtual proximity among participants only based on direct interactions among participants in the online community. However, participants are also likely to observe others' activities with whom they have not interacted directly. Incorporating these indirect relationships to measure virtual proximity may bring further interesting insights. Also, participants' tendency to herd is expected to influence evolution of online community population. A simulation study to examine this dynamics is expected to provide valuable insight on the fluid nature of online community.

5. References

Adamic, L. A., J. Zhang, E. Bakshy, M. S. Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows something. *Proceeding of the 17th international conference on World Wide Web*. 665–674.

Agrawal, A., C. Catalini, A. Goldfarb. 2011. Friends, family, and the flat world: The geography of crowdfunding. *NBER Working Paper* **16820**.

Alavi, M., D. E. Leidner. 2001. Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly* 107–136.

Allen, T. 1977. The technological gatekeeper.

Van Alstyne, M. W., E. Brynjolfsson. 2005. Global village or CyberBalkans: Modeling and measuring the integration of electronic communities. *Management Science* **51**(6) 851–868.

Amir, Y. 1969. Contact hypothesis in ethnic relations. *Psychological Bulletin* **71**(5) 319.

Aral, S., M. Van Alstyne. 2011. The Diversity-Bandwidth Trade-off. *American Journal of Sociology* **117**(1) 90–171.

Argote, L. 2012. *Organizational learning: Creating, retaining and transferring knowledge*. Springer.

Arguello, J., B. S. Butler, E. Joyce, R. Kraut, K. S. Ling, C. Rosé, X. Wang. 2006. Talk to me: foundations for successful individual-group interactions in online communities.

Proceedings of the SIGCHI conference on Human Factors in computing systems. ACM, 959–968.

Atchadé, Y. F. 2006. An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in applied Probability* **8**(2) 235–254.

Bagozzi, R. P., U. M. Dholakia. 2006. Open source software user communities: A study of participation in Linux user groups. *Management Science* **52**(7) 1099–1115.

Baker, N., J. Freeland. 1972. Structuring information flow to enhance innovation. *Management Science* **19**(1) 105–116.

Baker, N. R., J. Siegman, A. H. Rubenstein. 1967. The effects of perceived needs and means on the generation of ideas for industrial research and development projects. *Engineering Management, IEEE Transactions on* (4) 156–163.

Banerjee, A. V. 1992. A simple model of herd behavior. *The Quarterly Journal of Economics* 797–817.

Bayus, B. L. 2013. Crowdsourcing New Product Ideas over Time: An Analysis of the Dell IdeaStorm Community. *Management Science* **59**(1) 226–244.

Bikhchandani, S., D. Hirshleifer, I. Welch. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives* 151–170.

Bisgin, H., N. Agarwal, X. Xu. 2012. A study of homophily on social media. *World Wide Web* **15**(2) 213–232.

Bista, S. K., S. Nepal, N. Colineau, C. Paris. 2012. Using gamification in an online community. *CollaborateCom*. 611–618.

Bock, G.-W., R. W. Zmud, Y.-G. Kim, J.-N. Lee. 2005. Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. *MIS quarterly* 87–111.

Borgatti, S. P., R. Cross. 2003. A relational view of information seeking and learning in social networks. *Management Science* **49**(4) 432–445.

Bradner, E., G. Mark. 2002. Why distance matters: effects on cooperation, persuasion and deception. *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. 226–235.

Bramoullé, Y., H. Djebbari, B. Fortin. 2009. Identification of peer effects through social networks. *Journal of Econometrics* **150**(1) 41–55.

Brass, D. J. 1995. Creativity: It's all in your social network. *Creative Action in Organizations* **94** 99.

Brass, D. J., J. Galaskiewicz, H. R. Greve, W. Tsai. 2004. Taking stock of networks and organizations: A multilevel perspective. *Academy of Management Journal* **47**(6) 795–817.

Braun, M., A. Bonfrer. 2009. *Censoring, Interdependence and Scalability for Dyadic Social Media Data*. Massachusetts Institute of Technology.

Brewer, M. B., R. J. Brown. 1998. *Intergroup relations*. McGraw-Hill.

Bunderson, S. 2003. Recognizing and Utilizing Expertise in Work Groups: A Status Characteristics Perspective. *Administrative Science Quarterly* **48**(4) 557–591.

