Networks, Information & Social Capital

Sinan Aral,
New York University, Stern School of Business
sinan@stern.nyu.edu

Marshall Van Alstyne,
Boston University, School of Management
mva@bu.edu

Draft Version: March 15, 2008

DRAFT – COMMENTS WELCOME

This paper is based on a draft formerly titled "Network Structure & Information Advantage."

* We are grateful to Lada Adamic, Wayne Baker, Erik Brynjolfsson, Ron Burt, Paul Carlile, Emilio Castilla, Jerry Davis, Stine Grodal, Michael Macy, Erol Pekoz, Damon Phillips, Arun Sundararajan, Ezra Zuckerman and seminar participants at the Workshop on Information Systems Economics, the Sunbelt Social Networks Conference, the International Conference on Network Science, the Academy of Management Conference, Harvard, NYU, MIT, Stanford and the University of Chicago for valuable comments, and to the National Science Foundation (Career Award IIS-9876233 and grant IIS-0085725), Cisco Systems, France Telecom and the MIT Center for Digital Business for generous funding. We thank Tim Choe, Petch Manoharn, Cyrus-Charles Weaver and Jun Zhang for their tireless research assistance.

Networks, Information & Social Capital

Abstract

This paper investigates how information flows enable individuals to generate social capital from their social networks. By combining social network and performance data with the information content encoded in email communication, we examine the long held but empirically untested assumption that diverse networks drive performance by providing access to novel information. We demonstrate empirically that diverse networks do provide diverse, novel information, and that access to novel information predicts performance. But whether diverse networks deliver novel information depends on a tradeoff between network diversity and relationship channel bandwidth: as networks become more diverse, communication over each channel contracts. As network diversity and channel bandwidth both enable access to more novel information, diverse networks provide more novel information (a) when the topic space is large, (b) when topics are distributed non-uniformly across nodes and (c) when information in the network changes frequently. Diverse networks are not just pipes into diverse knowledge pools, but also inspire non-redundant communication even when the knowledge endowments of contacts are homogeneous. Consistent with theories of cognitive capacity, bounded rationality, and information overload, there are positive but diminishing performance returns to novel information. Network diversity also contributes to performance when controlling for the performance effects of novel information, suggesting additional non-information based benefits to structural diversity. These analyses unpack the mechanisms that enable information advantages in networks and serve as a 'proof-of-concept' for using email content to analyze relationships among information, networks and social capital.

Keywords: Social Networks, Social Capital, Information Content, Information Diversity, Network Size, Network Diversity, Performance, Productivity, Information Work.

Introduction

The micro-social processes that govern how people create and appropriate value from social capital remain a fundamental yet understudied feature of modern sociology. Social capital has been defined as "the sum of the resources, actual or virtual, that accrue to an individual or group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition" (Bourdieu & Wacquant 1992: 119). However, our theoretical and empirical understanding of the resources networks provide, the mechanisms linking networks to resources, and the processes governing the creation and appropriation of value from resources distributed through networks remain vague (Adler & Kwon 2002).[1] A prominent resource linking networks to value is information. Differential access to information, guided by differences in social structure, can not only create value, but may also help build other resources that contribute to social capital development. As information plays a central role in structural theories of social capital and brokerage, we focus on clarifying how networks create social capital and value through their influence on access to information.

One of the most prominent mechanisms theorized to drive the relationship between social structure and value (typically defined as realized social or economic advantage) is the existence of 'information benefits' to network structure. According to this argument, actors in diverse structural positions enjoy social and economic advantages based on their access to heterogeneous, novel information. Burt (1992) shows that individuals with structurally diverse networks (networks low in (a) cohesion, and (b) structural equivalence) are more successful in terms of wages, promotion, job placement, and creativity (Burt 2004a). He argues that these performance differentials can be explained in part by actors' access to diverse pools of knowledge, and their ability to efficiently gather non-redundant information – what he terms the 'vision advantage' (Burt 2004a).[2] Aral, Brynjolfsson and Van Alstyne (2006) demonstrate that

1 As Adler & Kwon (2002: 4) note "[s]keptics have … characterized the social capital concept as "a wonderfully elastic term," (Lappe & Du Bois 1997: 119), a notion that means "many things to many people" (Narayan & Pritchett 1997: 2), and [taking] on "a circus-tent quality" (De Souza Briggs 1997: 111)."

2 Coleman's (1988) argument, that focused information from cohesive networks provides more precise signals of actors' environments, also assumes that cohesive networks provide focused (while diverse networks provide diverse) information.

structural diversity is associated with higher levels of performance for task-based information workers. These studies, and numerous others, infer that network diversity is associated with performance in part because diverse contacts provide access to novel information (e.g. Sparrowe et al. 2001, Reagans and Zuckerman 2001, Cummings & Cross 2003). Novel information is thought to be valuable due to its local scarcity. Actors with scarce, novel information in a given network neighborhood are better positioned to broker opportunities, make better decisions, or apply information to problems that are intractable given local knowledge.

But, while there is abundant evidence linking social structure to performance (e.g. Burt 1992, 2004a, 2007, Reagans and Zuckerman 2001, Cummings 2004, Aral, Brynjolfsson & Van Alstyne 2006, 2007), direct evidence on information flowing through networked relationships is rarely used to validate information-based theories of social capital. As a consequence, our theories of the role of information in social capital formation and our understanding of how and why social structure explains economic outcomes remain underdeveloped. Comprehensive theories of the structure-performance relationship require a more thorough examination of the intermediate information mechanisms through which social structure affects economic advantage. As Burt (2005: 60) has recently noted: "There is abundant and accumulating empirical evidence of returns to brokerage. Evidence on the mechanism is not abundant. Initial research established the social capital potential of brokerage by focusing on returns to brokerage … The next phase of work is to understand the information-arbitrage mechanisms by which people harvest the value buried in structural holes … More generally, the sociology of information will be central in the work …"

We embark on this next phase of inquiry by directly examining the central premise of the information advantage argument: that diverse social structure drives performance by providing access to novel information. Using detailed empirical evidence on information content flowing through email networks in a medium sized executive recruiting firm, we lay the groundwork for a new, replicable mode of inquiry into how information flows generate social capital. Our research setting is well suited to this investigation: Since recruiters were frequently geographically dispersed and since instant messaging was rarely

used, recruiters relied on email as a primary means of communication.[3] To build theory relating network structure to information structure we investigate how topological properties of individuals' network positions (network size and network diversity) impact the diversity and novelty of the information they receive and distribute, and whether this in turn explains productivity and performance.[4]

The data and empirical methods also deserve special note. We collected and analyzed the topical content of email messages to determine the relative heterogeneity and novelty of the information passed between employees at the firm. Previous research by Wu et. al. (2004) and Kossinets & Watts (2006) has validated the usefulness of email data in characterizing and analyzing social networks in firms and academic institutions. We complement and extend this research by combining analysis of the social structure of email communication with evaluation of the information content of messages. We argue that combining analysis of message content and communication topology can open new avenues for answering questions at the heart of the sociology of information.

Our analyses uncover four primary results. First, we demonstrate that structural diversity predicts access to diverse novel information, and that diverse novel information in turn enables higher levels of productivity and performance. These findings provide empirical support for what has been a critical inference at the heart of social capital and brokerage theory. Our results also disentangle the effects of expertise endowments (the stocks of information, expertise and knowledge individuals build over time) from the effects of information flows (information shared between individuals in conversation and correspondence). While networks provide channels into the pools of information with which contacts are endowed, these links do not guarantee that information will flow. Individuals with diverse networks have contacts with more heterogeneous knowledge on average. But, because social structures contextualize and inform choices about which information is shared and with whom we share it, having a diverse communication network is a better predictor of access to diverse novel information than the heterogeneity of the knowl-

3 We thank Ron Burt for emphasizing this point.

4 The term "information structure" is used in the economics literature to denote the mapping of states of nature to signals i.e. news, received by a decision maker (see Arrow 1985). We use the term to mean the distribution of information across a network.

edge endowments of one's contacts. We explore why this is the case and the implications for information based theories of social capital.

Second, we find evidence of a tradeoff between network diversity and relationship channel bandwidth, the communication volume observed between individuals.[5] Although diverse networks should deliver information that is heterogeneous *across* channels (or alters), they are also made up of disproportionately weaker ties, implying lower bandwidth channels of communication (e.g. more infrequent communication across fewer topical dimensions). As networks become more diverse, channel bandwidth decreases. This tradeoff has implications for access to information because all else equal higher bandwidth channels deliver more diverse information due to the heterogeneity of the information flowing *within* each channel. As a result, whether diverse-low bandwidth networks deliver more diverse information than cohesive-high bandwidth networks ultimately depends on (a) the dimensionality of the information in the network (whether the total number of topics communicated across the entire network is large or small), (b) the distribution of information across the network (whether topics are uniformly or unevenly distributed across nodes) and (c) the rate at which the information in the network refreshes or changes.

Third, diminishing returns set in at two levels: Network size is a concave predictor of information diversity, and perhaps more importantly, there are diminishing marginal benefits to novel information. Part of the explanation for the decreasing contribution of network size to information diversity is that network diversity is increasing in network size, but with diminishing returns. As actors establish relationships with a finite set of possible contacts in an organization, the probability that a new relationship will be non-redundant and provide access to novel information decreases as possible connections in the network are exhausted. In addition, the marginal value of novel information is itself bounded. Consistent with evidence on the limits to human cognitive capacity, we find declining marginal benefits to more

5 We use the phrase relationship "channel bandwidth" carefully, and in preference to the more inclusive "strong tie" to draw attention to the volume of literal communication shared among people. In general, stronger ties imply greater bandwidth but the added precision allows us to also handle unusual cases. For example, adult children may have strong ties to parents based on emotional affinity, trust, or care-giving, yet be observed to communicate more frequently with co-workers who are less emotionally significant in their lives. We draw out the importance of focusing on information diversity, actually observed over a communications channel, in developing the theories that follow.

novel information. After a point, too much information is associated with decreasing and negative performance returns. This result implies limits to the value of more novel information and larger, more diverse networks.

Finally, network diversity contributes to performance even when controlling for the performance effects of novel information, suggesting additional benefits to diverse networks beyond those conferred through information based mechanisms. Somewhat surprisingly, traditional demographic and human capital variables (e.g. age, gender, industry experience, education) have little effect on access to diverse information, highlighting the importance of network structure for information advantage.

These results represent some of the first evidence on the relationship between network structure and information content. The findings advance our understanding of the economic value of information and the intermediate mechanisms driving the relationship between social structure and productivity. Our methods for analyzing network structure and information content in email data are replicable, opening a new line of inquiry into the relationships among networks, information and social capital.

Theory

Network Structure & Information Advantage: A Critical Inference

The assumption that network structure influences the distribution of information and knowledge in social groups (and thus characteristics of the information to which individuals have access) underpins a significant amount of theory linking social structure to economic outcomes (e.g. Simmel 1922 (1955), Moreno 1940, Granovetter 1973, Baker 1990, Burt 1992, Padgett & Ansell 1993, Podolny 2001, Reagans & Zuckerman 2001). Granovetter (1973) argues that topological properties of friendship networks, constrained by basic norms of social interaction, empower weak ties to deliver information about socially distant opportunities more effectively than strong ties. He posits that contacts maintained through weak ties typically "move in circles different from our own and thus have access to information different from that which we receive… [and are therefore]… the channels through which ideas, influence, or information socially distant from ego may reach him" (Granovetter 1973: 1371). Burt (1992) argues that networks rich

in structural diversity confer "information benefits" by providing access to diverse perspectives, ideas and information. As information in local network neighborhoods tends to be redundant, structurally diverse contacts provide channels through which novel information flows to individuals from distinct pools of social activity. Redundant information is less valuable because many actors are aware of it at the same time, reducing opportunities associated with its use. Structural redundancy is also inefficient because actors incur costs to maintain redundant contacts while receiving no new information from them (Burt 1992).

In contrast, exposure to diverse ideas, perspectives, and solutions is thought to enable information arbitrage, the creation of new innovations, and access to economic opportunities. Hargadon and Sutton (1997) describe how engineers use their structural positions between diverse engineering and scientific disciplines to broker the flow of information and knowledge from unconnected industrial sectors, creating novel design solutions. As Burt (2004b) puts it, "creativity is an import-export game,… not a creation game." The value of information in a network stems from its uneven distribution across actors and resides in pockets of distinct and diverse pools of information and expertise in local network neighborhoods. Actors with access to these diverse pools "benefit from disparities in the level and value of particular knowledge held by different groups…" (Hargadon & Sutton 1997: 717), and one of the key mechanisms through which network structures are theorized to improve performance is through access to novel, non-redundant information (Burt 1992).

While the argument that network structures influence performance through their effect on the distribution of information is intuitive and appealing, the vast majority of empirical work on networks and information advantage remains 'content agnostic' (Hansen 1999: 83), and infers the relationship between network structure and information structure from evidence of a link between networks and performance (e.g. Sparrowe et al. 2001, Cummings & Cross 2003). Reagans & Zuckerman (2001) infer that productivity gains from the external networks of corporate R&D teams are due in part to "information benefits," "a broader array of ideas and opportunities," and access to "different skills, information and experience." Burt (1992, 2004a) also makes this assumption, inferring that the observed co-variation of wages, promo-

tion, job placement, and creativity with network diversity is due in part to access to diverse and novel information. Some equate network content with the social function of relationships. For example, Burt (2000: 45) refers to "network content" as "the substance of relationships, qualities defined by distinctions such as friendship versus business versus authority." In one of the first studies to explore this type of network content, Podolny & Baron (1997) showed that while cohesive ties are beneficial in 'buy-in' networks and for those contacts that have control over the fate of employees, structural holes are important for collecting advice and information. Others define network content as the "attributes of the nodes" with respect to their self reported knowledge (Rodan & Galunic 2004).

We take a different view of network content, focused on the subject matter of communication flowing through the network rather than on the social function of relationships or actors' knowledge stocks. By focusing on the topical content of communication (as well as actors' knowledge) we distinguish information flows from knowledge endowments. As roles do not necessarily imply communication content and as not all knowledge possessed by an individual is necessarily shared or successfully transferred (Reagans & McEvily 2003), examining topical information flows between actors is critical to theories explaining how information and knowledge move and are distributed in networks, and how information in turn enables social capital development.

The limited research that empirically examines information transfer in networks has either focused on identifying tie and network characteristics that facilitate effective knowledge transfers; or on types of information (e.g. complex or simple; tacit or explicit) most effectively transferred through different types of ties. As a result, the fundamental assumption that structurally diverse network contacts provide access to diverse and novel information remains unexplored. For example, several studies examine how characteristics of dyadic relationships, like the strength of ties, impact the effectiveness of knowledge transfer, and how knowledge transfer processes in turn affect performance (Granovetter 1973, Uzzi 1996, 1997, Hansen 1999). These studies infer the impact of network structure on the effectiveness of knowledge sharing from the strength of individual dyadic relationships. Reagans & McEvily (2003) extend this work by simultaneously examining the effects of tie strength and network structure on the ease

of transferring knowledge between individuals. These studies either examine the strength of dyadic ties or the impact of network structure on discrete dyadic information transfer events, rather than on the information actors receive from all their network contacts in concert. Others examine characteristics of the information transferred across different types of ties. For example, Hansen (1999, 2002) and Uzzi (1996, 1997) explore the degree to which knowledge being transferred is tacit or codifiable, simple or complex, and related or unrelated to a focal actor's knowledge, and Fleming et al. (2007) examine brokerage and cohesion in patent collaboration and knowledge in patent content.

To complement this research, we ask a related, yet fundamentally different question: Do networks affect the acquisition of diverse and novel information and to what extent does this intermediate mechanism predict performance? In pursuing this question, we undertake three fundamental departures from the current literature. First, by exploring the relative information content differences among different network contacts, we explore actors' information diversity evaluated in relation to the body of information available in the entire network. Second, we focus on subject matter. Rather than characterizing the simplicity or complexity of information, or the degree to which knowledge is codifiable or tacit, we explore the topical content being discussed. Both simple and complex information can have either focused or diverse topical content. Complexity and codifiability do not describe whether information is topically similar or dissimilar, or novel relative to a larger body of knowledge. Third, we focus on the flow of information between actors, as well as the information that resides at each node. Disentangling information flows from information stocks is important because "knowledge transfer is a discretionary activity" (Regans & McEvily 2003: 243). A variety of social and organizational factors could distinguish the information with which actors are endowed from the information they choose to pass on and succeed in transferring (Aral et. al. 2007, Wu et. al. 2004). As the theoretical mechanism linking structure to performance through information rests on the relative novelty of the information actors send and receive, these three departures from previous research are critical to effectively exploring the dimensions of information theorized to drive value in networks.

In the absence of empirical evidence on the information processes that enable social networks to generate social capital, we develop theory about the relationships between network structure and access to heterogeneous, novel information, and between access to heterogeneous, novel information and performance.

----- FIGURE 1 -----

Our arguments problematize and unpack these relationships and our empirical analyses provide evidence on how networks generate social capital through their effects on information flow.

