Networks, Information and Brokerage: The Diversity–Bandwidth Tradeoff

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

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

Forthcoming in the American Journal of Sociology

* We are grateful to Lada Adamic, Wayne Baker, Erik Brynjolfsson, Ron Burt, Paul Carlile, Emilio Castilla, Jerry Davis, Stine Grodal, Jon Kleinberg, 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, the London School of Economics and the University of Chicago for valuable comments, and to the National Science Foundation (Career Awards IIS-9876233 and IIS-0953832 and grant IIS-0085725), Microsoft, Cisco Systems, France Telecom and the MIT Center for Digital Business for generous funding. We are also greatly indebted to Tim Choe, Petch Manoharn, Lev Muchnik, Cyrus-Charles Weaver and Jun Zhang for their tireless research assistance.

Networks, Information and Brokerage: The Diversity–Bandwidth Trade-off

Abstract

We propose that a trade-off between network diversity and communications bandwidth regulates access to novel information. As the structural diversity of a network increases, the bandwidth of its communication channels decrease, creating countervailing effects on the receipt of novel information. This trade-off occurs because more diverse networks, presumed to provide more information novelty, typically contain weaker ties. Weaker ties imply fewer opportunities for interaction and less total information flow. Information advantages to brokerage then depend on (a) whether the information overlap among alters is small enough to justify bridging structural holes, (b) whether the size of the topic space known to alters is large enough to consistently provide novelty, and (c) whether the knowledge stock of alters refreshes enough over time to justify updating what was previously known. We test these arguments by combining social network and performance data with direct observation of the information content flowing through e-mail at a medium-sized executive recruiting firm. We find that brokers with bridging ties to disparate parts of a social network can have disadvantaged access to novel information because their lower bandwidth communication curbs the total volume of novelty they receive. These analyses suggest that information benefits to brokerage depend on the information environments in which brokers find themselves and that we should embrace a more nuanced view of how information flows in social networks. The methods developed serve as 'proof-of-concept' for using e-mail content data to analyze relationships among information flows, networks and social capital.

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

Where does one find novel information? Most modern sociological theory suggests that we find novelty through weak ties that span structural holes. A more precise question, however, is where does one find the most novel information per unit time? That is, at what rate do we receive novelty from our different social contacts? We should get information with greater novelty from across a structural hole but at a slower rate because interactions with bridging ties are weak, infrequent and lower bandwidth. On the other hand, we should get information with less novelty from a cohesive embedded tie but at a faster rate because the tie is stronger, the interaction more frequent and the bandwidth higher. Contrary to conventional wisdom, this stronger tie can in certain circumstances provide greater total novelty over time. Since strong high bandwidth ties are more likely in cohesive networks and weak low bandwidth ties more likely in sparse networks, the two factors affecting the rate at which we find novel information – structural diversity and channel bandwidth – are likely to trade off, creating countervailing effects on access to novel information. We develop a theory of this trade-off and the contingencies of social structure and information environments that affect access to novelty. We test this theory on observed information content flowing through organizational e-mail networks. Results suggest that information benefits to brokerage depend on the information environments in which brokers find themselves and that we should embrace a more nuanced view of how information flows in social networks.

THE DIVERSITY-BANDWIDTH TRADE-OFF

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 outcomes such as wages, job placement, promotion, creativity, innovation, political success, social support, productivity and performance (Simmel 1922 (1955), Moreno 1940, Granovetter 1973, Baker 1990, Burt 1992, 2004, Padgett and Ansell 1993, Uzzi 1996, 1997, Podolny 2001, Reagans and Zuckerman 2001, Hansen 1999, 2002, Aral et. al. 2006, 2007). The central argument in this body of theory is that structurally diverse networks – networks low in cohesion and structural equivalence and rich in structural holes – provide access to diverse, novel information.