Burke, M., E. Joyce, T. Kim, V. Anand, R. Kraut. 2007. Introductions and requests: Rhetorical strategies that elicit response in online communities. *Communities and Technologies 2007*. Springer, 21–39.

Burt, R. S. 1992. *Structural holes*. Harvard Univ. Press.

Butler, B. 2001. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information Systems Research* **12**(4) 346–362.

Butts, C. 2010. Package "sna."

Butts, C. T. 2008. Social network analysis with sna. *Journal of Statistical Software* **24**(6) 1–51.

Byrne, D., G. L. Clore Jr, P. Worchel. 1966. Effect of economic similarity-dissimilarity on interpersonal attraction. *Journal of Personality and Social Psychology* **4**(2) 220.

Byrne, D. E. 1971. *The attraction paradigm*. Academic Pr.

Byrne, D., D. Nelson. 1964. Attraction as a function of attitude similarity-dissimilarity: The effect of topic importance. *Psychonomic Science* **1**(1-12) 93–94.

Byrne, D., T. J. Wong. 1962. Racial prejudice, interpersonal attraction, and assumed dissimilarity of attitudes. *The Journal of Abnormal and Social Psychology* **65**(4) 246.

Cairncross, F. 2001. *The death of distance: How the communications revolution is changing our lives*. Harvard Business Press.

Cameron, A., P. Trivedi. 2005. *Microeconometrics: methods and applications* 1st ed. Cambridge University Press.

Casakin, H. 2004. Visual analogy as a cognitive strategy in the design process: Expert versus novice performance. *Journal of Design Research* **4**(2) 124.

Chase, W. G., H. A. Simon. 1973. Perception in chess. *Cognitive Psychology* **4**(1) 55–81.

Clark, H. H., C. R. Marshall. 2002. Definite reference and mutual knowledge. *Psycholinguistics: Critical Concepts in Psychology* **414**.

Cohen, B. P., X. Zhou. 1991. Status processes in enduring work groups. *American Sociological Review* 179–188.

Collins, A. Burstein. 1989. After word: a framework for a theory of comparison and mapping. *Similarity and analogical reasoning*. Cambridge, MA, Cambridge University Press.

Constant, D., L. Sproull, S. Kiesler. 1996. The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization Science* **7**(2) 119–135.

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

Crandall, D., D. Cosley, D. Huttenlocher, J. Kleinberg, S. Suri. 2008. Feedback effects between similarity and social influence in online communities. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 160–168.

Cross, R., L. Sproull. 2004. More than an answer: Information relationships for actionable knowledge. *Organization Science* **15**(4) 446–462.

Dahl, Moreau. 2002. The influence and value of analogical thinking during new product ideation. *Journal of Marketing Research* **39**(1) 47–60.

Dahlander, L., L. Frederiksen. 2012. The Core and Cosmopolitans: A Relational View of Innovation in User Communities. *Organization Science* **23**(4) 988–1007.

Deterding, S., D. Dixon, R. Khaled, L. Nacke. 2011. From game design elements to gamefulness: defining gamification. *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. ACM, 9–15.

Doreian, P. 1989. Models of network effects on social actors. *Research methods in social network analysis* 295–317.

Dugosh, Paulus, Roland, Yang. 2000. Cognitive stimulation in brainstorming. *Journal of Personality and Social Psychology* **79**(5) 722–735.

Dunbar, K. 1995. How scientists really reason: Scientific reasoning in real-world laboratories. *The Nature of Insight*. Cambridge, MA, MIT Press, 65–395.

Eckert, C. M., M. K. Stacey, J. Wiley. 1999. Expertise and designer burnout. *Proceedings of the 12th International Conference on Engineering Design*. 195–200.

Ettlie, J. E., W. P. Bridges, R. D. O'keefe. 1984. Organization strategy and structural differences for radical versus incremental innovation. *Management science* **30**(6) 682–695.

Faraj, S., S. Javenpaa, A. Majchrzak. 2011. Knowledge collaboration in online communities. *Organization Science*.

Faraj, S., S. L. Johnson. 2011. Network exchange patterns in online communities. *Organization Science* **22**(6) 1464–1480.

Festinger, L., S. Schachter, K. Back. 1950. *Social Processes in Informal Groups*. Stanford, CA, Stanford Univ. Press.