Network Structure & Access to Heterogeneous, Novel Information

Two network characteristics in particular are theorized to drive access to diverse, novel information: network size and network diversity. These characteristics are discussed extensively in the literature and are fundamental because they represent the two dimensions of structure most directly related to information acquisition. As Burt (1992: 16) argues "everything else constant, a large, diverse network is the best guarantee of having a contact present where useful information is aired…" Network size is the most familiar structural feature of networks theorized to deliver more novel information (Burt 1992). A larger network puts individuals in the flow of ideas and connects them with more potentially useful sources of information. However, growing an expansive network with redundant contacts is unlikely to increase access to novel information. Structurally diverse networks provide channels into different pools of information and knowledge and put individuals in contact with diverse social groups. Diverse networks – networks low in constraint and structural equivalence – are therefore likely to deliver more diverse and novel information. These are the central inferences on which structural theories of social capital and brokerage rest. We therefore expect that *network size and network diversity are positively associated with receiving more diverse information and more total non-redundant information.*

Network Diversity & Knowledge Heterogeneity. The leading theoretical explanation for why diverse networks should deliver diverse information is couched in terms of access to diverse knowledge stocks – that diverse networks provide links into heterogeneous pools of information (e.g. Simmel 1922

(1955), Granovetter 1973, Caro 1982, Burt 1992, Reagans & Zuckerman 2001). Information shared within social groups is theorized to be more homogeneous than information shared across groups (Burt 2004a). This should create opportunities for individuals with diverse networks to access heterogeneous pools of information that are novel relative to the information in their own local network neighborhoods (Rodan & Galunic 2004). A necessary condition of this theory is that diverse networks connect individuals to alters with heterogeneous knowledge and experience. Whether structural diversity delivers diverse information is, according to this argument, a function of the extent to which diverse structure accesses heterogeneous subsets of the information space in a network. If the population of alters has relatively homogeneous information then structural diversity should matter relatively little. We therefore expect that *network diversity is positively associated with the knowledge heterogeneity of actors' contacts.*

Knowledge similarity in organizational settings can arise from educational or demographic similarities, overlapping expertise (Reagans & McEvily 2003) or project co-work (Rodan & Galunic 2004, Aral et. al 2007). Since executive recruiting does not require specialized educational attainment, employees in our firm develop expertise and accumulate stocks of knowledge in domains pertaining to the work they complete by working on projects of different types and developing expertise in specialized executive searches. Employees with similar backgrounds who work on similar projects or in similar functional areas develop overlapping sets of expertise and rely on similar bodies of knowledge to conduct their work. Common knowledge, derived from common prior experience, helps ease knowledge transfer (Reagans & McEvily 2003) and structures the topics of conversation employees are likely to share (Burt 1987, 1992, Cohen & Levinthal 1990, Rogers 1995, Smith-Loving, & Cook 2001). If knowledge and expertise within work groups is more homogeneous than across them, then access to different groups, with varied expertise should facilitate receipt of more diverse and (from the focal actor's point of view) novel information. For this reason we expect that *the knowledge heterogeneity of an actor's contacts is positively associated with receiving more diverse information and more total non-redundant information.*

Distinguishing Information Flows From Knowledge Stocks. While links to alters with diverse knowledge stocks should increase the likelihood of receiving more diverse and novel information, infor-

mation may not necessarily flow in direct proportion to information endowments. There are two prevailing views of network content that deserve special attention in the case of information flows. These approaches view networks either as "pipes" or "prisms" (Podolny 2001). In his study of venture capital markets, Podolny (2001: 33) argues that economic sociologists and organizational scholars have traditionally regarded networks as "plumbing" or "the channels or conduits through which "market stuff" (including information) flows."[6] But, as Podolny (2001: 34) (citing Baum & Oliver's (1992) study of day care centers and Podonly's (1993) study of investment banks) argues, ties are not merely pipes in the plumbing, but also "an informational cue on which others rely to make inferences…," (in his case about the quality of potential transaction partners). The same logic applies to information flows in networks – ties may not simply be conduits for information, but may also inform decisions about what information is shared and with whom.

We have little empirical evidence on how information moves through social groups (Wu et. al. 2004), or the relationships between social structure and information flow (for a review of the literature on information diffusion in networks see Aral et. al. 2007). Though most current views define network content as the "attributes of nodes" to which ties give access (e.g. Rodan & Galunic 2004), information exchange is a social process and knowledge transfer is a discretionary activity (Reagans & McEvily 2003, Wu et. al. 2004). A connection to an individual with a certain information endowment affords the possibility of receiving that information, but by no means guarantees it. As Wu et. al. (2004: 328) point out: "There are … differences between information flows and the spread of viruses. While viruses tend to be indiscriminate, infecting any susceptible individual, information is selective and passed by its host only to individuals the host thinks would be interested in it." In fact, the distinction can be expressed more starkly – information is sometimes withheld even when it is known to be of interest to others, reflecting complex selection and discretion in social choices concerning information sharing.

6 Parenthesis added by the authors. Podolny (2001) mentions in the next sentence that "market stuff" encompasses information as well as goods, services and transactions.

Most current models of information flow in networks implicitly assume that information will flow in proportion to the distribution of information in the network. An actor's likelihood of receiving information is generally assumed to be some function of the strength of ties to others endowed with that information or the number of ties in the actor's network that have the information in question (e.g. Granovetter 1978, Schelling 1978, Kleinberg, Kempe & Tardros 2003).[7] However, networks may not simply be pipes into different pools of information. Network structure may also correlate with the choices individuals make about which information they share. For example, we may hesitate to share sensitive information with someone we know is part of our own tightly clustered social circle for fear of it 'getting back to us,' or we may chose to share the subset of our information stock that pertains to our role vis-a-vis our communication partners (with roles reflected in structural similarities). An individual embedded in a cohesive social structure may be more likely to receive redundant information not only because their communication partners possess homogeneous knowledge and experience, but also because social cohesion and structural equivalence reflect social circumstances that inspire communication partners to pass on redundant information even when their knowledge stocks are heterogeneous. Prior theory and empirical evidence support this conjecture. Social cohesion and interdependence has been shown to promote solidarity (Durkeim 1933 (1893)), support (Durkheim 1951 (1897)) and behavioral homogeneity (Lazarsfeld et. al. 1944, Coleman et. al. 1966) primarily due to communication. Social cohesion may also reflect similarities in activity, purpose, task and the circumstances of a particular time and place. Localized knowledge and information pertaining to local tasks or circumstances are likely to be shared by those who communicate regardless of their prior experience. These types of information and knowledge are pertinent to production activities and are typically localized in social groups close to common activities (Hayek

7 Two core models have emerged to explain the diffusion of influence and contagion. Threshold models posit that individuals adopt innovations (or receive information) after surpassing their own private "threshold" (e.g. Granovetter 1978, Schelling 1978). Cascade models posit that each time an adjacent individual adopts, the focal actor adopts with some probability that is a function of their relationship (e.g. Kempe, Kleinberg, Tardos 2003). While both models assume an information transmission between adopters and non-adopters, they rarely specify the nature of the information or the conditions under which exchanges take place. Rather, the diffusion process is typically tested under various assumptions about the distribution of thresholds or dyadic adoption probabilities in the population. In fact, as Kempe, Kleinberg, Tardos (2003: 2) explain "the fact that [thresholds] are randomly selected is intended to model our lack of knowledge of their values."

1945). In organizational settings these activities involve working on the same projects or executing managerial responsibilities together. In addition, structural equivalence reflects role similarity (Friedkin 1984, Burt 1987, Borgatti & Everett 1993). Structurally equivalent alters are likely to perform similar social and organizational roles amongst their peers (Burt 1987). If structural equivalence is a proxy for role equivalence, then two structurally equivalent individuals may share the same subset of their information stock with a third party due to the similarity of their roles in their relationships to that third party even if their overall knowledge stocks are highly heterogeneous. The potential relationship between social structure and choices concerning information sharing suggests that in order to empirically substantiate theories of information-based social capital formation in networks we need evidence not only of actors' information endowments, but also of the information content actors actively share with one another.

Thus, a natural question arises: do diverse network ties deliver diverse information primarily by acting as pipes into heterogeneous pools of knowledge and expertise, or do individuals embedded in cohesive ties share similar information due to social similarities of time, place, task and role regardless of the heterogeneity of their knowledge and expertise? Both of these mechanisms could simultaneously be at play and there is also likely to be correspondence between knowledge endowment homogeneity and social cohesion as we have hypothesized. To determine the relative strength of these mechanisms, we estimate whether communication network diversity moderates the relationship between contact knowledge heterogeneity and access to diverse novel information or visa versa (Baron & Kenny 1986). If network diversity moderates the relationship between knowledge heterogeneity and access to diverse information, we have preliminary evidence that cohesive social structure influences the similarity of information flows above and beyond the influence of the similarity of the prior experience and knowledge of actors in cohesive networks, suggesting that networks are not merely 'pipes' into diverse pools of information, but also 'prisms' on which actors make inferences about which information to actively share.

Network Diversity & Channel Bandwidth. In examining the relationship between network diversity and information diversity, it is important to distinguish the diversity of information flows across ties from the diversity of the information flowing within each tie. Theoretical arguments concerning network

diversity and novel information have thus far focused almost exclusively on the relative diversity of the information received *across* alters in a network, rather than the diversity of information flowing *within* each tie (or channel) over time. When within-channel diversity is considered, the inference that network diversity drives performance by providing diverse, novel information is less straightforward in light of empirical evidence that diverse networks are made up of disproportionately weaker ties. While cohesive or constrained ties typically deliver information that is redundant *across* channels (which we refer to as 'information bias'), they are also typically stronger (Granovetter 1973, Burt 1992), implying greater bandwidth. Weak ties are by their nature lower bandwidth conduits for information (Granovetter 1973, Burt 1992). Information should flow less frequently (Granovetter 1973), with lower complexity (Hansen 1999) and detail (Uzzi 1999), and along fewer topical dimensions (see Granovetter 1973: p 1361) through weak ties. Social cohesion motivates individuals to devote time and effort to communicating with and assisting others due to the cooperative nature of ties surrounded by other third party ties (Granovetter 1985, Coleman 1988). The development of cooperative norms (Granovetter 1992) and the subsequent reduction in competition reflected in cohesive ties are likely to increase knowledge transfer between individuals embedded in cohesive networks (Szulanski 1996, Argote 1999, Reagans & McEvily 2003). Given evidence on the prevalence of weak ties in diverse networks and the likelihood of increased knowledge flow in cohesive networks, the bandwidth of communication channels should be lower in diverse networks. Thus, we expect that *network diversity is associated with lower channel bandwidth.*

When channel bandwidth is incorporated into the argument, whether diverse networks deliver diverse information depends not only on the diversity of the network around ego, but also on the bandwidth of the channels and the interplay between network diversity and channel bandwidth. Centola & Macy (2007) make a related yet distinct argument about contagions based on the number of ties rather than the bandwidth of channels. A simple model demonstrates that although a diverse network of weak ties ("diverse-weak") *can* provide access to more novel information than a constrained network of strong ties ("constrained-strong"), the converse is also possible. To illustrate, let E represent the event that an ego encounters *new* information through a new link. If n is a subset of all possible topics T ($n<T$), then an

actor receives "biased" content if she is more likely to receive news on one set of topics than another ($p_1 > p_2$), where p_1 and p_2 are the probabilities of receiving information from topic subsets n_1 and n_2. More precisely, a person with biased content has an asymmetric distribution over the likelihood of seeing different topics. Note that basic laws of probability require $n\,p_1 + (T\text{-}n)\,p_2 = 1$. Since the likelihood of encountering new information depends on what ego has learned from existing links, let L represent current contacts.[8] Then $P[E^c]$, the probability of encountering novel information from a new constrained link, can be described as:[9]

$$P[E^c] = p_1 n(1-p_1)^L + p_2(T-n)(1-p_2)^L \qquad [1]$$

Unbiased content implies $p_1 = p_2$, so that Equation 1 reduces to $P[E^D] = pT(1\text{-}p)^L$, where E^c and E^D represent the events of forging a constrained and a diverse link and getting new information.[10] To model the more frequent communication of the higher bandwidth tie, let B represent additional chances to cover new material over the constrained link during any given interval. Simplifying with $n_2 = T\text{-}n_1$ gives total accumulated probability of:

$$P[E^C] = \sum_{l=L}^{L+B} P[E^c] = p_1 n_1 (1-p_1)^L + p_2 n_2 (1-p_2)^L + \ldots p_1 n_1 (1-p_1)^{L+B} + p_2 n_2 (1-p_2)^{L+B} \qquad [2]$$

To see that a constrained-strong tie could offer more novel information, let $p_1 = p_2 + \varepsilon$ implying negligible bias so that $P[E^c] \approx P[E^D]$. Then choose any B large enough such that the following inequality is strict:

$$P[E_L^c] + P[E_{L+1}^c] + \ldots P[E_{L+B}^c] \approx P[E_L^D] + P[E_{L+1}^D] + \ldots P[E_{L+B}^D] > P[E_L^D] \qquad [3]$$

This demonstrates the original claim that a constrained-strong tie can supply a greater volume of novel information than a diverse-weak tie provides. When the advantage of bandwidth swamps the disadvantage

8 More precisely, l represents an information exchange with an existing link. In probabilistic terms it is a sample on link l such that ego receives information on a given topic n_i with probability p_i from each sample, making the likelihood of receiving new information a function of the number of samples (or analogously, the thickness of the communication channel).

9 Since our purpose is illustrative rather than proof theoretic, we refrain from presenting non-essential primitives and assumptions here and present the derivation of Equation 1 in Appendix A.

10 The likelihood of encountering novel information (for both constrained and unconstrained ties) decreases strictly and asymptotically toward 0 with each additional tie L. This exactly mirrors the pattern we observe empirically as shown later in Figure 5.

of bias, an ego *always* prefers the constrained-strong tie to the diverse-weak tie to increase the chances of encountering novel information.

To see when an diverse-weak tie could be preferred, let a "group think" network spread its bandwidth only over the subset of *n* topics with probability $p_1=B/T$ (such bias necessarily constrains $p_2 \approx \varepsilon$). For ease of simplification, let $n = T/B$. Then algebra reduces the relative probabilities to:

$$P[E_L^c] = \left(1-\frac{B}{L}\right)^L < \left(1-\frac{1}{T}\right)^L = P[E_L^D] \qquad [4]$$

This alternative case demonstrates the counterclaim, that a diverse-weak tie can supply a greater volume of novel information than a constrained-strong tie provides. Although $P[E_L^C] = P[E_L^c] + ...P[E_{L+B}^c]$ and increasing *B* adds more terms to $P[E_L^C]$ and none to $P[E_L^D]$, it also causes each term to approach *0* faster. No matter how large the bandwidth on constrained ties, there always exists a fixed number of links *L* such that link *L+1* should be an unconstrained tie. When the disadvantage of bias swamps the advantage of bandwidth, an ego *always* prefers the diverse-weak tie to the constrained-strong tie to increase chances of encountering novel information. While an enormous range of intermediate cases span these two extremes, conditions exist when a person could always prefer one or the other type of link depending on bias, bandwidth, and the number of links already present. All else equal, greater channel bandwidth should provide access to more diverse information. Stronger ties imply tighter relationships and thicker channels of communication. In relationships characterized by strong ties and high channel bandwidth, contacts are likely to be more willing to share information and to be similar across a greater number of dimensions, inspiring information exchanges across a larger number and a wider variety of topics. Our model implies that as the bandwidth of a channel increases, the topical diversity of the information flowing through it should also increase, providing recipients more chances to encounter novel information. We therefore

expect, all else equal, that *channel bandwidth is positively associated with receiving more diverse and novel information.* [11]

If greater structural diversity limits channel bandwidth and if channel bandwidth provides access to more novel information, then estimates of the impact of network diversity on access to novel information that exclude channel bandwidth will be downward biased. Controlling for variance explained by channel bandwidth should therefore increase the strength of the positive association between network diversity and access to diverse and novel information. In addition to bandwidth and bias, it follows from our model that access to more diverse information depends on the interaction of at least three additional factors beyond structural diversity: (a) the dimensionality of the information in the network (whether the total number of topics communicated across the entire network is large or small), (b) the distribution of topics over nodes (whether topics are uniformly or non-uniformly distributed across the network and (c) the rate at which the information in the network refreshes or changes. [12] Although in our empirical setting we

11 We distinguish channel bandwidth from the strength of ties although they are likely correlated. The strength of a tie may be a noisy reflection of the bandwidth of the channel. More detailed empirical work on the relationship between the strength of ties and the bandwidth of channels may provide evidence on how the social function of relationships (Podolny & Barron 1997, Burt 2000) is associated with the nature of the conduits of information flow they enable. We encourage this work although we do not focus on it here.

12 The first important factor determining the relative importance of bandwidth and bias in this theoretical framework is the dimensionality of the information space being considered. In a network in which there are relatively few topics and total information diversity is low, we expect bias to be more important than bandwidth. In contrast, when the dimensionality of the information space is high, with a large number of topics and a high degree of information diversity across the entire network, we expect bandwidth to play a more significant role in determining access to novel information. We show the robustness of our basic insight as information diversity increases or, in this case, as T increases without bound. For both the biased strong tie and the unbiased weak tie we have:

$$\lim_{T\to\infty}(P[E_L^C]) = \lim_{T\to\infty}(P[E_L^D]) = 1 \quad \text{; and} \quad \lim_{T\to\infty}(p_1) = \lim_{T\to\infty}(p_2) = 0 \text{ .}$$

This should make intuitive sense because new links are more likely to provide novel information when the number of possible topics vastly exceeds that covered by prior links. Yet, the likelihood of learning about any specific topic falls as the number of total topics grows. The ratio, however, is extremely informative. If the biased strong tie provides B times more bandwidth (see Appendix A), then the ratio simplifies to $\frac{p_1}{p_2} = \frac{(T-n_1)B}{T-n_1B}$ which means that $\lim_{T\to\infty}(\frac{p_1}{p_2}) = B$. Thus in a world of highly diverse information, i.e. *T* very large, an ego should prefer strong ties, even though they are biased, because they provide much more total new information. In our setting, we expect the disadvantage of bias to swamp the advantage of bandwidth. Interviews indicate that the dimensionality of information content in executive recruiting is limited (in the parlance of our model T, the space of topics, is small) meaning thicker channels are not as necessary to communicate information on more topics. Therefore, as individuals communicate with more contacts, and as individuals' networks connect them to actors that are

do not observe sufficient variation in these dimensions to definitively test their implications we feel they help bound and contextualize our generalizations, and we provide some empirical evidence of their importance in the discussion.

Non-Linearity in Information Acquisition. Finally, it is unlikely that there are constant or increasing returns to network size. The relationship between information diversity and network size is likely concave not only because there are costs to relationships, but also because in organizational networks, where the number of possible contacts is finite, there are natural constraints on the diversity of a network as it grows in size. While a greater number of contacts are likely to provide access to more diverse, non-redundant information, the probability that an additional contact will have novel information is likely decreasing in the size of an individual's network. This expectation is a direct result of our model and is also supported by prior empirical evidence on network formation. Social networks tend to cluster into homophilous cliques (for a review see McPherson, Smith-Loving, & Cook 2001). Since individuals usually make connections through contacts they already have, in bounded networks the likelihood that a marginal contact will be redundant should increase in the number of people already known.[13] As actors establish relationships with a finite set of alters, the probability that a new relationship will be structurally non-redundant should decrease as possible alters in the network are exhausted. We therefore expect that *the marginal increase in information diversity is decreasing in network size*. Part of the explanation for this limitation is that there exist natural constraints on network diversity in bounded organizational networks such that *the marginal increase in network diversity is decreasing in network size*.