Contacts maintained through weak ties are typically unconnected to other contacts and therefore more likely to "move in circles different from our own and thus [to] have access to information different from that which we receive…" (Granovetter 1973: 1371) These ties are "the channels through which ideas, influence, or information socially distant from ego may reach him." (Granovetter 1973: 1371) As Burt (1992: 16) argues, "[E]verything else constant, a large, diverse network is the best guarantee of having a contact present where useful information is aired…." Since information in local network neighborhoods tends to be redundant, structurally diverse contacts that reach across structural holes should provide channels through which novel information flows (Burt 1992).

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 and apply information to problems that are intractable given local knowledge (e.g. Hargadon and Sutton 1997, Reagans and Zuckerman 2001, Burt 2004a, Rodan and Gallunic 2004, Van Alstyne and Brynjolfsson 2005, Lazer and Friedman 2007). Access to novel information should increase the breadth of individuals' absorptive capacity, strengthen the ability to communicate ideas across a broader range of topics to a broader audience, and improve persuasion and the ability to generate broader support from subject matter experts (Cohen and Levinthal 1990, Simon 1991, Reagans and McEvily 2003, Rodan and Galunic 2004). For these reasons, networks rich in structural diversity are thought to confer "information benefits" or "vision advantages" that improve performance by providing access to diverse and novel perspectives, ideas and information (Burt 1992).

These are the central inferences on which structural theories of brokerage and the strength of weak ties rest, and it is therefore intuitive to expect that *having structurally diverse networks – networks low in cohesion and structural equivalence and rich in structural holes – is positively associated with receiving more diverse information and more total non-redundant information*, and that *access to more diverse information and more total non-redundant information is positively associated with individual performance.*[1] Over the last four decades, these two inferences have guided the way sociologists think about information flow in networks, motivating and informing thousands of empirical studies of innovation

(Hargadon and Sutton 1997, Burt 2004), academic output (Swedberg 1990), team performance (Regans and Zuckerman 2001), the formation of industry structures (Kogut, Walker and Shan 1997), the success of social movements (Centola and Macy 2007) and labor market outcomes (Montgomery 1991).

However, theoretical arguments linking network diversity to novel information have thus far focused almost exclusively on the relative diversity of the information received across different alters in a network, generally overlooking the diversity and volume of novel information flowing within each tie or channel over time. Although dense, cohesive networks tend to deliver information that is redundant *across* channels (with each alter providing the same or similar information), relationships in such networks are also typically stronger (Granovetter 1973, Burt 1992), implying greater frequency of interaction and richer information flows. Metaphorically, such ties have greater channel bandwidth. In contrast, weak ties offer less communication (Granovetter 1973, Burt 1992), and information should flow through them less frequently (Granovetter 1973), with lower complexity and detail (Hansen 1999, Uzzi 1997).[2]

Two mechanisms explain why socially distant weak ties should interact and communicate less: exposure and motivation. As contacts interact more frequently, they are more likely to be exposed to and to spend time with each others' contacts in cohesive embedded networks (Granovetter 1973). Cohesive embedded networks also motivate their members to interact with one another—social pressure, cognitive balance and the development of cooperative norms in embedded relationships inspire us to devote time and energy to communicating with embedded ties (Heider 1958, Newcomb 1961, Granovetter 1973, 1985, 1992, Coleman 1988).[3] In relationships among firms in New York's apparel industry for example, Uzzi (1997) reports that socially distant weak ties were "non-repeated … one shot deals," in which communication occurred much less frequently, while embedded ties were characterized by "constant communication." Similar evidence has been found in R&D organizations (Allen 1977, Reagans and Zuckerman 2001, Reagans and McEvily 2003), innovation labs (Hargadon and Sutton 1997), job seeking (Granovetter 1973), familial relations (Coleman 1988) and in relationships between firms' business units (Hansen 1999) and across firms (Helper et al 2000). Given evidence suggesting the prevalence of weak ties in structurally diverse networks and the likelihood of increased information flow in cohesive networks due

to motivation and exposure, the bandwidth of communication channels should be lower in diverse networks. Thus, network diversity and channel bandwidth should trade off such that *greater network diversity is associated with lower channel bandwidth.*