Fisher, C. 1982. *To Dwell among Friends*. Chicago, Univ. Chicago Press.

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

Friedman, T. 2007. *The World Is Flat 3.0: A Brief History of the Twenty-first Centry*. New York, Picador/Farrar, Straus and Giroux.

Friedman, T. L. 2006. *The world is flat [updated and expanded]: A brief history of the twenty-first century*. Macmillan.

Fulk, J., R. Heino, A. J. Flanagin, P. R. Monge, F. Bar. 2004. A Test of the Individual Action Model for Organizational Information Commons. *Organization Science* **15**(5) 569–585.

Fussell, S. R., R. M. Krauss. 1992. Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology* **62**(3) 378.

Galegher, J. R., R. E. Kraut, C. Egido. 1990. *Intellectual teamwork: The social and technological bases of cooperative work*. L. Erlbaum Associates Inc.

Garcia, D., P. Mavrodiev, F. Schweitzer. 2013. Social resilience in online communities: The autopsy of friendster. *Proceedings of the first ACM conference on Online social networks*. ACM, 39–50.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Rubin. 2003. *Bayesian Data Analysis*. Boca Raton, FL, Chapman and Hall/CRC.

Gelman, A., J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

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

Glaser, R. 1989. Expertise and learning: How do we think about instructional processes now that we have discovered knowledge structures. *Complex information processing: The impact of Herbert A. Simon*. 269–282.

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

Grant, R. M. 1996. Toward a knowledge-based theory of the firm. *Strategic Management Journal* **17**(10) 109–122.

Greene, W. 2011. *Econometric Analysis* 7th ed. Prentice Hall.

Hampton, Wellman. 2000. Examining community in the digital neighborhood: Early results from Canada's wired suburb. *Digital Cities: Technologies, Experiences, and Future Perspectives*. Heidelberg, Springer-Verlag, 194–208.

Hanneman, R. A., M. Riddle. 2005. *Introduction to social network methods*. University of California Riverside.

Hansen, M. T. 1999. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly* **44**(1) 82–85.

Hargadon, A., R. Sutton. 1997. Technology brokering and innovation in a product development firm. *Administrative Science Quarterly* **42**(4) 716–749.

Harrison, D. A., K. H. Price, J. H. Gavin, A. T. Florey. 2002. Time, teams, and task performance: Changing effects of surface-and deep-level diversity on group functioning. *Academy of Management Journal* 1029–1045.

Harrison, D., K. Price, M. Bell. 1998. Beyond relational demography: Time and the effects of surface- and deep-level diversity on work group cohesion. *Academy of Management Journal* **41**(1) 96–107.

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

Hayes. 1989. Cognitive processes in creativity. *Handbook of creativity*. New York, Plenum, 135–145.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica* 153–161.

Henkel, J., E. Von Hippel. 2005. *Welfare implications of user innovation*. Springer.

Hewstone, M., M. Rubin, H. Willis. 2002. Intergroup bias. *Annual review of psychology* **53**(1) 575–604.

Von Hippel, E. 1986. Lead Users: A Source of Novel Product Concepts. *Management Science* **32**(7) 791–805.

Von Hippel, E. 1988. *The sources of innovation*. Oxford University Press New York.

Von Hippel, E., G. von Krogh. 2003. Open source software and the “private-collective” innovation model: Issues for organization science. *Organization Science* **14**(2) 209–223.

Hollingshead, A. B., J. E. McGrath. 1995. Computer-assisted groups: A critical review of the empirical research. *Team Effectiveness and Decision Making in Organizations* 46–78.

Horton, W., B. Keysar. 1996. When do speakers take into account common ground? *Cognition* **59** 91–117.

Howe, J. 2008. *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Huang, Y., P. Vir Singh, K. Srinivasan. 2014. Crowdsourcing New Product Ideas Under Consumer Learning. *Management Science* 140702074200008.

Ibarra, H. 1993. Personal networks of women and minorities in management: A conceptual framework. *Academy of management Review* **18**(1) 56–87.

Jaccard, J., R. Turrisi. 2003. *Interaction effects in multiple regression*. Thousand oaks, CA, Sage Publications, Inc.