---

themselves unconnected and structurally non-equivalent, we expect the information they receive to be more diverse and we expect them to receive more total novel information as hypothesized. The second contingent factor is whether topics are uniformly or non-uniformly spread across the network. It follows from our theory and our model that if information is uniformly distributed, diverse contacts should not increase access to diverse information. This assumption lies at the heart of theories linking networks diversity to information advantages. Information turbulence provides the final contingent factor. It represents the extent to which topics within T obsolesce and require ego to refresh information on any given topic in any pool. Extreme turbulence, in which data obsolesce each period, favors seeking information diversity through a high bandwidth constrained-strong tie exclusively. Since all information is necessarily new information, $P[E_C] = P[E_D] = 1$, an ego always prefers the multiple samples provided by equation 2.

Access to Heterogeneous, Novel Information & Performance

Access to more heterogeneous and novel information should promote stronger managerial performance by enabling higher quality decisions, facilitating the development of managerial skills and providing a context for more effective political maneuverability. The most familiar benefit of access to information is improved decision making. Environmental awareness and knowledge of the variety of options available to a decision maker enables more optimal decisions (and actions) by increasing the accuracy of mental mappings from options or actions to expected consequences and outcomes (Marschak & Radner 1972). In the context of innovation, novel solutions outside localized social and intellectual spheres can enable new designs, products or solutions that cannot be generated from local knowledge (Burt 2004a). Some of the most groundbreaking innovations of this century have materialized through combinations of ideas from different disciplines (e.g. quantum computing, bioinformatics, nanotechnology), and a good deal of research into social capital has focused on the innovation benefits to diverse novel information (e.g. Hargadon & Sutton 1997, Reagans & Zuckerman 2001, Burt 2004a, Rodan & Gallunic 2004, Van Alstyne & Brynjolfsson, 2005, Lazer & Friedman 2005).

Access to diverse novel information also enables skill development by increasing familiarity and facility with different topics, improving the breadth of individuals' absorptive capacity and strengthening the ability to communicate across a wider array of subjects. As people are exposed to new ideas and information from a wide array of disciplines and topics, they are more able to absorb new ideas by associating them with what they already know. Developing the ability to absorb new ideas and concepts enables more effective knowledge transfer (Cohen & Levinthal 1990, Simon 1991) and makes it more likely that others will share more novel information with a given recipient (Reagans & McEvily 2003), reflecting the social selectivity of information sharing. As absorptive capacity is developed across a wider range of sub-

13 We focus on internal networks due to difficulties in collecting reliable data outside the firm and in estimating accurate network structures without access to whole network data (see Barnes 1979, Marsden 1990). As Burt (1992: 172) demonstrates however "little evidence of hole effects [are] lost... when sociometric choices [are] restricted to relations within the firm."

jects, individuals are better able to communicate ideas across a broader range of topics and to a broader audience, strengthening persuasion and the ability to generate broader support from subject matter experts in accomplishing managerial goals (Rodan & Galunic 2004).

Finally, access to diverse information, representing the perspectives and knowledge of varied social groups can create autonomy (Simmel 1922 (1955), Burgelman 1991, Burt 1992) and enable political maneuverability (Padgett & Ansell 1993), helping individuals get access to resources they need to do their jobs efficiently (Rodan & Galunic 2004). We therefore expect *access to non-redundant and diverse information is positively associated with recruiters' performance.*

Diminishing Returns to Novel Information. The performance benefits of access to more diverse and novel information depend not only on the value of novel information to the decision maker, but also on the ability of the decision maker to process and use the information they receive. Limits to human cognitive capacity make us susceptible to information overload (Simon 1991). Receiving more information should improve decision making to a point, after which new information may exceed the information processing capacity of the decision maker creating confusion, making information more difficult to find or recall and reducing the value of new information for decision quality and performance (Galbriath 1974, Tushman & Nadler 1978). Several theoretical results from information economics also predict non-linearities in the value of information. Arrow (1985) demonstrates that expected payoffs from decisions about uncertain events are concave in the amount of information the decision maker obtains, implying diminishing returns to more information. As measured by decision relevance, value only increases when new information leads to different and better decisions (Arrow 1985, Hirshleifer 1973). Information is novel if it provides an alternate perspective on a known topic or knowledge of an altogether new topic. As new information on known topics accumulates, beliefs tend to converge on a particular view of the world, making further confirmation unnecessary. Expected convergence under Bayes' Rule, for example, exhibits clear diminishing returns such that, beyond some threshold, more news has no more value. As new information on new topics accumulates, value is also likely to exhibit diminishing marginal returns due to decision irrelevance. As actors' information space becomes disparate, ideas are less likely to connect in

complementary ways and each bit of information is less likely to be relevant to the space of decisions and actions the actor is interested in. We find evidence of diminishing returns to novel information in our own analytical model. Collectively, these arguments suggest a non-linear relationship between access to more novel information and performance. We therefore expect that *the marginal increase in performance associated with access to novel information is decreasing in the amount of novel information actors receive.*

Alternative Mechanisms Linking Network Diversity & Performance. Other mechanisms could also explain the observed relationship between network diversity and performance. Network contacts could provide resources other than information (e.g. Podolny & Barron 1997), there could be power or control benefits to network structure independent of information flows (e.g. Burt 1992), and structural diversity could reduce dependence, place individuals in favorable trading relationships (e.g. Emerson 1962) or entitle them to benefits from informal reciprocity (e.g. Cook, Emerson & Gilmore 1983). These alternate mechanisms could also explain the link between structural diversity and performance without any prediction concerning actors' information access. While we expect network structure to impact performance through its effects on access to diverse and novel information, these other intermediate mechanisms could also tie network diversity to performance. We therefore expect that *network diversity is positively associated with performance, controlling for access to novel information.*

----- TABLE 1 -----

Empirical Setting – Executive Recruiting

To explore the information mechanisms that govern the creation of social capital, we identified a firm in an information intensive industry in which employees rely heavily on social information seeking through email communication. We studied a medium-sized executive recruiting firm with fourteen offices across the United States. As work was geographically dispersed and as recruiters relied heavily on email to seek information and coordinate tasks, we focused on analyzing the content of communication flowing in email networks. As one recruiter reported "[s]taff spend an enormous amount of time coordinating. We are *big* users of email," giving us confidence that our data cover a substantial portion of communication

traffic in the firm.[14] The email network of the firm displays a hub and spoke structure, with a dense core of thirty four recruiters at the firm's headquarters, and spokes in thirteen other offices located across the United States. The network of recruiters located at the headquarters is highly clustered (Clustering Coefficient = 25.32) and dense (Mean Density = 11.02, SD = 32.44) compared to the firm as a whole (Clustering Coefficient = 12.20; Mean Density = 5.41 (19.08).[15] There are two medium sized offices of six and seven recruiters each, with the rest comprised of one, two or three recruiters per office. This structure offers a unique window onto the value of network and information diversity as measured in email data, as geographic dispersion makes email a critical source of information.

----- FIGURE 2 -----

Our interviews revealed that the core of executive recruiters' work involves matching job candidates to clients' requirements; a process which is information-intensive and requires activities geared toward assembling, analyzing, and making decisions based on information gathered from team members, other firm employees, and contacts outside the firm. Access to information enables higher quality decisions in this setting. Recruiters report being more effective when they receive rich information from their colleagues about candidate qualifications, circumventing screening barriers, handling difficult placements, and team coordination.[16] Although recruiters also rely on databases and resumes to make decisions, social information seeking is critical to success. One recruiter told us that "[i]n meetings you'll hear 'Oh he looks good on paper, but he's an idiot,' or 'go talk to Simon about her," implying the importance of verifying what's "on paper" through social information exchanges with colleagues who have particular expertise or knowledge. Even subtle things such as a candidate's appearance can affect the likelihood of a successful match: "Mary has two inch nails. She doesn't present well in the LA market." In this case the

14 Although we did not collect phone conversation data or face to face information exchanges, our parameter estimates should be downward biased by the lack of such data, making statistically significant relationships between networks, information and performance more difficult to observe.

15 The Clustering Coefficient is the number of ties divided by the total number of possible ties that could exist (Watts & Strogatz 1998). The mean density is the average of the number of ties divided by the number of pairs in each ego network.

16 We conducted interviews over the course of a year beginning in October 2001.

recruiter conducting the search needed to seek information from colleagues with expertise on how a candidate's appearance might be perceived in the LA market. As another recruiter put it "[m]uch of the human data is private, never captured in the [corporate] database. It stays in my head." Even when the database contains relevant information, social verification is important because "[p]eople move or get promoted. [It is therefore best to] assume the database is 60% accurate."

Sharing procedural information can also improve efficiency and effectiveness (Szulanski 1996). For example, information exchanged through social communication helps recruiters navigate entry into client firms and candidate pools. One recruiter told us that "[c]all penetration can be really hard into private companies so researchers and consultants swap information to get through." Information sharing also enables coordination and eliminates the need for recruiters to 'reinvent the wheel' when they are searching for similar candidates or clients: "Communication within and across teams is a big success factor. It eliminates double work." In these ways, information helps recruiters fill different types of positions and perform complex matching of candidate strengths and weaknesses to client wants and needs.

Diverse and novel information is particularly useful in executive recruiting. Qualitative studies have shown that recruiters fill "brokerage positions" between clients and candidates and rely heavily on information flows to complete their work effectively (Finlay & Coverdill 2000). Information about a diverse pool of candidates, diverse markets and diverse client firms reduces time wasted interviewing unsuitable candidates and increases the quality of placement decisions by improving the fit between candidates and clients' requirements (Aral et. al. 2006). Recruiters report that tapping diverse information pools helps them do their work effectively. They emphasize the need for diverse contacts and report that "[d]iversity means more and better contacts" because "[s]kill sets are complementary and not perfectly overlapping." While interviewing recruiters, we also talked with several executive recruiter trainers. One trainer, who describes her job as 'helping recruiters learn to be better recruiters,' told us that "[To be a successful recruiter one should] develop relationships with people you don't know… Some folks join groups for their prestige but you should join clubs for their diversity. Whom do you what to know? Go

there. Neophytes hang out on their own." For these reasons we expect diverse and novel information to be particularly important for explaining variance in recruiter performance.

The process for executing a recruiting contract is relatively standard: A partner secures a contract with a client and assembles a project team (team size mean = 1.9, min = 1, max = 5). The team then establishes a universe of potential candidates including those in similar positions at other firms and those drawn from the firm's internal database of resumes and other leads.[17] Candidates are vetted on the basis of perceived quality, their match with the job description and other factors. After conducting initial due diligence, the team chooses a subset of candidates for internal interviews, approximately six of which are forwarded to the client along with detailed background information, notes and a formal report of the team's due diligence. The team then facilitates the client's interviews with each candidate, and the client, if satisfied with the pool, makes offers to one or more candidates. A contract is considered complete when a candidate accepts an offer.

Recruiters generate revenue on the basis of completed contracts. Therefore, the speed with which vacancies are filled is an important intermediate measure of workers' productivity. Contract completion implies that recruiters have met a client's minimum thresholds of candidate fit and quality and "[t]he longer a client delays, the lower the probability of job acceptance." Project duration can therefore be interpreted as a quality controlled measure of productivity. In assessing individual recruiters' performance, we measure revenues generated per month, projects completed per month and average project duration per month. Effective recruiters rely on being "in the know" and delivering candidates with professional and personal attributes that fit client needs. To accomplish this, recruiters must interact with several different information channels to match candidates' attributes with clients' requirements. We therefore expect recruiters with diverse, non-redundant information to complete more projects, to complete projects faster, and to generate more revenue for the firm per unit time.

Methods

By analyzing email communication patterns and message content, we are able to match information channels to the subject matter of the content flowing through them. Our empirical approach also addresses a methodological puzzle that has historically troubled network research. In traditional network studies, a fundamental tradeoff exists between comprehensive observation of whole networks and the accuracy of respondents' recall. Most research elicits network data from respondents who have difficulty recalling their networks (e.g. Bernard et. al 1981), especially among individuals socially distant to themselves (Krackhardt & Kilduff 1999). The inaccuracy of respondent recall and the bias associated with recall at social distance creates inaccurate estimates of network variables (Kumbasar, Romney & Batchelder 1994), forcing most empirical studies to artificially limit the boundary of estimated networks to local areas around respondents (e.g. Reagans & McEvily 2003). Such empirical strategies create estimation challenges due to the sensitivity of network metrics to the completeness of data (Marsden 1990). If important areas of the network are not captured, estimates of network positions can be biased. Furthermore, as our content measures consider the similarity of topics across the entire network, poor coverage of the firm could bias our estimates of the relative novelty or diversity of topics discussed via email. We therefore take several steps to ensure a high level of participation (described below). As 87% of eligible recruiters agreed to participate, and given that our inability to observe the remaining 13% is limited to messages between two employees who both opted out of the study, we collect email network and individual content data with nearly full coverage of the firm and there are no statistical differences between participants and those who opted out on dimensions of relevance to the study.[18]

---

17 We have also studied executive recruiters' use of information contained in the firm's internal database or 'Executive Search System.' For more detailed analyses on how use of the Executive Search System impacts performance see Aral et. al. (2006).

18 F-tests comparing performance levels of those who opted out with those who remained did not show statistically significant differences. F (Sig): Rev02 2.295 (.136), Comp02 .837 (.365), Multitasking .386 (.538).

Data

Our data come from four sources: (i) detailed accounting records of individual project assignments and performance, (ii) email data captured directly from the corporate server, (iii) survey data on demographic characteristics, human capital and information seeking behaviors, and (iv) data from the web site Wikipedia.org used to validate our analytical models of information diversity. The firm gave us access to their complete internal accounting and project databases for records spanning 2000 to 2005. Internal accounting project data describe: revenues generated by individual recruiters, contract start and stop dates, projects handled simultaneously by each recruiter, project team composition, and job levels of recruiters and placed candidates. These provide excellent performance measures that can be normalized for quality. Email data cover 10 months of complete email history at the firm. The data were captured from the corporate mail server during two equal periods from October 1, 2002 to March 1, 2003 and from October 1, 2003 to March 1, 2004. Participants received $100 in exchange for permitting use of their data, resulting in 87% coverage of eligible recruiters and more than 125,000 email messages captured.[19] Details of email data collection are described by Aral, Brynjolfsson & Van Alstyne (2006). The third data set contains survey responses on demographic and human capital variables such as age, education, industry experience, and information-seeking behaviors. Survey questions were generated from a review of relevant literature and interviews with recruiters. Experts in survey methods at the Inter-University Consortium for Political and Social Science Research vetted the survey instrument, which was then pre-tested for comprehension and ease-of-use. Individual participants received $25 for completed surveys and participation exceeded 85%. The fourth dataset is a set of 291 entries collected from Wikipedia.org, which we describe in detail in the section pertaining to the validity of our information diversity metrics (see Appendix

19 We wrote and developed email capture software specific to this project and took multiple steps to maximize data integrity. New code was tested at Microsoft Research Labs for server load, accuracy and completeness of message capture, and security exposure. To account for differences in user deletion patterns, we set administrative controls to prevent data expunging for 24 hours. The project went through nine months of human subjects review and content was masked using cryptographic techniques to preserve privacy (see Van Alstyne & Zhang 2003). Spam messages were excluded by eliminating external contacts who did not receive at least one message from someone inside the firm.

C). Descriptive statistics and correlations of all variables are provided in Tables 2 & 3 (we provide details on the construction of each variable in the next section). An observation is a person-month.

----- TABLE 2 -----

----- TABLE 3 -----

Variable Construction

Network Structure

Network Size. The size of *i*'s network (S_i) is simply the number of contacts with whom *i* exchanges at least one message. Size is the most familiar network characteristic related to information benefits and is a good proxy for a variety of characteristics, like degree centrality, betweenness centrality and network reach, which describe the breadth and range of actors' networks (see Burt 1992: 12). In our data, network size is significantly correlated with degree centrality ($\rho = .70$; $p < .001$), betweeness centrality ($\rho = .77$; $p < .001$), and reach ($\rho = .56$; $p < .001$), demonstrating its value as a proxy for network breadth.

Network Diversity. Network diversity describes the degree to which contacts are structurally 'non-redundant,' and there are both first order and second order dimensions of redundancy as shown in Figure 3. Individuals who are in contact are likely to share information and be aware of the same opportunities, ideas and expertise. Networks in which contacts are highly connected are termed 'cohesive.' Contacts that are themselves connected to the same people are termed 'structurally equivalent.' We measure redundancy in the first order by the lack of constraint in actors' networks, and in the second order by the average structural equivalence of actors' contacts.[20] We define the constraint C_i (Burt 1992: 55)[21] of an actor's network as the degree to which an individual's contacts are connected to each other, such that $C_i = \sum_j \left(p_{ij} + \sum_q p_{iq} p_{qj} \right)^2$, $q \neq i, j$; and the structural diversity D_i of an actor's network as $1 - C_i$. We use the standard definition of the structural equivalence of two actors, measured as the Euclidean distance

20 By measuring both first and second order network diversity we account for the possibility that small world networks (Watts 1999), clustered cliques linked by infrequent weak ties, could bring novel information into a cohesive clique from contacts two steps removed from ego.

of their contact vectors.[22] By measuring both first and second order network diversity we account for the possibility that small world networks, cohesive cliques linked by infrequent weak ties, could bring novel information into a clique

Knowledge Heterogeneity. We measure the knowledge heterogeneity of employee's contacts by evaluating the diversity of their expertise. In this setting recruiters' develop expertise as they complete projects of different types. As there is little in the way of formal training to become an executive recruiter, we do not use recruiters' educational backgrounds but rather the distributions of their prior project experience over project types to measure knowledge heterogeneity. The firm categorizes projects into the following categories: CEO, COO, CIO, Medical Executive, Human Resources Executive, Business Development Executive and 'Other.' We use these categories as the relevant areas of recruiters' expertise.[23] The *Knowledge Heterogeneity* variable is constructed using a Herfindahl Index of the expertise of an actor's contacts in each month, weighted by the strength of the tie to each alter. As the firm records each employee's effort share on each project, the expertise of a recruiter is share weighted by the amount of effort they recorded against any given project in the accounting data. The measure is constructed as follows:

$$KH_{it} = 1 - \sum_{k=1}^{9} \left(\frac{q_{ik}}{q_i} \right)^2 .$$

In this measure, $q_{ik} = \sum_{j=1}^{n} w_{ij} P_{jk}$ represents the total amount of prior experience in *i*'s network in project class *k*, weighted by the strength of the tie to each of *i*'s contacts w_{ij} (the number of messages exchanged between *i* and *j*) and summed over all of *i*'s contacts *j*. P_{jk} represents *j*'s prior experience in job class *k*, where *P* is an effort share weighted count of the number of projects of class *k* that *j* has completed. The

21 Where $p_{ij} + \sum p_{iq} p_{qj}$ measures the proportion of i's network contacts that directly or indirectly involve j and C_i sums this across all of i's contacts.