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

All else equal, greater channel bandwidth should also provide access to more diverse information and more total non-redundant information because interaction through rich high-bandwidth channels tends to be more detailed, cover more topics and address more complex, interdependent concepts. While unconnected alters may have more novel information, the amount of useful novel information delivered to ego should increase in cohesive networks, in which both the volume of the information flow and the motivation to share relevant novel information is greater. As Reagans and McEvily (2003: 262) argue, "It is easier to transfer all kinds of knowledge [codified and tacit, simple and complex] in a strong tie and more difficult to transfer all kinds … in a weak tie." If many interdependent ideas must be applied together, then throughput must increase to transfer them all. Even the seminal work favoring weak ties as a source of novel information foreshadows in a footnote that "one possible model would expect information to flow through ties in proportion to time expended in interaction; this model would predict much more information via strong ties." (Granovetter 1973: 1372) We consider how just such a model can reform conventional wisdom regarding the relationship between social structure and access to novel information.

SOCIAL PROCESSES AND ACCESS TO NOVEL INFORMATION

While most current theories describe networks as channels, pipes, bridges or conduits (e.g. Podolny 2001, Centola and Macy 2007); characterize content as "attributes of nodes" (e.g. Rodan & Galunic 2004); and implicitly assume that information flows in proportion to the distribution of information in the network (e.g. Granovetter 1978, Schelling 1978, Kleinberg, Kempe & Tardros 2003),[4] information exchange is fundamentally a social process and knowledge transfer a discretionary activity (Reagans & McEvily 2003, Wu et. al. 2004). A connection to any individual affords the possibility of receiving the information she possesses, but by no means guarantees it. As Wu et. al. (2004: 328) point out:

"[I]nformation is selective and passed by its host only to individuals the host thinks would be interested in it." In competitive settings, information is often withheld even when it is known to be of interest to others. Networks are not simply pipes into different pools of information; they reflect the nature of the relationships, interactions and information exchanges taking place among those they connect.

Although the channel, pipe, bridge and conduit metaphors are common in sociology, such terminology hides restrictive assumptions about network structure preceding information flow. Human interactions in fact define social network structure. So, to avoid problems with channel metaphors we argue first from social processes, using social distance as synecdoche for less frequent interaction, lower mutual commitment and limited understanding. Speaking metaphorically, social distance is inverse bandwidth. Five social mechanisms, summarized in Table 1, then explain why greater channel bandwidth and lower social distance should *increase* access to novel information.

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

Social Capital. In relationships characterized by strong cohesive ties, contacts are likely to be more willing to share information. Diverse, low bandwidth ties are typically opportunistic, functional and only selfishly cooperative (Granovetter 1973, Uzzi 1997), while cohesive, embedded ties are typically characterized by greater intimacy, trust, emotional intensity and mutual confiding (Coleman 1988, Uzzi 1996). Social cohesion motivates individuals to devote time and effort to communicating with and assisting one another (Granovetter 1985, Coleman 1988). The development of cooperative norms (Granovetter 1992) and the subsequent reduction in competition in cohesive networks are likely to increase knowledge transfer between individuals (Szulanski 1996, Argote 1999, Reagans and McEvily 2003). Social capital in strong high bandwidth relationships gives ego the standing to seek information and alter the comfort to offer information. It also engenders the levels of trust that allow contacts to share both sensitive and non-sensitive information. A weak-tie relationship will typically only provide access to the non-sensitive information. Similarly, in weak-tie relationships alters will be less willing to devote time and effort to information exchanges with ego, who will get less in return for placing burdensome requests and will receive less total novel information.