Jackson, May, Whitney. 1995. Understanding the dynamics of diversity in decision-making teams. *Team decision-making effectiveness in organizations*. R.A.Guzzo & E.Salas (Eds.), 204–261.

Jaffe, A. B. 1986. Technological opportunity and spillovers of R & D: evidence from firms' patents, profits, and market value. *American Economic Review* **76**(5) 984–1001.

Jansson, Smith. 1991. Design fixation. *Design Studies* **12**(1) 3–11.

Jehn, K., G. Northcraft, M. Neale. 1999. Why differences make a difference: A field study of diversity, conflict, and performance in workgroups. *Administrative Science Quarterly* **44** 741–763.

Jeppesen, L. B., L. Frederiksen. 2006. Why do users contribute to firm-hosted user communities? The case of computer-controlled music instruments. *Organization Science* **17**(1) 45–63.

Jeppesen, L. B., K. R. Lakhani. 2010. Marginality and problem-solving effectiveness in broadcast search. *Organization Science* **21**(5) 1016–1033.

Kiesler, S., J. Siegel, T. W. McGuire. 1984. Social psychological aspects of computer-mediated communication. *American Psychologist* **39**(10) 1123.

Krauss, R. M., S. Fussell. 1990. Mutual knowledge and communicative effectiveness. *Intellectual teamwork: Social and technological foundations of cooperative work*. Lawrence Erlbaum, Hillsdale, NJ, 111–145.

Kraut, R., T. Mukhopadhyay, J. Szczypula, S. Kiesler, B. Scherlis. 1999. Information and communication: Alternative uses of the internet in households. *Information Systems Research* **10**(4) 287–303.

Von Krogh, G., S. Spaeth, K. R. Lakhani. 2003. Community, joining, and specialization in open source software innovation: a case study. *Research Policy* **32**(7) 1217–1241.

Kulkarni, D., H. A. Simon. 1988. The processes of scientific discovery: The strategy of experimentation. *Cognitive science* **12**(2) 139–175.

Lakhani, K. 2009. InnoCentive (A).

Lakhani, K. R., E. Von Hippel. 2003. How open source software works:"free" user-to-user assistance. *Research Policy* **32**(6) 923–943.

Lakhani, K., R. Wolf. 2005. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Perspectives on Free and Open Source Software*. MIT Press.

Lancaster, F. W. 1978. *Toward paperless information systems*. Academic Press, Inc.

Latané, B. 1996. Dynamic social impact: The creation of culture by communication. *Journal of Communication* **46**(4) 13–25.

Laursen, K., A. Salter. 2006. Open for innovation: the role of openness in explaining innovation performance among U.K. manufacturing firms. *Strategic Management Journal* **27**(2) 131–150.

Lauw, H. W., J. C. Shafer, R. Agrawal, A. Ntoulas. 2010. Homophily in the digital world: A LiveJournal case study. *Internet Computing, IEEE* **14**(2) 15–23.

Lea, M., R. Spears. 1991. Computer-mediated communication, de-individuation and group decision-making. *International Journal of Man-Machine Studies* **34**(2) 283–301.

Lee, Y. J., K. Hosanagar, Y. Tan. Forthcoming. Do I follow my friends or the crowd? Information cascades in online movie ratings. *Management Science*.

Leenders, R. T. A. 2002. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks* **24**(1) 21–47.

Lerner, J., J. Tirole. 2002. Some simple economics of open source. *The Journal of Industrial Economics* **50**(2) 197–234.

Linstone, H. A., M. Turoff. 1976. *The Delphi method: techniques and applications*. Addison-Wesley.

Lu, Y., K. Jerath, P. V. Singh. 2013. The Emergence of Opinion Leaders in a Networked Online Community: A Dyadic Model with Time Dynamics and a Heuristic for Fast Estimation. *Management Science* **59**(8) 1783–1799.

Ludford, P. J., D. Cosley, D. Frankowski, L. Terveen. 2004. Think different: increasing online community participation using uniqueness and group dissimilarity. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 631–638.

Ma, M., R. Agarwal. 2007. Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities. *Information Systems Research* **18**(1) 42–67.

MacMillan, D. 2011. “Gamification”: A Growing Business to Invigorate Stale Websites. *BloombergView*.