22 Euclidean distance measures the square root of the sum of squared distances between two contact vectors, or the degree to which contacts are connected to the same people. We measure the average structural equivalence of actors' direct contacts.

denominator, $q_i = \sum_{k=1}^{9} q_{ik}$ represents the total project experience in *i*'s network summed over all project classes. Thus, the ratio $\left(\frac{q_{ik}}{q_i} \right)$ is the share of prior experience in project class *k* over the total project experience in *i*'s network. We then construct a Herfindahl Index of this ratio measuring the concentration of expertise across job classes among *i*'s contacts. To measure heterogeneity rather than concentration we subtract this measure of project experience concentration from one. As the expertise in *i*'s network becomes more concentrated in a few project classes, the knowledge heterogeneity measure decreases.[24] Reagans & McEvily (2003) construct a similar measure of '*expertise overlap,*' although our measure uses accounting records to record project experience (rather than self reports of expertise), and weights the expertise in an employee's network by the strength of their tie and the effort share of each alter on each project. Our measure of knowledge heterogeneity also changes over time as recruiters complete more projects of different types.

Channel Bandwidth. Bandwidth measures the volume of communication over a given channel. We record the average message traffic per communication channel or tie, operationalized as the amount of incoming email over the total number of contacts (the size of the actor's network) at time *t*:

$$B_{it} = \left(\frac{E_{it}^{I}}{S_{it}} \right).$$

----- FIGURE 3 -----

Non-Network Determinants of Information Advantage

---

23 We also ran specification controlling for other categorization schemes and sub-categories of 'Other' jobs clustered by their project descriptions, which returned similar results. We therefore retained the firm's original classification.

24 To normalize the Knowledge Heterogeneity measure so that its values range from zero to one, we scale the measure by multiplying the final metric by (9/8), creating this final metric: $KH_{it} = \frac{9}{8}\left[1 - \sum_{k=1}^{9} \left(\frac{q_{ik}}{q_i} \right)^2 \right]$. This scaling does not affect the distribution of the measure or the outcome of any of our analyses. It simply allows the measure to range from zero to one.

Several other factors could affect access to diverse information and individual performance other than our variables of interest. We therefore examine five possible alternative explanations for information advantage: demography, human capital, total communication volume, unobservable individual characteristics and temporal shocks to the flow of information in the firm.

Demography. That demography could influence performance, learning capabilities and the variety of ideas to which individuals have access has been well documented (e.g. Pfeffer 1983, Ancona & Caldwell 1992, Reagans & Zuckerman 2001). Older employees may have prior related knowledge on a wider variety of topics or may be more aware of experts in the organization. Employment discrimination and interpersonal difference could also impact the relative performance and information seeking and sharing habits of men and women. We therefore control for the age and gender of employees.

Human Capital. Greater industry experience, education or individuals' organizational status could also create variation in access to diverse and novel information and performance. As individuals gain experience, they may collect expertise across several domains, reflected in communications across multiple subjects or topics. It could also be that as individuals gain experience, they specialize and focus their work and their communication on a limited number of topics. We therefore control for the level of education, industry experience measured by the number of years employees have worked in executive recruiting, and organizational position. As employees occupy one of three positions in the firm – partner, consultant or researcher – we include dummy variables for each of these positions to account for authority and status differentials that could explain variation in both access to information and performance.

Total Communication Volume. We are interested both in the total amount of novel information and the importance of network structure holding communication volume constant. Other studies have demonstrated the importance of controlling for communication volume to isolate the effects of structural variables (e.g. Cummings & Cross 2003). We therefore control for total email communication.

Individual Characteristics & Temporal Shocks. Some employees may simply be more social or more ambitious, creating variation in information seeking habits and performance. To control for unobservable individual characteristics we test fixed effects specifications of each of our hypotheses. At the

same time, temporal shocks could affect demand for the firm's services and subsequent information seeking activities associated with more work. In our data, business exhibits seasonal variation. Demand for the firm's services picks up sharply in January and declines steadily through the next eight months. These exogenous shocks to demand could drive simultaneous increases in project workload, information seeking and revenue generation creating a spurious correlation between information flows and output. There could also be non-seasonal transitory shocks to demand in a given year or a given month of a given year. For this reason, we control for both seasonal and transitory variation in our data with dummy variables for each month and year.

Modeling Information Heterogeneity: A Vector Space Model of Communication Content

We model and measure the diversity of information in individuals' email using a Vector Space Model of the topics present in email content (e.g. Salton et. al. 1975).[25] Vector Space Models are widely used in information retrieval and search query optimization algorithms to identify documents that are similar to each other or pertain to topics identified by search terms. They represent textual content as vectors of topics in multidimensional space based on the relative prevalence of topic keywords. In our model, each email is represented as a multidimensional 'topic vector' whose elements are the frequencies of keywords in the email. The prevalence of certain keywords indicates that a topic that corresponds to those keywords is being discussed. For example, an email about pets might include frequent mentions of the words "dog," "cat," and "veterinarian;" while an email about statistics might mention the words "variance," "specification," and "heteroskedasticity." The relative topic similarity of two emails can then be assessed by topic vector convergence or divergence – the degree to which the vectors point in the same or

25 While email is not the only source of employees' communication, it is one of the most pervasive media that preserves content. It is also a good proxy for other social sources of information in organizations where email is widely used. In our data, the average number of contacts by phone ($\rho = .30$, $p < .01$) and instant messenger ($\rho = .15$, $p < .01$) are positively and significantly correlated with email contacts. Our interviews indicate that in our firm, email is a primary communication media.

orthogonal directions.[26] To measure content diversity, we characterize all emails as topic vectors and measure the variance or spread of topic vectors in individuals' inboxes and outboxes. Emails about similar topics contain similar language on average, and vectors used to represent them are therefore closer in multidimensional space, reducing their collective variance or spread.

Construction of Topic Vectors & Keyword Selection. Vector Space Models characterize documents D_i by keywords k_j weighted according to their frequency of use (or with 0 weights for words excluded from the analysis – called "stop words"). Each document is represented as an n-dimensional vector of keywords in topic space,

$$\overrightarrow{D_i} = (k_{i1}, k_{i2}, ..., k_{in}),$$

where k_{ij} represents the weight of the *j*th keyword.

----- FIGURE 4 -----

Weights define the degree to which a particular keyword impacts the vector characterization of a document. Words that discriminate topics are weighted more heavily than words less useful in distinguishing topics. As terms that appear frequently in a document are typically thematic and relate to the document's subject matter, we use the 'term frequency' of keywords in email as weights to construct topic vectors and refine our keyword selection with criteria designed to select words that *distinguish* and *represent* topics.[27]

In order to minimize their impact on the clustering process, we initialized our data by removing common "stop words," such as "a, "an," "the," "and," and other common words with high frequency across all emails that are likely to create noise in content measures. We then implemented an iterative, k-means clustering algorithm to group emails into clusters that use the same words, similar words or words

26 Each email may pertain to multiple topics based on keyword prevalence, and topic vectors representing emails can emphasize one topic more than another based on the relative frequencies of keywords associated with different topics. In this way, our framework captures nuances of emails that may pertain to several topics of differing emphasis.

27 Another common weighting scheme is the 'term-frequency/inverse-document frequency.' However, we use a more sophisticated keyword selection refinement method specific to this dataset described in detail in the remainder this section.

that frequently appeared together.[28] The result of iterative k-means clustering is a series of assignments of emails to clusters based on their language similarity. Rather than imposing exogenous keywords on the topic space, we extract topic keywords likely to characterize topics by using a series of algorithms guided by three basic principles.

First, in order to identify distinct topics in our corpus, keywords should *distinguish* topics from one another. We therefore chose keywords that maximize the variance of their mean frequencies across k-means clusters. This refinement favors words with widely differing mean frequencies across clusters, retaining words with an ability to distinguish between topics. In our data, we find the coefficient of variation of the mean frequencies across topics to be a good indicator of this dispersion.[29]

$$C_v = \frac{\sqrt{\frac{1}{n}\sum_i \left(m_i - \overline{M}\right)^2}}{\overline{M}}$$

Second, keywords should *represent* the topics they are intended to identify. In other words, keywords identifying a given topic should frequently appear in emails about that topic. To achieve this goal we chose keywords that minimize the mean frequency variance within clusters, favoring words that are consistently used across emails discussing a particular topic:[30]

$$ITF_i = \frac{\sqrt{\sum_c \sum_i \left(f_i - \overline{M_c}\right)^2}}{\overline{M_c}}$$

Third, keywords should not occur too infrequently. Infrequent keywords will not represent or distinguish topics and will create sparse topic vectors that are difficult to compare. We therefore selected

28 K-means clustering generates clusters by locally optimizing the mean squared distance of all documents in a corpus. The algorithm first creates an initial set of clusters based on language similarities, computes the 'centriod' of each cluster, and then reassigns documents to clusters whose centriod is the closest to that document in topic space. The algorithm stops iterating when no reassignment is performed or when the objective function falls below a pre-specified threshold.

29 The coefficient of variation is particularly useful due to its scale invariance, enabling comparisons of datasets, like ours, with heterogeneous mean values (Ancona & Caldwell 1992). To ease computation we use the square of the coefficient of variation, which produces a monotonic transformation of the coefficient without affecting our keyword selection.

30 i indexes emails and c indexes k-means clusters. We squared the variation to ease computation as in footnote 28.

high frequency words (not eliminated by the "stop word" list of common words) that maximize the inter-topic coefficient of variation and minimize intra-topic mean frequency variation. This process generated topical keywords from usage characteristics of the email communication of employees at our site.[31]

Measures of Information Diversity. Using the keywords generated by our usage analysis, we populated topic vectors representing the subject matter of the emails in our data. We then measured the degree to which the emails in an individual employee's inbox or outbox were focused or diverse by measuring the spread or variance of their topic vectors. We created five separate diversity measurement specifications based on techniques from the information retrieval, document similarity and information theory literatures (see Appendix B for detailed descriptions of each measure). The approach of all five measures is to compare individuals' emails to each other, and to characterize the degree to which emails are about a set of focused topics, or rather about a wider set of diverse topics. We used two common document similarity measures (Cosine similarity and Dice's coefficient) and three measures enhanced by an information theoretic weighting of emails based on their "information content."[32] We performed extensive validation tests of our diversity measures and their correlations, including application to an independent dataset from Wikipedia. A detailed description of the validation process and results appears in Appendix C. As all diversity measures are highly correlated (~ corr = .98; see Appendix B), our specifications use the average cosine distance of employees' incoming email topic vectors d_{ij}^{I} from the mean vector of their topic space M_i^I to represent incoming information diversity (ID_i^I):

$$ID_i^I = \frac{\sum_{j=1}^{N} \left(Cos\left(d_{ij}^{I}, M_i^{I}\right)\right)^2}{N}, \text{ where:}$$

31 We conducted sensitivity analysis of our keyword selection process by choosing different thresholds at which to select words based on our criteria and found results were robust to all specifications and generated keyword sets more precise than those used in traditional term frequency/inverse document frequency weighted vector space models that do not refine keyword selection.

32 Information Content is used to describe how informative a word or phrase is based on its level of abstraction. Formally, the information content of a concept c is quantified as its negative log likelihood –log p(c).

$$Cos(d_{ij}, M) = \frac{d_i \bullet M_i}{|d_i||M_i|} = \frac{\sum_j w_{ij} \times w_{Mj}}{\sqrt{\sum w_{ij}^2}\sqrt{\sum w_{Mj}^2}}, \text{ such that } 0 \leq ID_i^I \leq 1.$$

This measure aggregates the cosine distance of email vectors in an inbox from the mean topic vector of that inbox, approximating the spread or variance of topics in incoming email for a given individual. We measure the total amount of *i*'s incoming email communication as a count of incoming email messages, $E_i^I = \sum_j m_{ji}$, where m_{ji} represents a message sent from *j* to *i*; and the total amount of non-redundant information flowing to each actor *i* (NRI_i^I) as diversity times total incoming email: $NRI_i^I = (E_i^I * ID_i^I)$.

Model Specifications

We began by examining the structural determinants of access to diverse and novel information. We first estimated an equation relating network structure to the diversity of information flowing into actors' email inboxes using pooled OLS and fixed and random effects models on monthly panels of individuals' networks and information diversity.[33] The estimating equation is specified as follows:

$$ID_{it}^I = \gamma_i + \beta_1 E_{it}^I + \beta_2 NS_{it} + \beta_3 NS_{it}^2 + \beta_3 ND_{it} + \beta_4 SE_{it} + \beta_5 KH_{it} + \beta_6 B_{it} + \sum_j \beta_j HC_{ji} + \sum_m \beta_m M_{it} + \varepsilon_{it} \quad [5],$$

where ID_{it}^I represents the diversity of the information in a given individual's inbox, E_{it}^I represents the total number of incoming messages received by *i*, NS_{it} represents the size of *i*'s network, NS_{it}^2 represents network size squared, ND_{it} represents network diversity (measured by one minus constraint), SE_{it} represents average structural equivalence, KH_{it} represents the knowledge heterogeneity of *i*'s contacts, B_{it} represents channel bandwidth, $\sum_j \beta_j HC_{ji}$ represents controls for human capital and demographic

33 We focus in this paper on incoming information for two reasons. First, we expect network structure to influence incoming information more than outgoing information. Second, the theory we intend to test is about the information to which individuals have access as a result of their network structure, not the information individuals send. These dimensions are highly correlated.

variables (Age, Gender, Education, Industry Experience, and Managerial Level), and $\sum_m \beta_m M_{it}$ represents temporal controls for each month/year.

We then examined the relationship between network structure and the total amount of novel information flowing into actors' email inboxes (NRI_{it}^{I}), again testing pooled OLS and fixed and random effects specifications using the following model:

$$NRI_{it}^{I} = \gamma_i + \beta_1 NS_{it} + \beta_2 NS_{it}^2 + \beta_3 ND_{it} + \beta_4 SE_{it} + \beta_5 KH_{it} + \beta_6 B_{it} + \sum_j \beta_j HC_{ji} + \sum_m \beta_m M_{it} + \varepsilon_{it} \quad [6].$$

To explore the mechanisms driving the creation and appropriation of information advantages from network structure we explicitly considered a) the tradeoff between network diversity and channel bandwidth, b) whether diverse networks provide access to contacts with heterogeneous knowledge and c) whether a non-linear relationship between network size and network diversity could explain why the marginal increase in information diversity is decreasing in network size. To test these intermediate mechanisms we specified the following two equations.

$$ND_{it} = \gamma_i + \beta_1 NS_{it} + \beta_2 NS_{it}^2 + \beta_3 KH_{it} + \sum_j B_j HC_{ji} + \sum_m B_m M_{it} + \varepsilon_{it} \quad [7].$$

If the probability of contact redundancy is increasing in network size (implying that network diversity is bounded in organizational networks of finite size), we should see a non-linear positive relationship between network size and structural diversity, such that the marginal increase in structural diversity is decreasing in network size.

To explore the relationship between network diversity and channel bandwidth, and to test whether diverse networks are associated with lower channel bandwidth, we specified the model expressed in equation [8]. If network diversity is associated with lower channel bandwidth, we would expect to observe parameter estimates such that $\beta_1 < 0$ and $\beta_2 > 0$.

$$B_{it} = \gamma_i + \beta_1 ND_{it} + \beta_2 SE_{it} + \beta_3 NS_{it} + \beta_4 NS_{it}^2 + \sum_j B_j HC_{ji} + \sum_m B_m M_{it} + \varepsilon_{it} \quad [8].$$

Finally, we tested the relationship between non-redundant information (NRI_{it}^{I}) and performance (P_{it}), and included our measure of structural network diversity (ND_{it}) in the specification.

$$P_{it} = \gamma_i + \beta_1 NRI_{it}^{I} + \beta_2 \left(NRI_{it}^{I}\right)^2 + \beta_3 ND_{it} + \sum_j B_j HC_{ji} + \sum_m B_m Month + \varepsilon_{it} \quad [9].$$

If information benefits to network diversity exist, network diversity should be positively associated with access to diverse and non-redundant information, and non-redundant information should be positively associated with performance. If network diversity confers additional benefits beyond information advantage (such as power or favorable trading conditions) network diversity should contribute to performance beyond its contribution through information diversity.[34] Finally, if there are diminishing returns to novel information, we should see a concave relationship between novel information and productivity. As a robustness check we also estimated equation [9] replacing the non-redundant information variable (NRI_{it}^{I}) with incoming information diversity (ID_{it}^{I}).

We estimate relationships between network structure and information access, and between information access and performance using panel data. We are interested in how variation in network structure explains performance differentials across individuals, and also in how changes in actors' networks explain variation in their own performance over time. If network structure generates social capital by influencing information access, actors with larger, more diverse networks with higher channel bandwidth should receive more novel information and perform better than their counterparts. However, evidence of variation across individuals cannot exclude the possibility that unobservable characteristics of individuals, such as ambition or social intelligence, could simultaneously drive variation in network diversity and performance. If unobserved characteristics of individuals are correlated with the error terms in our models, pooled OLS estimation will produce biased parameter estimates. We therefore examine variation within and

34 We were unable to reject the hypothesis of no heteroskedasticity and report standard errors according to the White correction (White 1980). White's approach is conservative. Estimated coefficients are unbiased but not efficient. In small samples, we may observe low t-statistics even when variables exert a real influence. As there may be idiosyncratic error at the level of individuals, for OLS analyses we report robust standard

across individuals over time using both fixed effects and random effects models to control for bias created by this unobserved heterogeneity and to examine variation within and across observations of individuals over time. The statistical procedures used to estimate our specifications are detailed in Appendix D.