In the context of job seeking, Granovetter (1973: 1371) nicely sets up the open empirical question we seek to address: “A natural a priori idea is that those with whom one has strong ties are motivated to help with job information. Opposed to this greater motivation are the structural arguments I have been making: those to whom we are weakly tied… will have access to information different from that which we receive.” Social capital, developed through prior information sharing, enables ego to seek and alters to share more novel information in high bandwidth relationships. Whether strong or weak ties deliver more total novel information therefore remains a critical open question.

Transactive Memory. Wegner (1987) introduced the term "transactive memory" to describe intimate relationships in which individuals have organized into mutually determined and understood domains of expertise. Although developing a relationship can be understood "as a process of mutual … disclosure … it can also be [understood] as a necessary precursor to transactive memory." (Wegner 1987: 200) As relationships develop, contacts become more familiar with each other’s areas of interest and expertise. Knowing who knows what makes embedded relationships with high bandwidth communication channels a more likely source of novel information.

Discovery of remote information is more likely when ego knows whom to ask for it (Wegner 1987). Stronger ties are more familiar with each other’s catalog of knowledge, inspiring information exchanges on a larger number and wider variety of topics. The greater the social distance between two people, the lower the likelihood that ego knows what an alter knows, limiting ego’s ability to seek information effectively and alters’ ability to proactively offer relevant novel information to ego. Knowing who has the most information about job opportunities or where to seek funding facilitates the search process even if the information to be transferred is itself not known beforehand. In Uzzi’s study of the fashion industry, knowing who possesses information on how get the best price for wool precedes discovery of that price and where it is offered. Building catalogues of expertise requires prior shared experience, which is a characteristic of strong-tie relationships (Wegner 1987, Liang et. al. 1995, Cramton 2001). More frequent interaction also gives alters a broader catalog of ego's knowledge and interests, making it easier for them to volunteer relevant non-redundant information. For example, alters are more likely to volunteer

information about potentially relevant job opportunities if they know that ego is looking for a job and in which industry ego is interested in working.

Search-Transfer. While transfer of simple news might be efficient in weak ties, that is not the case for complex information on multiple interdependent topics. Weak ties are inherently limited in the set of novel information they can transfer to the subset of 'simple' novel information (Hansen 1999). As Reagans and McEvily (2003: 242) demonstrate, strong embedded ties create a favorable social environment for information transfer: "Cohesion around a relationship can ease knowledge transfer by decreasing the competitive and motivational impediments that arise, specifically the fact that knowledge transfer is typically beneficial for the recipient but can be costly for the source." Awareness of a previously unknown software module can pass easily via an infrequent social contact. But, transferring that module together with interdependent instructions and contextual information requires a level of expert assistance that implies a helping relationship (Hansen 1999).

Information exchanges in embedded relationships are likely to be more detailed, and also more holistic in the sense that they not only convey discrete bits of information but also meta information about how each discrete idea connects with others, as well as discussion of the conceptual implications of each idea. On the other hand, structurally diverse bridging ties are usually formed for a particular purpose and in order to deliver information on a single or a limited number of dimensions. Such information is likely to be more discrete, summarizing a number of dimensions in a single signal, such as the price of goods in an economic relationship. Uzzi (1997) describes how representatives of firms engaged in embedded relationships go beyond exchanging price information to also discussing more detailed implications concerning profit margins, fashion sense and strategy. People can absorb ideas more easily on topics matching their expertise (Cohen & Levinthal 1990), and cohesive embedded ties, in effect those with high bandwidth, have been shown to produce higher rates of complex knowledge transfer in contract R&D (Reagans & McEvily, 2003) and product innovation firms (Hansen 1999). Bandwidth therefore affects the ability to share complex forms of novelty.