Makela, K., H. K. Kalla, R. Piekkari. 2007. Interpersonal similarity as a driver of knowledge sharing within multinational corporations. *International Business Review* **16**(1) 1–22.

March, J. G. 1991. Exploration and exploitation in organizational learning. *Organization Science* **2**(1) 71–87.

Markus, L. M. 2001. Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of management information systems* **18**(1) 57–93.

Marquaridt, D. W. 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* **12**(3) 591–612.

Martin, X., W. Mitchell. 1998. The influence of local search and performance heuristics on new design introduction in a new product market. *Research Policy* **26**(7) 753–771.

Mason, C. H., W. D. Perreault Jr. 1991. Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research* 268–280.

McCallum, A. K. 2002. *MALLET: A Machine Learning for Language Toolkit*.

McKinsey. 2013. *Evolution of the networked enterprise: McKinsey Global Survey Results*.

McPherson, J. M., L. Smith-Lovin. 1987. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*.

McPherson, M. 1983. An ecology of affiliation. *American Sociological Review* **48**(4) 519–532.

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

Mislove, A., B. Viswanath, K. P. Gummadi, P. Druschel. 2010. You are who you know: inferring user profiles in online social networks. *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 251–260.

Monteiro, L. F., N. Arvidsson, J. Birkinshaw. 2008. Knowledge Flows Within Multinational Corporations: Explaining Subsidiary Isolation and Its Performance Implications. *Organization Science* **19**(1) 90–107.

Moore, G. 1990. Structural determinants of men's and women's personal networks. *American Sociological Review* 726–735.

Moreland, R. L., J. M. Levine. 1982. Socialization in small groups: Temporal changes in individual-group relations. *Advances in experimental social psychology* **15** 137–192.

Nelson, R. R., S. G. Winter. 1982. *Evolutionay Theory of Economic Change*. Cambridge, MA, Belknap Press.

Newell, A., H. A. Simon. 1972. *Human problem solving*. Englewood Cliffs, NJ, Prentice-Hall.

Nonnecke, B., J. Preece. 2000. Lurker demographics: Counting the silent. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 73–80.

Novick, L. R. 1988. Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14**(3) 510.

O'Brien, R. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity* **41**(5) 673–690.

O'Reilly, C. A. 1983. The use of information in organizational decision making: A model and some propositions. *Research in Organizational Behavior*.

Oh, W., S. Jeon. 2007. Membership herding and network stability in the open source community: The ising perspective. *Management Science* **53**(7) 1086–1101.

Oreg, S., O. Nov. 2008. Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values. *Computers in Human Behavior* **24**(5) 2055–2073.

Orlikowski, W. J. 1996. Improvising organizational transformation over time: A situated change perspective. *Information Systems Research* **7**(1) 63–92.

Pal, A., R. Farzan, J. Konstan, R. Kraut. 2011. Early detection of potential experts in question answering communities. *User Modeling, Adaption and Personalization* 231–242.

Pelled, L. 1996. Demographic Diversity, Conflict, and Work Group Outcomes: An Intervening process theory. *Organization Science* **7**(6) 615–631.

Reicher, S. D. 1984. Social influence in the crowd: Attitudinal and behavioural effects of de-individuation in conditions of high and low group salience. *British Journal of Social Psychology* **23**(4) 341–350.

Ren, Y., R. Kraut, S. Kiesler. 2007. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies* **28**(3) 377–408.

Riahi, F., Z. Zolaktaf, M. Shafiei, E. Milios. 2012. Finding expert users in community question answering. *Proceedings of the 21st international conference companion on World Wide Web*. 791–798.

Rice, R. E. 1982. Communication networking in computer-conferencing systems: A longitudinal study of group roles and system structure. *Communication yearbook* **6** 925–944.

Rosenberg. 1982. Learning by using. *Inside the Black Box: Technology and Econoimcs*. Cambridge, UK, Cambridge University Press, 120–140.

Salganik, M. J., P. S. Dodds, D. J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**(5762) 854–856.

Sampson, R. C. 2007. R&D alliances and firm performance: the impact of technological diversity and alliance organization on innovation. *Academy of Management Journal* **50**(2) 364.

Scharfstein, D. S., J. C. Stein. 1990. Herd behavior and investment. *The American Economic Review* 465–479.