Results

Network Structure & Access to Diverse, Non-Redundant Information

We first estimated the relationships between network size, network diversity, knowledge heterogeneity, channel bandwidth and access to diverse information controlling for demographic factors, human capital, unobservable individual characteristics, temporal shocks and total communication volume in hierarchical regressions, adding variables to the specification in succession (see Table 4 Models 1-6).

----- TABLE 4 -----

The knowledge heterogeneity of recruiters' contacts is positively correlated with the diversity of the information they receive in both pair wise correlations (.23 $p < .05$, see Table 3) and regression results (see Table 4 Models 1-2). When network size is added to regression models, the magnitude of the positive association between knowledge heterogeneity and information diversity decreases from .28 to .12, indicating that as recruiters add network contacts the heterogeneity of their contacts' expertise increases, with size accounting for some of the variation in information diversity originally attributed to knowledge heterogeneity. Controlling for network size and total communication volume, a one standard deviation increase in the knowledge heterogeneity of recruiters' contacts is associated with a .12 standard deviation increase in incoming information diversity (Model 2, $p < .01$). Table 4 Models 2-6 all demonstrate that the diversity of information flowing to an actor is increasing in the actor's network size and network diversity, while the marginal increase in information diversity is decreasing in network size, supporting hypotheses 1 and 4a. A one standard deviation increase in the size of recruiters' networks (approximately 8 additional contacts) is associated with a 1.2 standard deviation increase in information diversity (Model 2,

---

errors clustered by individual. Clustered robust standard errors are robust to correlations within observations of each individual, but are never

$p < .01$); while the coefficient on network size squared is negative and significant indicating diminishing marginal diversity returns to network size.[35] As actors add network contacts, the contribution to information diversity lessens, implying that information benefits to network size are constrained. Network diversity is also positively and significantly associated with greater information diversity in incoming email. The first order diversity variable which measures the lack of constraint in an actor's network is highly significant in all specifications, while the average structural equivalence of actors' contacts does not influence access to diverse information controlling for network size and first order structural diversity.[36] These results demonstrate that large diverse networks provide access to diverse, novel sets of information. When the network diversity and structural equivalence terms are added to the estimation (Models 3), the positive contribution of knowledge heterogeneity to incoming information diversity disappears, implying that network diversity and knowledge heterogeneity are positively correlated and that network diversity is a stronger predictor of access to diverse information than the knowledge heterogeneity of recruiters' contacts. A one standard deviation increase in network diversity is associated with ~ .15 standard deviation increase in incoming information diversity. Finally, channel bandwidth is associated with access to more diverse information, confirming hypothesis H3b. A one standard deviation increase in channel bandwidth is associated with a .085 standard deviation increase in information diversity ($p < .01$). When channel bandwidth is added to the specification (Model 4), the magnitude of the estimated relationship between network diversity and information diversity increases implying a negative correlation between network diversity and channel bandwidth – providing preliminary evidence of a tradeoff between the two. Random effects specifications mirror the results of fixed effects specifications and demonstrate that traditional demographic and human capital factors have little effect on access to diverse information (see Model 6).

---

fully efficient. They are conservative estimates of standard errors.

35 We also tested a negative exponential specification of this relationship with very similar results. Both models fit well.

36 Structural equivalence does have a positive and significant correlation with access to diverse information when the network diversity variable is left out of the estimation.

We then tested relationships between network size, network diversity, knowledge heterogeneity, channel bandwidth and the total amount of novel information that accrues to recruiters in incoming email. Our results, shown in Table 4 Models 7-12, demonstrate that the amount of novel information flowing to an actor is increasing in the actor's network size, network diversity and channel bandwidth. Knowledge heterogeneity has a strong positive relationship with total non-redundant information (Model 7, $p < .01$)), until the network diversity and structural equivalence variables are added to the specification (Model 8) again demonstrating that network diversity is a stronger predictor of access to diverse novel information. Network diversity also has a strong positive relationship with the total amount of novel information flowing into actors' inboxes (Model 8, $p < .01$), but is not significant when controlling for network size (Models 9). The impact of size on total novel information dominates that of structural diversity because the total non-redundant information variable takes into account the *volume* of novel information, which increases significantly as the total number of contacts increases. Channel bandwidth is also a strong predictor of the volume of novel information (Model 10), with a one standard deviation increase in bandwidth associated with a .35 standard deviation increase in total novel information received ($p < .01$). These results highlight the importance of information flows over time. The amount of novel information flowing in networks of similar structural diversity is greater in larger networks with greater bandwidth. As network size and the thickness of channels increase, the total volume of novel information flowing into recruiters' inboxes also increases. These results also demonstrate the importance of considering channel bandwidth (and the relationship between network diversity and channel bandwidth) when estimating relationships between network structure and access to diverse novel information. Bandwidth seems to tradeoff with network diversity and has a strong positive relationship with incoming information diversity and total non-redundant information. The random effects models again mirror the results of fixed effects specifications, demonstrating that demographic and human capital factors have little effect on the amount of novel information recruiters receive. We would also expect network diversity to drive greater access to total non-redundant information, controlling for network size. However, our model and results imply that while structural diversity has a strong impact on the *diversity* of the information actors receive (per unit of

information), variation in the total *amount* of novel information received is determined mostly by network size and channel bandwidth, again drawing attention to the thickness of communication channels and the number of contacts in providing novel information. These results also suggest nuanced relationships between network diversity, network size and channel bandwidth, which we explore in detail in the next section.

Tradeoffs between Network Size, Network Diversity & Channel Bandwidth

Our model and prior research demonstrating a positive correspondence between weak ties and diverse networks lead us to predict a tradeoff between network diversity and channel bandwidth (the thickness of communication channels as measured by communication volume per channel). Our parameter estimates in Table 4 present some initial evidence of this tradeoff as the inclusion of the bandwidth variable increases the magnitude of the positive relationship between network diversity and information diversity, indicating that the exclusion of bandwidth downward biases parameter estimates of the relationship between network diversity and access to diverse information. In Table 5 Models 1-5, we explicitly test the tradeoff between network diversity and channel bandwidth holding the size of networks constant. There is a strong negative relationship between network diversity and channel bandwidth (β = -.314, p < .01) and a positive relationship between structural equivalence and channel bandwidth (β = .107, p < .10), indicating that as networks become more diverse the thickness of communication channels narrows. These results hold even when we control for network size (Table 5, Models 3-5) demonstrating that in networks of the same size, more diverse networks have lower bandwidth communication channels. Interestingly, the relationships do not seem to be driven by time and effort costs to network maintenance, but rather by the nature of relationships in sparse networks. The positive parameter estimate on the network size variable indicates that as actors cultivate more contacts the bandwidth of their communication channels widens rather than narrowing. If constraints on time and effort devoted to relationship maintenance were driving average channel bandwidth we would expect channel bandwidth to decrease as size in-

creased. On the contrary, as actors communicate with more people they also exchange more messages per contact. Taken together, these results indicate that the thickness of communication channels is narrower in diverse networks populated with weak ties. The network size squared estimate is negative and significant in random effects specifications indicating that the marginal increase in channel bandwidth is decreasing in network size. Knowledge heterogeneity is negatively associated with channel bandwidth in both pair wise correlations (ρ = - .25 p < .05) and random effects models (β = -.21 p < .01, Model 5) providing some insight into why diverse networks may be associated with lower channel bandwidth. Individuals whose contacts have diverse knowledge and experience communicate more infrequently and with lower volume per channel. This result is consistent with characterizations of weak ties in previous research (Granovetter 1973, Uzzi 1996) and provides new evidence that information flows may be weaker in diverse networks due to the experience and knowledge dissimilarity of individuals' contacts – effects which are more pronounced in random effects models that consider variation across individuals. Demographic variables have no effect on channel bandwidth in Models 4-5, while education has a consistently negative relationship, perhaps indicating an ability of more educated employees to communicate more efficiently with fewer messages per channel. Fixed and random effects models are relatively consistent, except that network size and knowledge heterogeneity variables only effect channel bandwidth in random effects models, indicating that persistent variation in network size and knowledge heterogeneity across individuals explains part of the variation in channel bandwidth while changes in individuals' network size and network knowledge heterogeneity over time does not. This is most likely because variation in network size and knowledge heterogeneity across individuals explains differences in channel bandwidth and because there are relatively smaller changes in these variables in a given individual's network over time.

----- TABLE 5 -----

The results in Table 5 Models 6-9 demonstrate that knowledge heterogeneity and network diversity are positively correlated. Diverse networks are populated with contacts whose prior experience and knowledge is heterogeneous, providing evidence of the first possible mechanism through which diverse

networks deliver diverse information – by providing pipes into pools of heterogeneous information. The results in Table 4 Models 3 - 6 and 8 however demonstrate that network diversity moderates the relationship between knowledge heterogeneity and access to diverse novel information. When network diversity is added to the regressions the relationship between knowledge heterogeneity and access to diverse novel information disappears. The positive relationship between knowledge heterogeneity and network diversity (Table 3 and Table 5) and the disappearance of the relationship between knowledge heterogeneity and diverse information when controlling for network diversity provides evidence that network diversity moderates the influence of knowledge heterogeneity on access to diverse information (Baron & Kenny 1986). Something about the social connection between people, beyond their similarity in prior experience, explains the redundancy of the information they share with others. We suspect this reflects similarities in activity, purpose, task and the circumstances of a particular time and place and discuss the implications further in the conclusion.

The results in Table 5 Models 6-9 also show a strong positive, but non-linear relationship between network size and network diversity: structural diversity is increasing in network size, but with diminishing marginal returns. This result supports hypothesis 4b and demonstrates why information benefits to larger networks may be constrained in bounded organizational networks. As recruiters contact more colleagues, the contribution of a marginal contact to the structural diversity of a focal actor's network is increasing, but with diminishing marginal returns. The implications of this tradeoff between size and structural diversity complement Burt's (1992: 167) concepts of "effective size" and "efficiency."[37] Figure 5 displays graphs relating network size, network diversity and information diversity, clearly showing the positive, non-linear relationships.

----- FIGURE 5 -----

Network Structure, Information Diversity & Performance

37 In fact, Burt (1992: 169) finds stronger evidence of hole effects with the constraint measures we employ than with effective size, demonstrating "exclusive access is a critical quality of relations that span structural holes."

Finally, we test the performance implications of network structure and access to diverse, non-redundant information as measured by revenues generated per month, projects completed per month, and the average duration of projects.[38] Table 6 displays strong evidence of a positive relationship between access to non-redundant information and performance.

----- TABLE 6 -----

In random effects models, which incorporate variation within and between individuals, a one standard deviation increase in the amount of non-redundant information flowing to individuals is associated with on average with just over $4,600 more revenue generated (Model 10, $p < .01$), an extra one tenth of one project completed (Model 6, $p < .01$), and 15 days shorter average project duration per person per month (Model 2, $p < .01$). These results support Hypothesis 5a and provide evidence of 'information advantages' to network structure. Tables 4 and 6 together demonstrate that diverse networks provide access to diverse, non-redundant information, which in turn drives performance in information intensive work. As a robustness check, we estimated the relationships between information diversity and performance with very similar results. A one standard deviation increase in information diversity is associated with increases in revenues ($\beta_{FE} = 1322.97$, N.S.; $\beta_{RE} = 2254.75$, $p < .01$) and project completions ($\beta_{FE} = .036$, $p < .05$; $\beta_{RE} = .049$, $p < .01$), and reductions in average project duration ($\beta_{FE} = -16.04$, $p < .01$; $\beta_{RE} = -15.78$, $p < .01$).

We also uncovered evidence of alternative mechanisms linking network structure to performance. Table 6 shows network diversity is positively associated with performance even when holding access to novel information constant, providing preliminary evidence of additional benefits to network structure beyond those conferred through information advantage. Holding access to novel information constant, network diversity is associated with greater revenue generation (Model 10, $p < .10$), more completed projects (Model 6, $p < .05$), and faster project completion (Models 2-4, $p < .01$, $p < .05$, $p < .10$). These results leave open the possibility that some benefits to network diversity come not from access to novel,

38 As there are some employees who do not take on projects or who are not involved in any projects in a given month, we only estimate equations for individuals with non-zero revenues in a given month.

non-redundant information, but rather from other mechanisms, like access to job support, power or organizational influence. Finally, we tested whether the positive relationship between access to novel information and performance was strictly linear, or rather whether access to novel information displayed diminishing marginal performance returns (Hypothesis 5b). We found across the board that access to non-redundant information had diminishing marginal performance returns in each of our performance measures (Tables 6 & 7).

----- TABLE 7 -----

These parameter estimates suggest that the marginal performance impacts of novel information are lower when employees already have access to significant amounts of novel information. In fact, as the graphs in Figure 6 demonstrate, there seem to be negative returns to more novel information beyond the normalized mean.[39] These non-linearities in the value of novel information likely arise for the reasons outlined. First, beyond the threshold for decision relevance, new information adds no value. Second, employees' capacity to process or act on new information may be constrained, making them less able to get the most out of novel information after having received too much of it. This explanation is consistent with theories of bounded rationality, cognitive capacity and information overload, as well as economic theories of the marginal value of new information. This result not only implies limits to the value of more novel information and larger and more diverse networks, it also may suggest a causal relationship between access to novel information and performance. If causality ran in the other direction, such that the most productive employees were magnets for more novel information, we would expect to observe increasing (or at least constant) returns to scale, meaning employees should continue to receive more novel information as they become more productive.[40]

39 For novel information greater than the normalized mean, coefficients in revenue regressions are negative and significant (β_{FE} = -3340.33, $p < .05$; β_{RE} = - 3207.06, $p < .05$) and in completed projects regressions are negative, though not significant (β_{FE} = -.04, N.S.; β_{RE}=-.04, N.S.).

40 We note that this is not definitive evidence of the direction of causality. We could witness diminishing returns for instance if novel information is finite or if recruiters send novel information to all above average recruiters in the same relative proportions but neglect to send novel information to non-star recruiters. Other contingencies may also exist. However, we propose this result as one piece of evidence suggesting a causal relationship rather than claiming it as definitive. In ongoing work we are exploring identification strategies explicitly and encourage more of this type of work.

----- FIGURE 6 -----

In fixed effects models, which control for variation explained by unobserved, time invariant characteristics of individuals, a one unit increase in the amount of non-redundant information flowing to individuals is associated on average with just over $3,800 more revenue generated ($p < .01$), an extra one tenth of one project completed ($p < .01$), and 14 days shorter average project duration per person per month ($p < .01$), again supporting the argument that network diversity drives performance through its effects on access to non-redundant information (see Table 7). The relationship between non-redundant information and performance is non-linear in the case of all three performance metrics indicating diminishing marginal returns to more novel information. Network diversity is associated with faster average project completion even when controlling for access to novel information, while access to non-redundant information explains most of the variance in project completion and revenue generation in fixed effects models, indicating that small changes in individuals' networks over time have less of an effect on performance through non-information based mechanisms than does variation in networks across individuals.[41]

Discussion & Conclusions

We analyzed email topology and content data, detailed accounting records of project expertise, output and performance, and survey data on human capital and information seeking practices in an executive recruiting firm with fourteen offices across the United States. The results bring empirical evidence to bear on both the economic value of information and the information based mechanisms through which social networks generate social capital. There is evidence that diverse networks provide access to diverse

41 Given the core-periphery structure of the email network of this firm (displayed in Figure 3), we compared the effects of network diversity on performance for those employees physically located at the headquarters to those who worked in peripheral offices. Our estimates of pooled OLS regressions provide weak evidence that network diversity enhances performance on average, that being in a peripheral office reduces performance, and that the interaction effect of being in a peripheral office and having a diverse network is positive, implying the potential for network diversity to be even more important for the geographically isolated. We do not report these results in this paper due to space and focus considerations and because estimated relationships are not robust to panel data procedures given being in the periphery is a time invariant binary variable. However, these results indicate that future work on the importance of network diversity for the geographically isolated may be fruitful.

non-redundant information and that access to non-redundant information in turn predicts productivity and performance. There are also important nuances in how information advantages from social structure are created and appropriated. Access to diverse, novel information is driven not only by structural diversity, but also by the bandwidth of communication channels and the knowledge heterogeneity of individuals' contacts. A key insight is the need to incorporate within channel information diversity (as well as across channel diversity) into theory relating social structure to information access. Holding network diversity constant, greater channel bandwidth delivers more diverse novel information. But, as networks become more diverse, the bandwidth of communication channels narrows creating countervailing influences on access to non-redundant information. The relationship between network diversity and channel bandwidth seems to be driven not by relationship maintenance costs, but rather by the nature of the relationships developed in diverse networks. Although we do not directly observe these costs, if time and effort costs of relationship maintenance were driving reductions in channel bandwidth associated with more diverse networks, we would expect bandwidth to also decrease as network size increased – however, we find the opposite: channel bandwidth increases with network size, implying that there is something about the nature of unconstrained relationships (other than time and effort costs) that reduces the thickness of information flows between structurally diverse contacts – a finding consistent with prior evidence on the nature of weak tie relationships (Granovetter 1973, Hansen 1999, Uzzi 1999).

The estimated tradeoff between network diversity and channel bandwidth implies three contingencies governing the ability of diverse networks to deliver diverse novel information. Although we cannot robustly test these contingencies in our setting (because we observe only one network and one set of information), our analytical model implies that diverse networks should provide more novel information (a) when the space of topics discussed in a social network is large, (b) when the distribution of topics over nodes is non-uniform, and (c) when information in the network changes frequently. These testable predictions follow from our analytical model and provide a theoretical starting point for future work. We observe some initial evidence of the first contingency – the importance of the size of the topic space – in our data. Business at this firm is cyclical. As demand picks up sharply in January and declines through the

next eight months, more emails are sent and received (Jan. = 166, Feb. = 139; *t-statistic* = 1.23, p < .11) and there is more total average non-redundant information present in the network (Jan. = 55.47, Feb. = 45.74; *t-statistic* 1.34, p < .10) in January 2003 than in December 2003, indicating a larger topic space in the first month of the year.[42] If diverse networks provide more novel information when the topic space in a network is large, then we should see a stronger relationship between network diversity and access to diverse novel information in January than in December. Although our data are limited when restricted to observations in these two months, we find that in January the relationship between network diversity and incoming information diversity is positive and significant (β_{Jan} = .46, S.E. = .15, p < .01), whereas in December the relationship is not discernable from zero (β_{Dec} = - .186, S.E. = .19, N.S.). Given this preliminary evidence, we encourage further investigation of these contingencies.