Knowledge Creation. Creating new knowledge also injects more novelty into the network and often requires rich interaction through thick communication channels. Songwriters and artists benefit from community embeddedness as their ideas feed on one another. In creative works such as Broadway musicals a team combines initially separate ideas through a creative process of brainstorming, problem solving and collaboration (Uzzi & Spiro 2005). Such idea-generating collaborations are rarely socially remote. They commonly arise in apprenticeship relationships for example between professors and graduate students or between colleagues interacting based on common interests (Lave and Wenger 1991). Obstfeld (2005) finds that brokers who bring together disconnected alters, in effect increasing the frequency of their interactions, promote innovation more than those who keep their contacts separated. Successful innovation teams coordinate their knowledge and actions, intentionally pushing new knowledge to all team members. For instance, initiating design changes to the set of a stage play requires collaborators to update team members quickly and often. These updates to one's social network bring people together and coordinate group action, representing a "union" strategy. In the context of an automotive engineering firm, this strategy was more conducive to trust, cooperation, transfers of complex knowledge, and ultimately to idea generation than "disunion" strategies that kept contacts apart (Obstfeld 2005). In Obstfeld's setting, new social knowledge generated from prolonged contact between engineers helped create innovation, demonstrating one way dense cohesive social networks outperform sparse networks with structural holes.[5]

Homophily. Homophily among those in cohesive embedded networks makes them more likely to share mutual interests across a wider variety of topics due to similarities across a greater number of distinct social dimensions (Blau 1986, McPherson et. al. 2001). Though overlapping interests across a greater number of dimensions have been theorized to create redundancy, they can in fact inspire more multifaceted communication, creating opportunities for high bandwidth channels to deliver more of the different dimensions of information known to each contact. We are more likely to be inspired to cover more topical ground in conversation with those with whom we share a greater number of common interests. Individuals connected by cohesive ties are more likely to engage each other more deeply and to participate in cooperative activities such as joint problem solving, so they are more likely to discover topics of

mutual interest in their discussions and to subsequently continue to both generate and exchange information on those additional dimensions (Uzzi 1997, Helper et al 2000).

In summary, these five social phenomena (social capital, transactive memory, search-transfer, knowledge creation, and homophily) imply that as the bandwidth of a channel increases the topical diversity of information and the total volume of novel information flowing through it should also increase. We therefore expect that *channel bandwidth is positively associated with receiving more diverse information and more total non-redundant information*.

INFORMATION ENVIRONMENTS AND THE CONTINGENCY OF VISION ADVANTAGES

If network diversity and channel bandwidth tradeoff and if both provide access to novel information, then which provides greater information advantages to brokers will depend on the information environments in which brokers find themselves. Although a diverse network of weak, low bandwidth ties ("diverse-low bandwidth") *can* provide access to more novel information than a cohesive network of strong, high bandwidth ties ("cohesive-high bandwidth"), the converse is also possible and in many cases more likely. Three characteristics of information environments should affect the degree to which bandwidth delivers more novel information to ego. First, the more information overlaps among people in the network, the less structural diversity should confer information advantages. Second, the larger the total size of the topic space, the more important bandwidth should be. Third, the more information changes over time, the more cohesive-high bandwidth networks should deliver novel information.

In the following section we translate our theory into probabilistic expectations of access to novel information in different information environments. These expectations describe how social motivations to exchange more information and the likelihood of greater redundancy in densely connected groups affect the likelihood of receiving novel information from both diverse-low bandwidth and cohesive-high bandwidth networks. Each alter has information on certain topics (represented by numbers), which together comprise the set of topics or ideas that exist in the network. The numbers of arrows between actors represent the bandwidths of communication channels (which parallels tie strength).

Consider two actors Alex (*A*) and Beth (*B*) depicted in Figure 3, Panel 1. Alex has weak, low bandwidth ties to unconnected alters Isaac (*i*) and Jake (*j*), while Beth has strong, high bandwidth ties to alters Kim (*k*) and Lauren (*l*), who connect to each other via strong ties. Alex's ties to Isaac and Jake are more likely to be low bandwidth because he is less likely to have sufficient social capital with them to inspire them to share more, he is less likely to know what they know (as are they to know what he needs), they are likely to have less in common and thus are less likely to share information, and they are less likely to create new knowledge together. However, because all three are socially distant they are more likely to have different information from each other. This scenario captures classic arguments about network structure and information access as well as the diversity-bandwidth trade-off.