Schlosser, A. E. 2005. Posting versus Lurking: Communicating in a Multiple Audience Context. *Journal of Consumer Research* **32**(2) 260–265.

Schumpeter, J. A. 1939. *Business cycles*. Cambridge Univ Press.

Shah, C. 2010. Collaborative information seeking: A literature review. *Advances in Librarianship* **32** 3–33.

Shrum, W., N. H. Cheek Jr, S. MacD. 1988. Friendship in school: Gender and racial homophily. *Sociology of Education* 227–239.

Siegel, J., V. Dubrovsky, S. Kiesler, T. McGuire. 1986. Group processes in computer-mediated communication. *Organizational Behavior and Human Decision Processes* **37** 157–187.

Simon, H. A. 1981. *The sciences of the artificial*. MIT press.

Simpson, W. 2001. The Quadratic Assignment Procedure (QAP).

Singh, R., L. S. Tan. 1992. Attitudes and attraction: A test of the similarity-attraction and dissimilarity-repulsion hypotheses. *British Journal of Social Psychology* **31**(3) 227–238.

Skopek, J., F. Schulz, H.-P. Blossfeld. 2011. Who Contacts Whom? Educational Homophily in Online Mate Selection. *European Sociological Review* **27**(2) 180–195.

Smith. 2003. The constraining effects of initial ideas. *Group Creativity: Innovation Through Collaboration*. Oxford, UK, Oxford University Press, 15–31.

Smith, Ward, Schumacher. 1993. Constraining effects of examples in a creative generation task. *Memory and Cognition* **21**(6) 837–845.

Sproull, L., S. Kiesler. 1986. Reducing social context cues: Electronic mail in organizational communications. *Management Science* **32**(11) 1492–1512.

Sproull, L., S. Kiesler. 1991. New ways of working in the networked organization. *Cambridge, MA*.

Stangor, C., L. Lynch, C. Duan, B. Glas. 1992. Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology* **62**(2) 207.

Sternberg, R. J., T. I. Lubart. 1995. *Defying the crowd: Cultivating creativity in a culture of conformity*. New York, Free Press, New York.

Stewart, D. 2005. Social status in an open-source community. *American Sociological Review* **70**(5) 823–842.

Tajfel, H., M. G. Billig, R. P. Bundy, C. Flament. 1971. Social categorization and intergroup behaviour. *European Journal of Social Psychology* **1**(2) 149–178.

Takhteyev, Y., A. Gruzd, B. Wellman. 2012. Geography of Twitter networks. *Social Networks* **34**(1) 73–81.

Tausczik, Y. R., J. W. Pennebaker. 2012. Participation in an online mathematics community: differentiating motivations to add. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 207–216.

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

Thelwall, M. 2009. Homophily in MySpace. *Journal of the American Society for Information Science and Technology* **60**(2) 219–231.

Urban, G., E. von Hippel. 1988. Lead user analyses for the development of new industrial products. *Management Science* **34**(5) 569–582.

Usher, A. P. 1954. *A History of Mechanical Inventions*. Cambridge, MA, Courier Dover Publications.

von Hippel. 1976. The dominant role of users in scientific instrument innovaton process. *Research Policy* **5**(3) 212–239.

Vosniadou. 1989. Analogical reasoning as a mechanism in knowledge acquisition: a developmental perspective. *Similarity and analogical reasoning*. Cambridge, MA, Cambridge University Press, 413–437.

Wang, X., B. S. Butler, Y. Ren. 2013. The Impact of Membership Overlap on Growth: An Ecological Competition View of Online Groups. *Organization Science* **24**(2) 414–431.

Wasko, M. M. L., S. Faraj. 2005. Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly* 35–57.

Weisberg, R. W. 1999. Creativity and Knowledge: A Challenge to Theories. *Handbook of creativity* 226.

Welser, H. T., E. Gleave, D. Fisher, M. Smith. 2007. Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure* **8**(2) 1–32.

Whittaker, S., L. Terveen, W. Hill, L. Cherny. The dynamics of mass interaction.

Williams, K. Y., C. A. O'Reilly. 1998. Demography and diversity in organizations: A review of 40 years of research. *Research in Organizational Behavior* **20** 77–140.

Wittenbaum, G. M., G. Stasser. 1996. Management of information in small groups.