Network diversity is a stronger predictor of access to diverse novel information than the knowledge heterogeneity of contacts, which have been the primary focus of prior research examining 'network content' (e.g. Burt 2000, Rodan & Galunic 2004). We speculate that network diversity better predicts information flow because unlike the spread of disease, information diffusion is a discretionary process in which individuals make choices about the types of information they distribute among their contacts. The social nature of information sharing is influenced in part by social structure – meaning the connections between our contacts inform our decisions about the information we share. The relationship between social structure and social choices concerning information sharing make the diversity of communication networks a better predictor of access to information than the heterogeneity of the knowledge with which our contacts are endowed. A connection to an individual with a certain knowledge endowment affords the possibility of receiving that information but by no means guarantees it. The similarities of the subsets of information endowments that are shared are a function not only of the similarity of the knowledge and prior expertise of contacts but also of their similarities in current activity, purpose, task and role, which are reflected in cohesive social structure.

42 Statistics are reported per employee.

Network size is a concave predictor of access to diverse novel information in part because network diversity is increasing in network size, but with diminishing returns. As actors establish relationships with a finite set of possible contacts in an organization, the probability that a marginal relationship will be non-redundant, and provide access to novel information, decreases as possible alters in the network are exhausted. These findings establish limits to the value of building larger and larger networks. The marginal benefit of adding contacts to a network is decreasing in network size because network diversity is bounded when the number of potential contacts is finite, as is the case in organizational networks. We expect the marginal value of additional contacts declines more rapidly in smaller organizations and in settings where diversity is limited by other environmental constraints such as a high degree of collocation.

We also find diminishing marginal productivity returns to novel information, a result consistent with anecdotal evidence of information overload, and theories of bounded rationality, limits to cognitive capacity and economic decision making. The value of additional information should decline in the amount of information individuals receive because there are constraints on our ability to process and act on more information. From the standpoint of economic theory, the marginal value of novel information should decline due to decision irrelevance, belief convergence and declining complementarities amongst disparate ideas. Future empirical work may uncover how much diversity is too much or the points at which additional novel information provides no additional value. We suspect that while the basic shape of the relationship between novel information and performance generalizes relatively broadly, the specific points at which the value of additional contacts or information begins to decline will depend on the setting under investigation, including the nature of the task, the size of the population and the characteristics of the information being shared, among other things. In the case of executive recruiting, diverse information is valuable because the quality of the match between a candidate and an open position is a function of the multitude of different options that are considered and because recruiters must fill different types of positions over time, necessitating access to different types of information on possible options. These two points emerged in our interviews with recruiters. We suspect that in some contexts information diversity

could have a negligible or even negative correspondence with performance. For example, in situations where the task (or decision space) is narrow or where precision and therefore redundancy are more important than breadth, the relationship between diverse information and performance could be reversed. Establishing what types of work benefit most from diverse information and the contingencies that govern these relationships can help shape theory describing when and how diverse networks enable social capital development.

In our context, network diversity contributes to performance even when controlling for the positive performance effects of access to novel information, suggesting additional benefits to network diversity beyond those conferred through information advantage. This finding leaves room for other mechanisms linking network diversity to social capital and performance. Network contacts may provide non-information based resources (e.g. Podolny & Barron 1997) such as power or control (e.g. Burt 1992), favorable trading relationships (e.g. Emerson 1962) or benefits from informal reciprocity (e.g. Cook, Emerson & Gilmore 1983). Surprisingly, traditional demographic and human capital variables (e.g. age, gender, industry experience, education) have little effect on access to diverse information, highlighting the importance of network structure for information advantage.

These results represent some of the first evidence on the relationship between network structure and the information content flowing in networks. But, relationships between social structure, information access and economic outcomes are subtle and complex and require more detailed theoretical development and empirical inquiry across different contexts. Our methods for analyzing network structure and information content in email data are replicable, opening a new line of inquiry into the relationships among networks, information and economic performance.

References

Adler, P. S., & Kwon, S. W. 2002. "Social capital: Prospects for a new concept." *Academy of Management Review*, 27(1): 17–40.

Ancona, D.G. & Caldwell, D.F. 1992. "Demography & Design: Predictors of new Product Team Performance." *Organization Science*, 3(3): 321-341.

Aral, S., Brynjolfsson, E., & Van Alstyne, M. 2006. "Information, Technology and Information Worker Productivity: Task Level Evidence." *Proceedings of the 27th Annual International Conference on Information Systems*, Milwaukee, Wisconsin.

Aral, S., Brynjolfsson, E., & Van Alstyne, M. 2007. "Productivity Effects of Information Diffusion in Networks." *Proceedings of the 28th Annual International Conference on Information Systems*, Montreal, CA.

Argote, L. 1999. *Organizational Learning: Creating, Retraining & Transferring Knowledge*. Kluwer Academic, Boston, MA.

Arrow, K.J. 1985. "Informational Structure of the Firm." AEA Papers and Proceedings, 75(2): 303-307.

Baker, Wayne E. 1984. "The Social Structure of a National Securities Market." *American Journal of Sociology* 89:775-811.

Baker, W. 1990. "Market Networks & Corporate Behavior." *American Journal of Sociology* 96:589-625.

Barron, R.M. & D.A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic & Statistical Considerations." Journal of Personality & Social Psychology, (51:6): 1173-1182.

Baum, J.A.C., & Oliver, C. 1992. "Institutional Embeddedness & the Dynamics of Organizational Populations. *American Sociological Review*, (57:4): 540-559.

Bernard, H.R., Killworth, P., & Sailor, L. 1981. "Summary of research on informant accuracy in network data and the reverse small world problem." *Connections*, (4:2): 11-25.

Borgatti, S. & M. Everett. 1993. "Two Algorithms for Computing Regular Equivalence." *Social Networks*, 15: 361-376.

Bulkley, N. & Van Alstyne, M. 2004. "Why Information Influence Should Productivity" *The Network Society: A Global Perspective*; Manuel Castells (ed.). Edward Elgar Publishers. pp: 145-173.

Burgelman, R.A. 1991. "Intraorganizational Ecology of Strategy Making & Organizational Adaptation: Theory & Field Research." *Organization Science*, (2:3):239-261.

Burt, R. 1987. "Social Contagion & Innovation: Cohesion versus Structural Equivalence." *American Journal of Sociology*, 92: 1287-1335.

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

Burt, R. 2000. "The network structure of social capital" In B. Staw, & Sutton, R. (Ed.), *Research in organizational behavior* (Vol. 22). New York, NY, JAI Press.

Burt, R. 2004a. "Structural Holes & Good Ideas" *American Journal of Sociology*, (110): 349-99.

Burt, R. 2004b. "Where to get a good idea: Steal it outside your group." As quoted by Michael Erard in *The New York Times*, May.

Burt, R. 2005. Brokerage & Closure: An Introduction to Social Capital. Oxford University Press. New York, NY.

Caro, R. 1982. *The Path to Power*. Alfred A. Knopf. New York, NY.

Centola, D., & Macy, M. (Forthcoming). "Complex Contagions & the Weakness of Long Ties." *American Journal of Sociology.*

Coleman, J.S., Katz, E., Menzel, H. 1966. *Medical Innovation*. Bobbs-Merrill, New York, NY.

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

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

Cook, K.S., Emerson, R.M., Gilmore, M.R., & Yamagishi, T. 1983. “The distribution of power in exchange networks.” *American Journal of Sociology*, 89: 275-305.

Cummings, J., & Cross, R. 2003. “Structural properties of work groups and their consequences for performance.” Social Networks, 25(3):197-210.

Cummings, J. 2004. “Work groups, structural diversity, and knowledge sharing in a global organization." *Management Science*, 50(3), 352-364.

De Souza Briggs, X. 1997. “Social capital and the Cities: Advice to change agents.” *National Civic Review*, 86, 2: 111-117.

Durkheim, E. (1893) 1933. *The Division of Labor in Society*, Free Press, New York, NY.

Durkheim, E. (1897) 1951. *Suicide*, Free Press, New York, NY.

Emerson, R. 1962. “Power-Dependence Relations.” *American Sociological Review*, 27: 31-41.

Finlay, W. & Coverdill, J.E. 2000. “Risk, Opportunism & Structural Holes: How headhunters manage clients and earn fees.” *Work & Occupations*, (27): 377-405.

Fleming, L., Mingo, S., & D. Chen. 2007. “Collaborative Brokerage, Generative Creativity & Creative Success.” *Administrative Science Quarterly* (52:3): 443-475.

Friedkin, N.E. 1984. “Structural Cohesion & Equivalence Explanations of Social Homogeneity.” *Sociological Methods & Research* 12: 235-261.

Galbraith, J.R. 1974. *Organizational Design: An Information Processing View*. M. Wiener.

Granovetter, M. 1973. “The strength of weak ties.” *American Journal of Sociology* (78):1360-80.

Granovetter, M. 1978. “Threshold models of collective behavior.” *American Journal of Sociology* (83:6):1420-1443.

Granovetter, M. 1985. “Economic Action & Social Structure: The Problem of Embeddedness.” *American Journal of Sociology* (91):1420-1443.

Granovetter, M. 1992. “Problems of Explanation in Economic Sociology.” In N. Nohria & R.G. Eccles (eds.), *Networks & Organizations*: 25-56. Harvard Business School Press, Boston.

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

Hansen, M. 2002. "Knowledge networks: Explaining effective knowledge sharing in multiunit companies." *Organization Science* (13:3): 232-248.

Hargadon, A. & R, Sutton. 1997. “Technology brokering and innovation in a product development firm.” *Administrative Science Quarterly*, (42): 716-49.

Hayek, F. 1945. “The Use of Knowledge in Society.” *American Economic Review*, 35: 519-530.

Hirshleifer, J. 1973. Where are we in the theory of information? *American Economic Review*, (63): 31-39.

Lappe, F.M., & Du Bois, P.M. 1997. “Building social capital without looking backward.” National Civic Review, 86: 119-128.

Lazarsfeld, P., Berelson, B., H. Gaudet. 1944. *The People's Choice*. Columbia University Press, New York, NY.

Lazer, D. & Friedman, A. 2005. *The Parable of the Hare and the Tortoise: Small Worlds, Diversity & System Performance*. JFK School of Government, Havard University, Working Paper RWP 05058.

Kempe, D., Kleinberg, J., & E. Tardos. 2003. "Maximizing the spread of influence through a social network" *Proceedings of the 9th ACM SIGKDD*, Washington, D.C.: 137-146.

Krackhardt, D. & Kilduff, M. 1999. "Whether close or far: Social distance effects on perceived balance in friendship networks." *Journal of personality and social psychology*, (76) 770-82.

Kossinets, G. & D. Watts. 2006. "Empirical Analysis of an Evolving Social Network." *Science* (311:5757): 88-90.

Kumbasar, E., Romney, A.K., and Batchelder, W.H. 1994. "Systematic biases in social perception." *American Journal of Sociology*, (100): 477-505.

Marschak, J., & R. Radner. 1972. *Economic Theory of Teams*. Yale University Press, New Haven, CT.

Marsden, P. 1990. "Network Data & Measurement." *Annual Review of Sociology* (16): 435-463.

McPherson, M., L. Smith-Lovin & J. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415-444.

Narayan, D. & L. Pritchett. 1997. "Cents and sociability: Household income and social capital in rural Tanzania." *Social Development and Development Research Group, Poverty and Human Resources*. World Bank, Washington, D.C.

Padgett, J.F., & C.K. Ansell. 1993. "Robust Action & the Rise of the Medici." *American Journal of Sociology*, (98:6): 1259-1319.

Podolny, J. 1993. "A Status-Based Model of Market Competition." *American Journal of Sociology*, (98:4): 829-872.

Podolny, J. 2001. "Networks as the Pipes and Prisms of the Market." *American Journal of Sociology*, (107:1): 33-60.

Podolny, J., Baron, J. 1997. "Resources and relationships: Social networks and mobility in the workplace." *American Sociological Review* (62:5): 673-693.

Reagans, R. & McEvily, B. 2003. "Network Structure & Knowledge Transfer: The Effects of Cohesion & Range." *Administrative Science Quarterly*, (48): 240-67.

Reagans, R. & Zuckerman, E. 2001. "Networks, diversity, and productivity: The social capital of corporate R&D teams." *Organization Science* (12:4): 502-517.

Reagans, R. & Zuckerman, E. 2006. "Why Knowledge Does Not Equal Power: The Network Redundancy Tradeoff" *Working Paper Sloan School of Management,* 2006, pp. 1-67.

Rodan, S. & D. Galunic. 2004. "More Than Network Structure: How Knowledge Heterogeneity Influences Managerial Performance & Innovativeness." *Strategic Management Journal* (25): 541-562.

Rogers, E. 1995. *The Diffusion of Innovation*. Free Press. New York, NY.

Salton, G., Wong, A., & Yang, C. S. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM*, 18(11): 613-620.

Schelling, T.C. 1978. *Micromotives & Macrobehavior*. George J. McLeod Ltd. Toronto, CA.

Simmel, G. (1922) 1955. *Conflict & the Web of Group Affiliation*. Free Press. New York, NY.

Simon, H. 1991. "Bounded Rationality & Organizational Learning." *Organization Science*. (2:1): 125-134.

Sparrowe, R., Liden, R., Wayne, S., & Kraimer, M. 2001. "Social networks and the performance of individuals and groups." *Academy of Management Journal*, 44(2): 316-325.

Szulanski, G. 1996. "Exploring internal stickiness: Impediments to the transfer of best practice within the firm." *Strategic Management Journal* (17): 27-43.

Tushman, M.L., & D.A. Nadler. 1978. "Information Processing as an Integrating Concept in Organizational Design." *Academy of Management Review*, 3: 613-624.

Uzzi, B. 1996. "The Sources and Consequences of Embeddedness for the Economic Performance of Organizations: The Network Effect." *American Sociological Review*, (61):674-98.

Uzzi, B. 1997. "Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness." *Administrative Science Quarterly*, 42: 35-67.

Van Alstyne, M. & Brynjolfsson, E "Global Village or CyberBalkans? Measuring and Modeling the Integration of Electronic Communities". *Management Science;* 51 (6) (June 2005): pp. 851-868

Van Alstyne, M. & Zhang, J. 2003. "EmailNet: A System for Automatically Mining Social Networks from Organizational Email Communication," NAACSOS.

Watts, D. & S. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature*, 393: 440-442.

Watts, D. 1999. "Networks, Dynamics and the Small World Phenomenon." *American Journal of Sociology*, 105(2):493-527.

White, H. 1980. "A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity." *Econometrica* (48:4): 817-838.

Wu, F., Huberman, B., Adamic, L., & J. Tyler. 2004. "Information Flow in Social Groups." *Physica A*, 337: 327-335.

Tables

Table 1. Summary of Hypotheses

Domain	Hypothesis	Hypothesized Relationship
Network Structure	**H1**	*Network size and network diversity are positively associated with receiving more diverse information and more total non-redundant information.*
Knowledge Heterogeneity	**H2a**	*Network diversity is positively associated with the knowledge heterogeneity of actors' contacts.*
	H2b	*The knowledge heterogeneity of an actor's contacts is positively associated with receiving more diverse information and more total non-redundant information.*
Channel Bandwidth	**H3a**	*Network diversity is associated with lower channel bandwidth.*
	H3b	*Channel bandwidth is positively associated with receiving more diverse information and more total non-redundant information.*
Concavity of Information Benefits to Network Size	**H4a**	*The marginal increase in information diversity is decreasing in network size.*
	H4b	*The marginal increase in network diversity is decreasing in network size*.
Performance	**H5a**	*Access to non-redundant and diverse information is positively associated with recruiters' productivity and performance.*
	H5b	*The marginal increase in productivity and performance is decreasing in the amount of novel information actors receive.*
	H5c	*Network diversity is positively associated with performance, controlling for access to novel information.*

Table 2: Descriptive Statistics

Variable	Obs.	Mean	SD	Min	Max
Age	522	42.36	10.94	24	67
Gender (1=male)	657	.56	.50	0	1
Industry Experience	522	12.52	9.52	1	39
Years Education	522	17.66	1.33	15	21
Total Incoming Emails	563	80.31	59.67	0	342
Information Diversity	563	.57	.14	0	.87
Total Non-Redundant Information	563	47.94	35.97	0	223.30
Network Size	563	16.81	8.79	1	58
Structural Holes	563	.71	.17	0	.91
Structural Equivalence	563	77.25	16.32	27.35	175.86
Knowledge Heterogeneity	560	.86	.07	.51	.97
Channel Bandwidth	555	5.87	4.13	0	51
Revenue	630	20962.03	18843.16	0	80808.41
Completed Projects	630	.39	.36	0	1.69
Average Project Duration (Days)	630	225.23	165.77	0	921.04

Table 3: Pair Wise Correlations Between Independent Variables

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Age	1.00														
2. Gender (1=male)	.11*	1.00													
3. Industry Experience	.73*	.20*	1.00												
4. Years Education	.38*	.06	.15*	1.00											
5. Total Incoming Email	-.33*	-.10*	-.28*	-.15*	1.00										
6. Information Diversity	.09	.05	.16*	.05	.29*	1.00									
7. Non-redundant Information	-.32*	-.09*	-.27*	-.12*	.98*	.36*	1.00								
8. Network Size	-.07	.02	-.01	.09	.63*	.45*	.64*	1.00							
9. Network Diversity	.12*	.02	.25*	.01	.34*	.71*	.35*	.62*	1.00						
10. Structural Equivalence	-.19*	-.06	-.24*	-.06	.23*	-.08	.23*	-.05	-.16*	1.00					
11. Knowledge Heterogeneity	.11*	.20*	.27*	.12*	.03	.23*	.04	.38*	.46*	-.21*	1.00				
12. Channel Bandwidth	-.24*	-.10*	-.24*	-.20*	.19*	.52*	.50*	-.02	-.02	.29*	-.25*	1.00			
13. Revenue	.44*	-.02	.33*	.15*	-.09*	.23*	-.12*	-.12*	.27*	-.16*	.12*	-.05	1.00		
14. Completed Projects	.41*	-.01	.29*	.11*	-.09*	.23*	-.11*	-.09*	.25*	-.14*	.10*	-.07	.92*	1.00	
15. Average Project Duration	.50*	.12*	.49*	.21*	-.30*	.14*	-.31*	-.07	.18*	-.21*	.07	-.14*	.54*	.47*	1.00