Alex's weak tie contacts, being separated by a structural hole, have no redundant information, while Beth's strong tie contacts, being strongly connected, have redundant information. To demonstrate the importance of the diversity-bandwidth tradeoff in even extreme settings that are least favorable to our theory, we invoke the most conservative version of Granovetter's original forbidden triad argument. Although, according to Granovetter, the strong ties connecting *B*-*k* and *B*-*l* imply the *k*-*l* tie "is always present (whether strong or weak)" (Granovetter 1973: 1363), we represent the *k*-*l* tie as a strong connection and assume complete information homogeneity between Kim (*k*) and Lauren (*l*). This same basic scenario holds across all Panels 1-6, yet Kim and Lauren frequently provide more novel information to Beth than Isaac and Jake provide to Alex because they furnish a greater overall volume of information. Due to the high bandwidth nature of their relationships, they are more willing and have more opportunities to provide Beth more samples of their respective information spaces. In social terms, whether this extra volume contains extra novelty per unit of information is a tradeoff that depends on (i) how much the information of alters overlaps with one another (ii) the total number of topics in alters' catalog of knowledge, and (iii) the rate at which information in the network refreshes or updates.

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

The classic weak-tie, structural hole argument sets the baseline in Panel 1, which represents weak and strong tie strengths by two arrows and three arrows respectively. Each alter has information on four

topics (i, k, l = {1, 2, 3, 4} and j = {5, 6, 7, 8}), but only Alex's contacts have no overlap in their information. Alex's weak, low bandwidth ties to Isaac and Jake allow him to secure two samples each from their topic spaces. Beth secures more information samples from Kim and Lauren than Alex does from his alters because of the social processes that characterize their respective information exchanges. Beth has more opportunities to talk with her alters, who are more motivated to share information due to the social pressure, cooperative norms and cognitive balance that have developed in their embedded relationships. Those factors also make Kim and Lauren less likely to withhold information and more likely to proactively offer information to Beth.

Assuming alters do not offer the same piece of information twice, Alex samples two non-redundant items from Isaac and two non-redundant items from Jake, receiving four total novel pieces of information overall. Beth on the other hand will receive three novel pieces of information from her first contact Kim, but there is only a ¼ probability that she will receive a novel piece of information in her subsequent exchange with Lauren. If Beth's first draw from Lauren is novel, Lauren has no more non-redundant information to share.[6] Assuming redundant information on her first exchange (which occurs with probability ¾), Beth then has a one in three chance of receiving non-redundant information on her second exchange with Lauren. Over these two exchanges, Beth receives novel information with cumulative probability ½ (as given by ¼ + (¾)*(1/3) = ½). If Beth has not received new information by the third exchange (which occurs with probability ½), she retains a ½ chance of receiving non-redundant information in her last exchange. The total chance of Beth receiving novel information over three exchanges is ¾ (given by ¼ + (¾)*(1/3) + ½(½) = ¾). The total number of non-redundant pieces of information Beth expects to receive is thus 3 and ¾ given that she started by receiving 3 non-redundant items from Kim.

If each bit of novel information represents a job opening, then Alex's social network spans eight different opportunities and he can expect to receive news about four of them. In contrast, Beth's social network includes only four opportunities and she can expect, on average, to receive news of fewer opportunities. This is due to the heterogeneity of information among Alex's contacts, and demonstrates the value of structural diversity in delivering novel information. Even though Alex has fewer opportunities to

exchange information with his contacts, he still expects to receive more novel information because his social network bridges non-overlapping information pools separated by structural holes.