Zammuto, R. F., T. L. Griffith, A. Majchrzak, D. J. Dougherty, S. Faraj. 2007. Information Technology and the Changing Fabric of Organization. *Organization Science* **18**(5) 749–762.

Zander, U., B. Kogut. 1995. Knowledge and the speed of the transfer and imitation of organizational capabilities: An empirical test. *Organization Science* **6**(1) 76–92.

Zhang, J. 2010. The sound of silence: Observational learning in the US kidney market. *Marketing Science* **29**(2) 315–335.

Zhang, J., M. S. Ackerman, L. Adamic. 2007. Expertise networks in online communities: structure and algorithms. *Proceedings of the 16th international conference on World Wide Web*. 221–230.

Zhang, J., P. Liu. 2012. Rational Herding in Microloan Markets. *Management Science* **58**(5) 892–912.

Zipf, G. 1949. *Human Behavior and the Principle of Least Effort*. Menlo Park, CA, Addison-Wesley.

6. Technical Appendix for Chapter 2

Parameter Estimation Procedure

For the procedures below, letters with superscript u represent the values of the updated corresponding parameters.

Step 1: Estimating $\boldsymbol{\gamma}$ **(**$\boldsymbol{\gamma}$ represents homogeneous coefficients**)**

$$\boldsymbol{\gamma}^u | a_i, b_i, \alpha_0, \alpha_1, d_{ij}, data$$

$$f\left(\boldsymbol{\gamma}^u \middle| a_i, b_i, \alpha_0, \alpha_1, d_{ij}, data\right)$$

$$\propto \left|\Sigma_{\boldsymbol{\gamma}\mathbf{0}}\right|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\gamma}^u - \overline{\boldsymbol{\gamma}_\mathbf{0}})'\Sigma_{\boldsymbol{\gamma}0}^{-1}(\boldsymbol{\gamma}^u - \overline{\boldsymbol{\gamma}_\mathbf{0}})\right] L(\boldsymbol{Y})$$

where $\overline{\boldsymbol{\gamma}_\mathbf{0}}$ and $\Sigma_{\boldsymbol{\gamma}\mathbf{0}}$ are diffused priors. Because there is no closed form for this, we use the Metropolis-Hastings algorithm to draw from this conditional distribution of $\boldsymbol{\gamma}^u$. The probability of accepting $\boldsymbol{\gamma}^u$ is:

$$\Pr(acceptance) = \min\{\frac{\exp\left[-\frac{1}{2}(\boldsymbol{\gamma}^u - \overline{\boldsymbol{\gamma}_\mathbf{0}})'\Sigma_{\boldsymbol{\gamma}0}^{-1}(\boldsymbol{\gamma}^u - \overline{\boldsymbol{\gamma}_\mathbf{0}})\right] L(\boldsymbol{Y}|\boldsymbol{\gamma}^u)}{\exp\left[-\frac{1}{2}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}_\mathbf{0}})'\Sigma_{\boldsymbol{\gamma}0}^{-1}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}_\mathbf{0}})\right] L(\boldsymbol{Y}|\boldsymbol{\gamma})}, 1\}$$

We define diffuse priors by setting $\overline{\boldsymbol{\gamma}_\mathbf{0}}$ to be a vector of zeros and $\Sigma_{\boldsymbol{\gamma}\mathbf{0}} = 30I$.[14]

Step 2: Generate a_i^u, b_i^u :

[14] Our estimation is not sensitive to the setting of the diffuse hyperprior.

$$f(a_i^u, b_i^u | \boldsymbol{\gamma}^u\ , \alpha_0^u, \alpha_1^u, d_{ij}, data)$$

$$\propto N\left(\left(a_i^u, b_i^u | \boldsymbol{\beta}^u\ , \alpha_0^u, \alpha_1^u, d_{ij}\right), \Sigma_{ab}\right) L(\boldsymbol{Y})$$

$$\propto |\Sigma_{ab}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(a_i^u, b_i^u)\Sigma_{ab}^{-1}(a_i^u, b_i^u)'\right] L(\boldsymbol{Y})$$