* $p < .05$

Table 4. Network Structure & Access to Diverse, Novel Information: Panel Data Estimates

Dependent Variable:	***Information Diversity***						***Non-Redundant Information***					
Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Specification	*FE*	*FE*	*FE*	*FE*	*RE*	*RE*	*FE*	*FE*	*FE*	*FE*	*RE*	*RE*
Age						-.004						-.004
						(.008)						(.008)
Gender						.147						-.053
						(.096)						(.105)
Education						-.013						-.045
						(.041)						(.045)
Industry Experience						.002						-.008
						(.009)						(.009)
Partner						.161						-.269
						(.197)						(.214)
Consultant						.159						-.293*
						(.149)						(.162)
Total Email Incoming	.002**	-.000	-.001	-.002**	-.002***	-.001						
	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)						
Knowledge Heterogeneity	.281***	.123***	.028	.037	-.023	-.025	.181***	.099*	-.045	-.011	-.063*	-.060
	(.051)	(.052)	(.041)	(.042)	(.035)	(.041)	(.044)	(.052)	(.047)	(.042)	(.036)	(.042)
Network Size		1.176***	.439***	.505***	.612***	.587***			.733***	.668***	.746***	.834***
		(.144)	(.117)	(.118)	(.105)	(.122)			(.131)	(.112)	(.102)	(.119)
Network Size-Squared		-.815***	-.250***	-.259***	-.361***	-.335***			-.120	-.081	-.108	-.182*
		(.115)	(.090)	(.089)	(.083)	(.092)			(.104)	(.089)	(.083)	(.094)
Network Diversity			.145***	.155***	.230***	.162***		.206***	-.048	.066	.032	.022
			(.055)	(.055)	(.051)	(.063)		(.067)	(.063)	(.054)	(.052)	(.064)
Structural Equivalence			.019	.021	.028	.060		-.032	.027	-.007	.025	-.014
			(.034)	(.034)	(.031)	(.038)		(.045)	(.039)	(.034)	(.032)	(.039)
Channel Bandwidth				.085***	.099***	.067**				.352***	.377***	.360***
				(.031)	(.031)	(.033)				(.026)	(.025)	(.029)
Constant	-.223**	.067	.064	.457***	.442***	.207	-.172**	.371***	.189***	-.101	.112	1.425†
	(.105)	(.113)	(.089)	(.085)	(.092)	(.700)	(.076)	(.081)	(.071)	(.062)	(.075)	(.745)
Temporal Controls	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year
F-Value / Wald χ^2	7.51***	12.75***	5.43***	5.54***	148.9***	108.3***	11.73***	10.02***	23.13***	41.75***	755.5***	646.6***
(d.f.)	(10)	(12)	(14)	(15)	(15)	(21)	(9)	(11)	(13)	(14)	(14)	(20)
R^2	.14	.24	.14	.16	.27	.21	.18	.19	.40	.56	.65	.68
Observations	556	556	538	535	535	429	556	538	538	535	535	429

Hausman Test Results (RE Consistent & Efficient - Models 4 & 5): $\chi^2 = 24.33$**, $p < .05$; Hausman Test Results (RE Consistent & Efficient - Models 10 & 11): $\chi^2 = 36.36$***, $p < .01$. Note: * p <.10; ** p <.05; *** p <.01.

Table 5. Information Advantage Mechanisms: Relationships Between Network Size, Network Diversity & Channel Bandwidth

Dependent Variable:	***Channel Bandwidth***					***Network Diversity***			
Model:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Specification:	*FE*	*FE*	*FE*	*RE*	*RE*	*FE*	*FE*	*RE*	*RE*
Age					-.006				-.003
					(.011)				(.006)
Gender					-.113				-.183**
					(.136)				(.079)
Education					-.123**				-.048
					(.059)				(.034)
Industry Experience					-.014				.018***
					(.012)				(.007)
Partner					.145				-.006
					(.284)				(.158)
Consultant					-.231				.107
					(.217)				(.118)
Network Diversity	-.314***	-.288***	-.335***	-.286***	-.190*				
	(.083)	(.089)	(.097)	(.089)	(.111)				
Structural Equivalence	.107*	.101*	.105*	.167***	.149**				
	(.057)	(.059)	(.060)	(.054)	(.068)				
Knowledge Heterogeneity		-.074	-.095	-.209***	-.141**	.384***	.185***	.189***	.126***
		(.072)	(.075)	(.058)	(.068)	(.044)	(.043)	(.033)	(.037)
Network Size			.213	.476***	.398**		1.229***	1.443***	1.369***
			(.203)	(.171)	(.199)		(.115)	(.087)	(.089)
Network Size-Squared			-.123	-.345**	-.333**		-.812***	-1.002***	-.885***
			(.160)	(.142)	(.161)		(.096)	(.079)	(.080)
Constant	.062	.060	.072	.105	2.793***	-.127*	.020	-.158**	.702
	(.109)	(.110)	(.111)	(.124)	(.972)	(.075)	(.069)	(.074)	(.557)
Temporal Controls	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year
F-Value / Wald χ^2	4.52***	4.19***	3.65***	74.63***	76.63***	18.55***	30.48***	622.29***	541.42***
(d.f.)	(10)	(11)	(13)	(13)	(19)	(9)	(11)	(11)	(17)
R^2	.09	.09	.10	.20	.24	.26	.41	.61	.62
Observations	536	535	535	535	429	556	556	556	442

Hausman Test Results (RE Consistent & Efficient - Models 3 & 4): 21.23*, p < .10; Hausman Test Results (RE Consistent & Efficient - Models 7 & 8): 411.38***, p < .01. Note: * p <.10; ** p <.05; *** p <.01.

Table 6. Network Diversity, Non-Redundant Information and Individual Performance - Random Effects Estimates

Dependent Variable:	***Project Duration***				***Completed Projects***				***Revenue***			
Model:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Specification	*RE*	*RE*	*RE*	*RE*	*RE*	*RE*	*RE*	*RE*	*RE*	*RE*	*RE*	*RE*
Age				-.124				-.006				-257.49
				(2.56)				(.006)				(297.91)
Gender (Male=1)				-2.33				-.117				-7583.31**
				(32.96)				(.081)				(3841.14)
Education				10.34				-.017				-14.84.2
				(13.52)				(.033)				(1584.94)
Industry Experience				3.50				.003				26.64
				(2.39)				(.006)				(280.04)
Partner				-17.20				.239				18811.08**
				(75.86)				(.191)				(9124.73)
Consultant				-56.69				.286*				12761.9*
				(60.89)				(.154)				(7351.83)
Network Diversity	-12.61***	-10.91***	-9.13**	-9.46*	.054***	.039**	.019	.004	2276.6***	1656.50*	656.15	-407.21
	(4.07)	(4.08)	(4.14)	(4.97)	(.017)	(.017)	(.016)	(.020)	(856.13)	(852.79)	(851.49)	(1050.23)
Non-Redundant Information		-14.67***	-20.03***	-23.23***		.104***	.152***	.189***		4678.81***	6984.64***	8974.12***
		(5.33)	(5.85)	(6.91)		(.021)	(.023)	(.027)		(111.20)	(1168.87)	(1421.87)
Non-Redundant Information Squared			16.44**	13.53			-.166***	-.213***			-8249.0***	-10396***
			(7.59)	(9.85)			(.031)	(.041)			(1619.59)	(2112.13)
Constant	289.16***	282.08***	282.87***	75.61	.589***	.639***	.629***	1.031*	31620***	33891.6***	33340.6***	64041.51**
	(14.17)	(14.27)	(14.30)	(239.66)	(.039)	(.040)	(.038)	(.591)	(1933.77)	(1963.95)	(1903.88)	(28019.77)
Temporal Controls	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year
Wald χ^2 (d.f.)	24.53***	32.47***	37.53***	44.24***	20.93***	45.16***	76.83***	81.69***	15.38*	33.83***	62.01***	72.89***
	(9)	(10)	(11)	(17)	(9)	(10)	(11)	(17)	(9)	(10)	(11)	(17)
R^2	.004	.03	.03	.25	.07	.14	.20	.31	.07	.14	.20	.29
Obs.	420	420	420	320	420	420	420	320	420	420	420	320

Note: * p <.10; ** p <.05; *** p <.01.

Table 7. Network Diversity, Non-Redundant Information and Individual Performance - Fixed Effects Estimates

	Project Duration			Project Completions			Revenue		
Model:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Specification	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*
Network Diversity	-14.23*** (4.18)	-12.74*** (4.19)	-11.03*** (-19.94)	.032* (.018)	.022 (.018)	.005 (.018)	565.33 (934.08)	165.14 (931.51)	-564.46 (926.08)
Non-Redundant Information		-14.21*** (5.44)	-19.95*** (6.02)		.097*** (.023)	.150*** (.025)		3806.12*** (1211.06)	6258.47*** (1316.03)
Non-Redundant Information Squared			16.61** (7.64)			-.154*** (.032)			-7105.14*** (1669.37)
Constant	295.10*** (6.08)	288.93*** (6.48)	290.10*** (6.47)	.617*** (.027)	.659*** (.028)	.649*** (.027)	33585.03*** (1359.91)	35238.48*** (1442.79)	34736.2*** (1414.51)
Temporal Controls	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year
F-Value (d.f.)	2.89*** (9)	3.32*** (10)	3.48*** (11)	1.54 (9)	3.15*** (10)	5.11*** (11)	1.27 (9)	2.16** (10)	3.70*** (11)
R^2	.07	.09	.10	.04	.08	.14	.03	.06	.10
Obs.	420	420	420	420	420	420	420	420	420

Hausman Test Results (Tables 6 & 7, Model 3): 3.69, N.S.; Hausman Test Results (Tables 6 & 7, Model 6): 13.27, N.S.; Hausman Test Results (Tables 6 & 7, Model 9): 29.20***, p < .01. Note: * p <.10; ** p <.05; *** p <.01.

Figures

Figure 1. Logic of the Information Advantage argument.

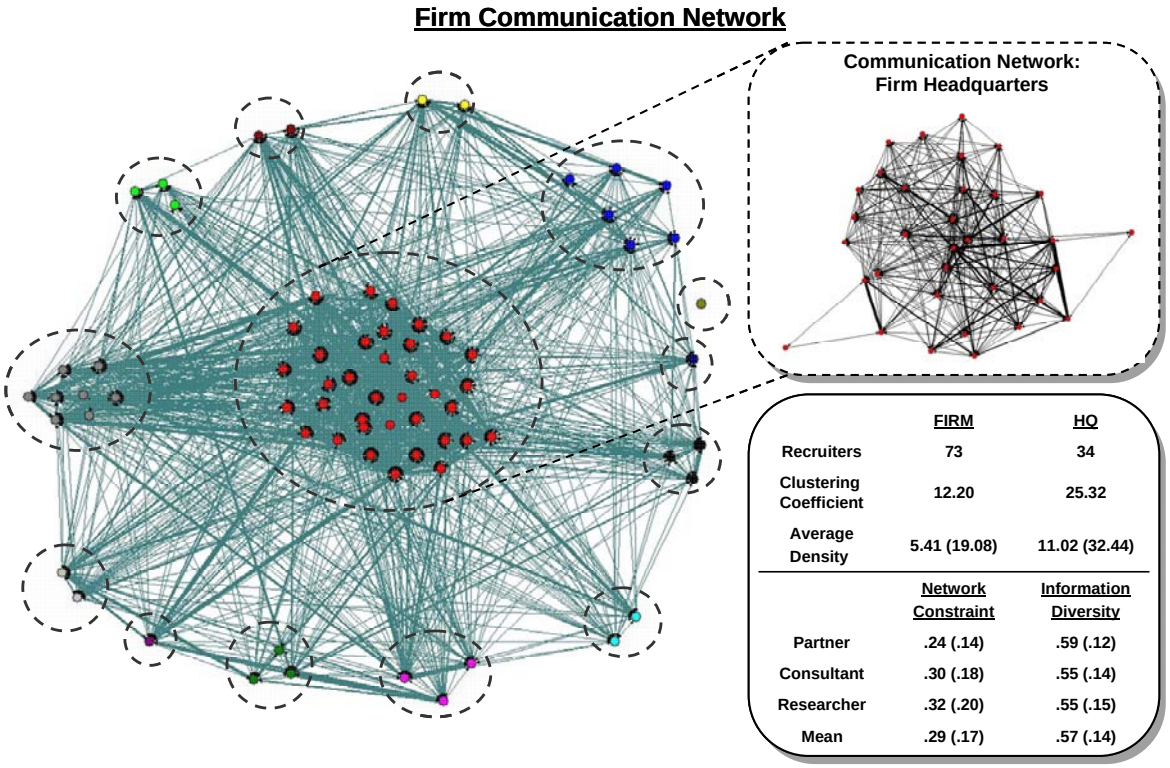


Figure 2. The email network of the firm displays a hub and spoke structure, with a dense core in the firm headquarters and spokes in various offices located across the U.S.

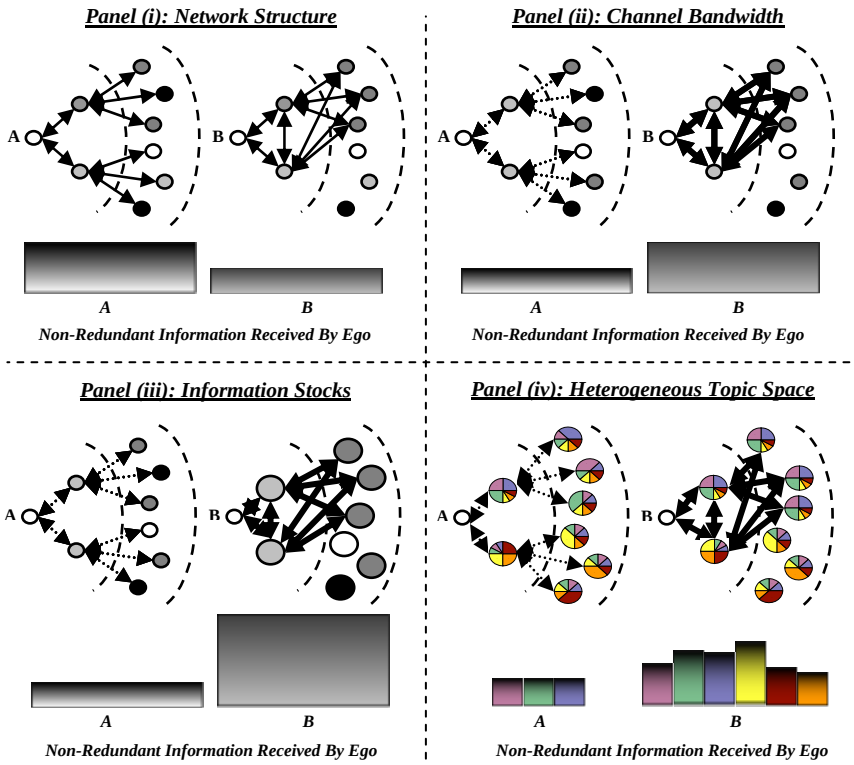


Figure 3. Our theory and analytical model propose that information access is a function not only of network structure (Panel (i)), but also of relationship channel bandwidth (Panel (ii)), and knowledge stocks (Panel (iii)) across

heterogeneous topics (Panel (iv)). Panel (ii) depicts that actor B may receive more total non-redundant information in a cohesive network due to greater channel bandwidth. Panel (iv) shows that B may receive more total non-redundant information on a greater number of topics due to greater channel bandwidth even though A's contacts have greater diversity in their information stocks.

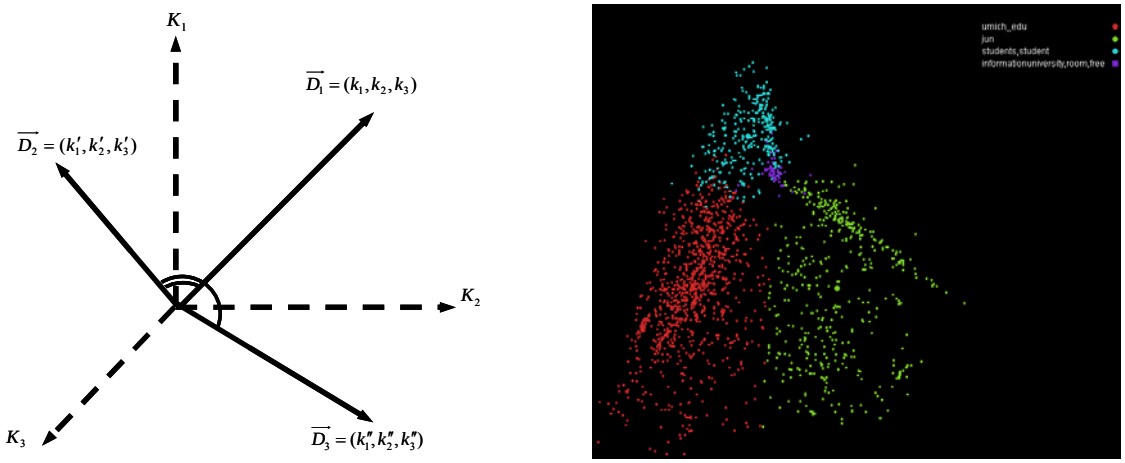


Figure 4. A three dimensional Vector Space Model of three documents is shown on the left. A Vector Space Model containing a test inbox with emails clustered along three dimensions is shown on the right.

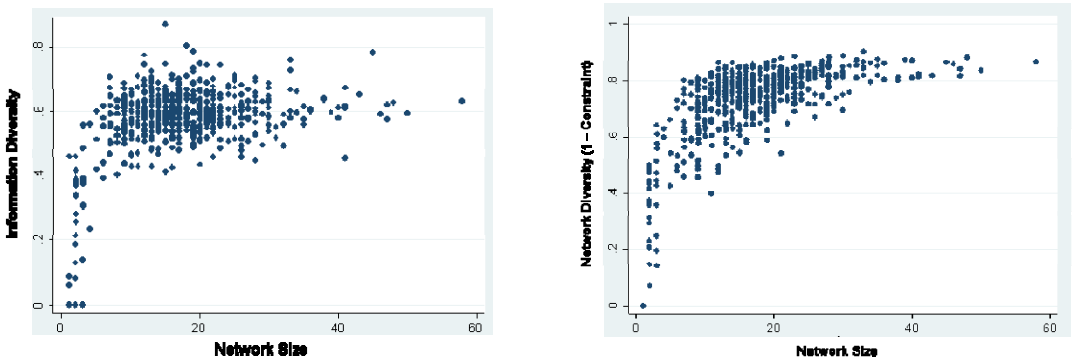


Figure 5. Graphs showing relationships between network size, network diversity and information diversity.