In Panel 2, we examine the same scenario but raise the bandwidth of Beth's ties by one and reduce the bandwidth of Alex's ties by one. The power of bandwidth becomes immediately apparent. While we maintain the same conservative assumptions about the distribution of information across alters (Kim and Lauren have completely redundant information, while Isaac and Jake have completely non-redundant information), the increased bandwidth of Beth's ties is enough to provide her with more expected novel information. In fact, the example is trivial. While Alex expects to receive two pieces of non-redundant information (one each from Isaac and Jake), Beth expects to receive four pieces of novel information simply because the bandwidth of her communication channels with Kim and Lauren is higher. In fact, the relative benefit of bandwidth is based on a model that is socially conservative. In their study of R&D transfer, Reagans and McEvily (2003) found that cohesion improves the willingness and ability to transfer information by reducing competition and costs of sharing. Here, Isaac and Jake might have preferred to hoard their unique information either to use themselves or because alters in their positions are more likely to compete, while Kim has less incentive to keep from Beth what Lauren can also share.

In Panel 3, we relax the conservative assumption of complete information heterogeneity between Isaac and Jake by introducing partial overlap in their information sets.[7] Although Kim and Lauren continue to have completely homogeneous information, the scenario again tips in favor of channel bandwidth – the cohesive-high bandwidth ties yield more novel information. The only difference in this panel is that Jake's information overlaps with Isaac's information by 50%. Alex still receives two novel pieces of information from Isaac but then on contact with Jake, only receives novel information with probability ½. Assuming Alex receives no novel information during his first interaction with Jake (which occurs with symmetric probability ½) he will receive novel information during his second interaction with probability 2/3^{rds} as two of the three remaining information items available from Jake are novel. If however, he does receive novel information in his first interaction, the chance of receiving novel information on his second interaction falls to 1/3rd. The total probability of Alex receiving novel information over both draws from

Jake is 1 (based on interaction one: ½ + interaction two: ½ (1/3) + ½ (2/3)). So, Alex expects to receive three total items of novel information, one from Jake and two from Isaac. As Beth's likelihood of receiving novel information has not changed relative to Panel 1 (3 and 3/4^{ths} from Kim and Lauren respectively), Beth expects to receive novel information with greater likelihood and than Alex in Panel 3. This example demonstrates the value of channel bandwidth in delivering novel information even when one's alters have completely overlapping information, which arises from the ability to exchange a greater volume of information with each contact. Panels 1-3 imply the following: *All else equal, we expect that the greater the information overlap among alters, the less valuable structural diversity will be in providing access to novel information.*[8]

In Panel 4 we illustrate the effect of a complex or high dimensional information environment by broadening the overall topic space. Now, alters are aware of twelve topics instead of four. The bandwidths of ties are as they were in Panel 1. Alex's contacts Isaac and Jake again have non-redundant information sets and Beth's contacts Kim and Lauren have redundant information sets. As in Panel 1, Alex expects four items of novel information, but in this case, because Beth's high-bandwidth ties sample from a broader information space with less chance of collision, she expects more novel information overall. In her first three interactions with Kim, Beth receives three novel items of information, but it is apparent after only her second interaction with Lauren that Beth's total expected novel information exceeds that of Alex. The chain can be established by summing the probabilities of receiving novel information from each of the three interactions. Reduce the denominator once for each draw; reduce the numerator once for each success.[9] On average, receiving 3 pieces of novel information from Kim and 2¼ from Lauren, Beth expects to do better than Alex based on a larger topic space. The difficulty of transferring complex information makes bandwidth even more important in this case. If three units of interdependent information need to be transferred together to be useful, then Beth's benefit of bandwidth is understated. Alex may not be able use the two pieces of novel information he receives from Isaac and Jake if he has insufficient context to understand them. Likewise, social capital theory also predicts Beth is better off. It is easier to ask for one item than ten. Alex must be willing to ask for more and his contacts must be willing to share but