Because this distribution does not have a closed form, we use the Metropolis-Hastings algorithm to draw from the conditional distribution of a_i, b_i: a_i, b_i is the draw of the random effect from the previous iteration, and we draw a_i^u, b_i^u by $\begin{bmatrix} a_i^u \\ b_i^u \end{bmatrix} = \begin{bmatrix} a_i \\ b_i \end{bmatrix} + \Delta \begin{bmatrix} a \\ b \end{bmatrix}$, where $\Delta \begin{bmatrix} a \\ b \end{bmatrix}$ is a draw from N(0,$\Delta^2\Lambda$), and Δ and Λ are chosen adaptively to reduce autocorrelation among MCMC draws following Atchade (2006). The probability of accepting this $\begin{bmatrix} a_i^u \\ b_i^u \end{bmatrix}$, the updated value for $\begin{bmatrix} a_i \\ b_i \end{bmatrix}$ is:

$$\Pr(acceptance) = \min\{\frac{\left[\exp\left(-\frac{1}{2}(a_i^u, b_i^u)\Sigma_{ab}^{-1}(a_i^u, b_i^u)'\right)\right] L(\boldsymbol{Y}|a_i^u, b_i^u)}{\left[\exp\left(-\frac{1}{2}(a_i\ , b_i\ )\Sigma_{ab}^{-1}(a_i\ , b_i\ )'\right)\right] L(\boldsymbol{Y}|a_i\ , b_i\ )}, 1\}$$

Step 3: $\Sigma_{\mathrm{ab}}^{\mathrm{u}} | a_i^u, b_i^u$

$$(\Sigma_{\mathrm{ab}}^{\mathrm{u}} | a_i^u, b_i^u) \sim IW_2(7 + N, G_0^{-1} + \sum_{i=1}^{N} (a_i^u, b_i^u)(a_i^u, b_i^u)')$$

Step 4: $d_{ij}^u d_{ij}^u, d_{ji}^u | \alpha_0^u, \boldsymbol{\gamma}^u\ , a_i, b_i, \alpha_1^u, \sigma_d^2, data$

$$f(d_{ij}^u, d_{ji}^u | \alpha_0^u, \boldsymbol{\gamma}^u\ , a_i, b_i, \alpha_1^u, \sigma_d^2, data\ )$$

$$\propto N\left(\left(d_{ij}^{u}, d_{ji}^{u} \middle| \alpha_0^u, \boldsymbol{\gamma}^u, a_i, b_i, \alpha_1^u\right), \sigma_d^2\right) L(\boldsymbol{Y})$$

$$\propto \sigma_d^{-1} \exp\left[-\tfrac{1}{2}\left(d_{ij}^{u} + d_{ji}^{u}\right)^2 \sigma_d^{-2}\right] L(\boldsymbol{Y})$$

We use the Metropolis-Hastings algorithm to draw from this conditional distribution of d_{ij}^{u} and d_{ji}^{u} : d_{ij} and d_{ji} are the draw of the unobservable similarity effects from the previous iteration, and we draw d_{ij}^{u}, d_{ji}^{u} by $\begin{bmatrix} d_{ij}^{u} \\ d_{ji}^{u} \end{bmatrix} = \begin{bmatrix} d_{ij} \\ d_{ji} \end{bmatrix} + \Delta\boldsymbol{d}$, where $\Delta\boldsymbol{d}$ is a draw from N(0,$\Delta^2\Lambda$), and Δ and Λ are chosen adaptively to reduce autocorrelation among MCMC draws following Atchade (2006).

The probability of accepting $\begin{bmatrix} d_{ij}^{u} \\ d_{ji}^{u} \end{bmatrix}$ is:

$$\Pr(acceptance) = \min\{\frac{\left[\exp\left(-\frac{1}{2}\left(d_{ij}^{u} + d_{ji}^{u}\right)\sigma_d^{-2}\right)\right] L(\boldsymbol{Y}|d_{ij}^{u}, d_{ji}^{u})}{\left[\exp\left(-\frac{1}{2}\left(d_{ij} + d_{ji}\right)\sigma_d^{-2}\right)\right] L(\boldsymbol{Y}|d_{ij}, d_{ji})}, 1\}$$

Step 5: Generating σ_d^{u}

$$(\sigma_d^u |\, d_{ij}^u, d_{ji}^u) \sim IW_1(1 + N(N-1), 1 + \textstyle\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (d_{ij}^u + d_{ji}^u)^2)$$

Step 6: Go to step 1.