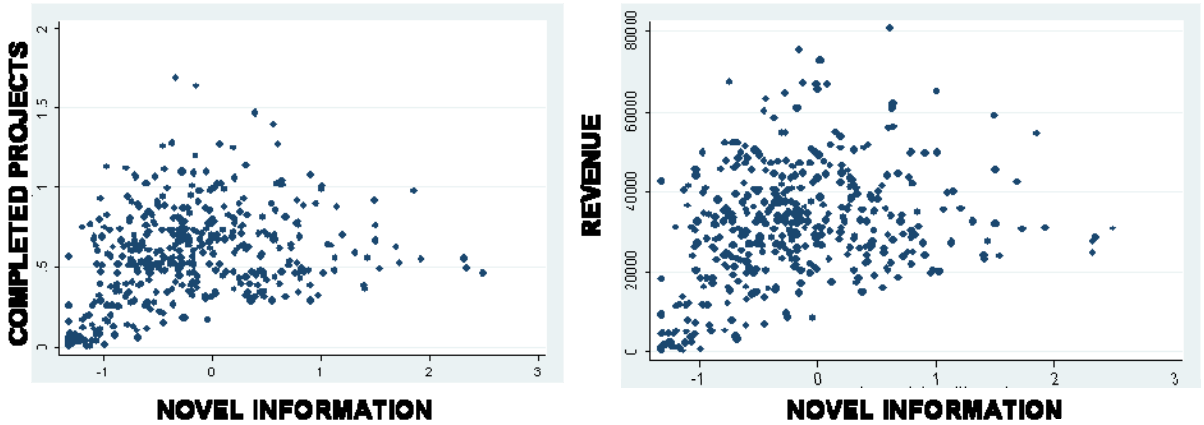


Figure 6. Graphs of the relationships between novel information, completed projects and revenue.

Appendix A. Model Derivation

This short section provides the derivation for Equation 1. Let there be 1 … n_1 topics in topic set n_1 and 1 … n_2 topics in topic set n_2 for a total of $n_1+n_2 = T$. Define the likelihoods of encountering n_1 and n_2 topics as p_1 and p_2 respectively. It follows that $n_1p_1 + n_2p_2 = 1$. Further, define the following:

$$I_{lk} = 1 \text{ if link } l \text{ connects to idea } k, \text{ 0 otherwise.}$$

$$J_k = \begin{Bmatrix} 1 & if & \sum_{l=1}^{L} I_{lk} = 0 \\ 0 & otherwise & \end{Bmatrix}$$

Ψ= {Event that link L+1 connects to a new idea}

Here, J_k indicates whether idea k has failed to appear among the information provided by any of the links 1 … L. With this terminology, we can now derive $P(\Psi)$, the probability of encountering a new idea given that there are k ideas remaining to be seen.

$$\begin{aligned} P(\Psi) &= E[P(\Psi \mid J_1 \ldots J_k)] \\ &= E[\sum_{i=1}^{n_1} J_i p_1 + \sum_{h=n_1+1}^{T} J_h p_2] \\ &= n_1 p_1 E[J_i] + n_2 p_2 E[J_h] \\ &= n_1 p_1 (1-p_1)^L + n_2 p_2 (1-p_2)^L \end{aligned}$$

The last step arises because an idea that occurs with probability p must not have occurred in any of the previous L draws. This completes the derivation. It is useful to note three properties. First, having no prior links L=0 implies that a new idea is encountered with certainty. Second, increasing links without bound $L\to\infty$ implies the chances of encountering a new idea approach 0. Third, unbiased information implies $p_1= p_2$=1/T. Further, if ideas in n_1 become B times more likely to appear among in-group communications, then $p_1=B/T$ which implies that $p_2=\frac{1-n_1 B/T}{T-n_1}$ (with $n_1 < T$, $B<T$, and $n_1B \le T$) which simplifies Equation 4 in the main text.

Appendix B. Descriptions & Correlations of Information Diversity Metrics

1. Cosine Distance Variance

Variance based on cosine distance (cosine similarity):

$$ID_i^I = \frac{\sum_{j=1}^{N} \left(Cos\left(d_{ij}^I, M_i^I\right)\right)^2}{N}, \text{ where } Cos(d_{ij}, M) = \frac{d_i \bullet M_i}{|d_i||M_i|} = \frac{\sum_j w_{ij} \times w_{Mj}}{\sqrt{\sum w_{ij}^2}\sqrt{\sum w_{Mj}^2}}$$

We measure the variance of deviation of email topic vectors from the mean topics vector and average the deviation across emails in a given inbox or outbox. The distance measurement is derived from a well-known document similarity measure – the cosine similarity of two topic vectors.

2. Dice's Coefficient Variance

Variance based on Dice's Distance and Dice's Coefficient: $VarDice_i^I = \frac{\sum_{j=1}^{N} \left(DistDice\left(d_{ij}^I\right)\right)^2}{N}$, where

$$DistDice(d) = DiceDist(d, M) = 1 - Dice(d, M)$$, and where

$$Dice(D1, D2) = \frac{2\sum_{i=1}^{T}(t_{D1j} \times t_{D2j})}{\sum_{i=1}^{T} t_{D1j} + \sum_{i=1}^{T} t_{D2j}}$$

Similar to VarCos, variance is used to reflect the deviation of the topic vectors from the mean topic vector. Dice's coefficient is used as an alternative measure of the similarity of two email topic vectors.

3. Average Common Cluster

AvgCommon measures the level to which the documents in the document set reside in different k-means clusters produced by the eClassifier algorithm:

$$AvgCommon_i^I = \frac{\sum_{j=1}^{N}\left(CommonDist\left(d_{1j}^I, d_{2j}^I\right)\right)}{N},$$

where (d_{1j}^I, d_{2j}^I) represents a given pair of documents (1 and 2) in an inbox and j indexes all pairs of documents in an inbox, and where:

$$CommonDist\left(d_{1j}^I, d_{2j}^I\right) = 1 - CommonSim\left(d_{1j}^I, d_{2j}^I\right)$$

$$CommonSim\left(d_{1j}^I, d_{2j}^I\right) = \frac{\sum Iterations\_in\_same\_cluster}{\sum Iterations}$$

AvgCommon is derived from the concept that documents are similar if they are clustered together by k-means clustering and dissimilar if they are not clustered together. The k-means clustering procedure is repeated several times, creating several clustering results with 5, 10, 20, 30, 40 … 200 clusters. This measures counts the number of times during this iterative process two emails were clustered together divided by the number of clustering iterations. Therefore, every two emails in an inbox and outbox that are placed in separate clusters contribute to higher diversity values.

4. Average Common Cluster with Information Content

AvgCommonIC uses a measure of the "information content" of a cluster to weight in which different emails reside. AvgCommonIC extends the AvgCommon concept by compensating for the different amount of information provided in the fact that an email resides in the same bucket for either highly diverse or tightly clustered clusters. For example, the fact that two emails are both in a cluster with low intra-cluster diversity is likely to imply more similarity between the two emails than the fact that two emails reside in a cluster with high intra-cluster diversity.

$$CommonICSim(D_1, D_2) = \frac{1}{\log(\frac{1}{\|all\_documents\|})} \cdot \frac{\sum_{D_1, D_2 in\_same\_bucket} \log\left(\frac{\|documents\_in\_the\_bucket\|}{\|all\_documents\|}\right)}{total\_number\_of\_bucket\_levels}$$

$$CommonICDist(D_1, D_2) = 1 - CommonICSim(D_1, D_2)$$

$$AvgCommonIC = \underset{d_1, d_2 \in documents}{average}\ \{CommonICDist(d_1, d_2)\}$$

5. Average Cluster Distance

AvgBucDiff measures diversity using the similarity/distance between the clusters that contain the emails:

$$AvgBucDiff = \underset{d_1, d_2 \in documents}{average}\ \{DocBucDist(d_1, d_2)\}$$, where

$$DocBucketD\ ist(D_1, D_2) = \frac{1}{\|cluster\_iterations\|} \cdot \sum_{i \in cluster\_iterations}\left(BucketDist\ (B_{iteration=i, D_1}, B_{iteration=i, D_2})\right)$$, and:

$$BucketDist(B_1, B_2) = CosDist(m_{B_1}, m_{B_2}).$$

AvgBucDiff extends the concept of AvgCommon by using the similarity/distance between clusters. While AvgCommon only differentiates whether two emails are in the same cluster, AvgBucDiff also considers the distance between the clusters that contain the emails.

Correlations Between the Five Measures of Information Diversity

Measure	1	2	3	4	5
1. VarCosSim	1.0000				
2. VarDiceSim	0.9999	1.0000			
3. AvgCommon	0.9855	0.9845	1.0000		
4. AvgCommonIC	0.9943	0.9937	0.9973	1.0000	
5. AvgClusterDist	0.9790	0.9778	0.9993	0.9939	1.0000

Appendix C: External Validation of Diversity Measures

We validated our diversity measurement using an independent, publicly available corpus of documents from Wikipedia.org. Wikipedia.org, the user created online encyclopedia, stores entries according to a hierarchy of topics representing successively fine-grained classifications. For example, the page describing "genetic algorithms," is assigned to the "Genetic Algorithms" category, found under "Evolutionary Algorithms," "Machine Learning," "Artificial Intelligence," and subsequently under "Technology and Applied Sciences." This hierarchical structure enables us to construct clusters of entries on diverse and focused subjects and to test whether our diversity measurement can successfully characterize diverse and focused clusters accurately.

We created a range of high to low diversity clusters of Wikipedia entries by selecting entries from either the same sub-category in the topic hierarchy to create focused clusters, or from a diverse set of unrelated subtopics to create diverse clusters. For example, we created a minimum diversity cluster (Type-0) using a fixed number of documents from the same third level sub-category of the topic hierarchy, and a maximum diversity cluster (Type-9) using documents from unrelated third level sub-categories. We then constructed a series of document clusters (Type-0 to Type-9) ranging from low to high topic diversity from 291 individual entries as shown in Figure 3.[43] The topic hierarchy from which documents were selected appears at the end of this section.

If our measurement is robust, our diversity measures should identify Type-0 clusters as the least diverse and Type-9 clusters as the most diverse. We expect diversity will increase relatively monotonically from Type-0 to Type-9 clusters, although there could be debate for example about whether Type-4 clusters are more diverse than Type-3 clusters.[44] After creating this independent dataset, we used the Wikipedia entries to generate keywords and measure diversity using the methods described above. Our methods were very successful in characterizing diversity and ranking clusters from low to high diversity. Figure 3 displays cosine similarity metrics for Type-0 to Type-9 clusters using 30, 60, and 90 documents to populate clusters. All five diversity measures return increasing diversity scores for clusters selected from successively more diverse topics.[45] Overall, these results give us confidence in the

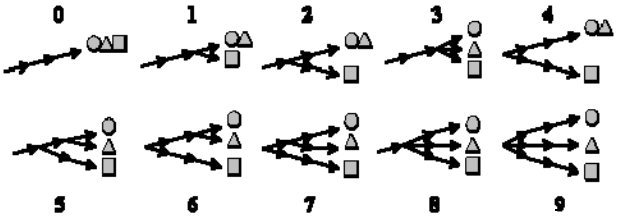
43 We created several sets of clusters for each type and averaged diversity scores for clusters of like type. We repeated the process using 3, 6 and 9 document samples per cluster type to control for the effects of the number of documents on diversity measures.

44 Whether Type-3 or Type-4 clusters are more diverse depends on whether the similarity of two documents in the same third level sub category is greater or less than the difference of similarities between documents in the same second level sub category as compared to documents in categories from the first hierarchical layer onwards. This is, to some extent, an empirical question.

45 The measures produce remarkably consistent diversity scores for each cluster type and the diversity scores increase relatively monotonically from Type-0 to Type-9 clusters. The diversity measures are not monotonically increasing for all successive sets, such as Type-4, and it is likely that the information contained in Type-4 clusters are less diverse than Type-3 clusters due simply to the fact that two Type-4 documents are taken from the same third level sub category.

ability of our diversity measurement to characterize the subject diversity of groups of text documents of varying sizes.

Document clusters selected from Wikipedia.org



Diversity measurement validation results

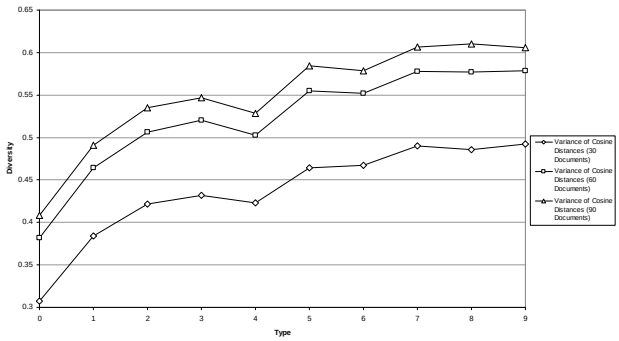


Figure C1. Wikipedia.org Document Clusters and Diversity Measurement Validation Results.

Wikipedia.org Categories

+ **Computer science >**	+ **Geography >**	+ **Technology >**
+ Artificial intelligence	+ Climate	+ Robotics
+ Machine learning	+ Climate change	+ Robots
+ Natural language processing	+ History of climate	+ Robotics competitions
+ Computer vision	+ Climate forcing	+ Engineering
+ Cryptography	+ Cartography	+ Electrical engineering
+ Theory of cryptography	+ Maps	+ Bioengineering
+ Cryptographic algorithms	+ Atlases	+ Chemical engineering
+ Cryptographic protocols	+ Navigation	+ Video and movie technology
+ Computer graphics	+ Exploration	+ Display technology
+ 3D computer graphics	+ Space exploration	+ Video codecs
+ Image processing	+ Exploration of Australia	+ Digital photography
+ Graphics cards		

Appendix D: Estimation Procedures

The fixed effects estimator uses variation within observations of a single individual over time. The basic specification includes observations of dependent and independent variables for each individual in each cross sectional time period *t*, and a time invariant vector of individual characteristics α_i representing unobserved heterogeneity across individuals:

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it} . \quad [10]$$

The fixed effects transformation, also called the within transformation, is obtained by first averaging equation 10 over $t = 1,\ldots, T$, to create the cross section equation, also called the between estimator:

$$\overline{y_i} = \alpha_i + \overline{x_i}\beta + \overline{\varepsilon_i} , \quad [11]$$

$$\text{where } \overline{y_i} = \frac{\sum_1^T y_{it}}{T}, \ \overline{x_i} = \frac{\sum_1^T x_{it}}{T} \text{ and } \overline{\varepsilon_i} = \frac{\sum_1^T \varepsilon_{it}}{T} .$$

By subtracting equation 11 from equation 10, the fixed effects transformation removes unobserved time invariant individual specific heterogeneity embodied in α_i :

$$y_{it} - \overline{y_i} = (x_{it} - \overline{x_i})\beta + \varepsilon_{it} - \overline{\varepsilon_i} \; . \qquad [12]$$

The fixed effects estimator produces estimates using variation within observations of the same individuals over time and allows us to estimate the effects of network structure controlling for unobserved omitted variables that could bias our estimates.

While the fixed effects estimator helps us estimate the effects of network structure on information access and performance controlling for unobservable omitted variables, it has several drawbacks. First, we are also interested in the effects of observable time invariant characteristics of individuals, such as demography (e.g. age, gender), human capital (e.g. education, industry tenure), and organizational hierarchy (e.g. individuals position in the firms formal organizational structure), on access to information and performance. More precisely, we are interested in the relative effects of network structure on information access and performance compared to these traditional factors. As the fixed effects estimator washes away variation in time invariant characteristics, it makes estimation of these parameters impossible. Second, we believe that variation across individuals also helps explain differences in information access and performance correlated with network structure. An individual may be able to manipulate the information they receive by changing their communication patterns over time, but persistent structural differences between individuals could also explain performance differentials. We therefore estimate both pooled OLS and random effects models of our specifications.

The OLS estimator on pooled data estimates an unweighted average of the within and between estimators. Although we do not report these results in the tables, we produced pooled OLS estimates of our specifications with very similar results, which most closely resembled the random effects estimates we report. We estimated the pooled OLS specifications with robust clustered standard errors in order to control for the fact that repeated observations of the same individuals over time in panel data may artificially constrict the standard errors. Clustered robust standard errors treat each individual as a super-observation for part of its contribution to the variance estimate (e.g. $\varepsilon_{ci} = \eta_c + \upsilon_{ci}$, where η_c is an individual effect and υ_{ci} the idiosyncratic error). They are robust to correlations within the observations of each individual, but are never fully efficient. They represent conservative estimates of standard errors.

When variables of interest do not vary much over time, fixed effects methods can produce imprecise estimates. In our case, we are not only interested in estimating the impact of time invariant characteristics of individuals on access to information and performance (e.g. age, gender, education), but we also know that certain aspects of network structure change relatively little over time. We therefore estimate both fixed effects and random effects specifications. The random effects model estimates a matrix weighted average of the between [11] and within [12] estimators where the weighting matrix λ accounts for correlation across observations in the residuals, as follows:

$$y_{it} - \overline{\lambda y_i} = (x_{it} - \overline{\lambda x_i})\beta + \varepsilon_{it} - \overline{\lambda \varepsilon_i} \; . \qquad [13]$$

We estimate λ as a function of the idiosyncratic error variance and the group specific error variance. When $\lambda = 0$, the procedure is equivalent to estimating OLS, and when $\lambda = 1$ we are estimating fixed effects. The random effects model brings efficiency gains and the ability to estimate parameters of time invariant covariates at the risk of inconsistency. To test the consistency of the random effects estimator, we conduct Hausman tests (Hausman 1978) comparing fixed and random effects models and report our results in the table notes for each set of results.