Beth is better positioned to both ask and receive. Further, creativity is often higher when there are more ideas to work with (Weitzman 1998), implying the value of novel information is higher in the presence of a greater volume of novelty. Panel 4 implies that, *all else equal, the broader the topic space, the more valuable channel bandwidth will be in providing access to novel information.*

Thus far, we have presented the diversity-bandwidth trade-off in purely static contexts where colleagues' information does not change. A more realistic scenario involves dynamic updating. As we become aware of news concerning our workplaces, our friends and changes in the world around us, we revise our understanding of basic facts as well as complex know-how. The advance of Internet technologies, mobile service applications for personalized news and the 'always on' nature of online social networks can in fact accelerate the pace at which our knowledge of the world refreshes. Information simultaneously obsolesces as it updates. Environmental turbulence inspires adaptation (Galbraith 1974, March 1991) and changing information makes learning from experience more difficult (Weick 1979). As prior knowledge becomes obsolete more quickly, accessing timely information requires gathering news more frequently.

Reinterpreting a classic example (Granovetter 1973), suppose that highly desirable job openings fill quickly but that undesirable jobs remain open longer. Information drawn from weak ties about the jobs currently available can sample disproportionately from undesirable jobs. By the time a weak tie delivers information about a desirable job, information about that job is already well known to competing alters whose strong ties update them more quickly. If information about jobs refreshes often or obsolesces quickly, frequent communication is essential to getting news before others. This speaks directly to the issue of the information refresh rate relative to channel bandwidth. High bandwidth ties are more likely to deliver time-critical information, and are thus more likely to deliver non-redundant information in turbulent information environments. Panels 5 and 6 therefore introduce time.

To reestablish the weak-tie/structural-hole baseline, Panel 5 shows that diverse low-bandwidth ties can provide more novel information. In both Panels 1 and 5, Beth's contacts' knowledge overlaps while Alex's does not; Beth has bandwidth three while Alex has bandwidth two; and information sets

span a topic space of four. But, in Panel 5, information refreshes. Dashed lines separate changes in information. Since Panel 5 spans two periods (T_1 and T_2), expected access to novel information exactly doubles that of Panel 1. Panel 6, however, shows a more turbulent environment. Updates occur twice per period as shown for example by the fact that Isaac's information set changes from {1, 2, 3, 4} to {5, 6, 7, 8} within period T_1. Although Beth might learn of three news items (among 1, 2, 3 or 4) from Kim, by the time she checks with Lauren, the context has already changed such that she learns three new items (from among 5, 6, 7 or 8). This gives her six novel pieces of information per period, a full dozen across both periods.[10] High bandwidth ties can therefore provide more access to new information in more turbulent information environments, despite being more structurally constrained.

In a slow-moving information environment such as roof repair (a roof needs repair roughly once every twenty years), a roofer's network of weak ties is sufficient to deliver information about potential jobs (Podolny 2001). But in turbulent environments such as stock market arbitrage, minute advantages can be critical and people must shift from exploiting what they know to exploring what they do not know quickly and often (March 1991). In communications terms, this means interacting more frequently and increasing communication channel bandwidth. For transactive memory systems, change renders the catalog of others' knowledge obsolete and a person searches less effectively without updates. In the creativity literature, the chance at Schumpeterian recombination of ideas rises as individuals are exposed to change and design changes must be shared with team mates more quickly for projects to be successful (Obstfeld 2005). Constantly changing information implies that ego does not need to change channels to receive incremental novelty because what their contacts have to tell them is itself changing, refreshing or updating. The greater the bandwidth of communication channels, the more of this newly updated information will be passed on to ego in a timely manner. *We therefore expect, all else equal, that the higher the refresh rate, the more valuable channel bandwidth will be in providing access to novel information.*

Since stylized examples depend heavily on assumptions and initial conditions, we extend these illustrations by developing a more general analytical model of our arguments in Appendix A. We formally prove there that each of the factors previously discussed can make either a diverse-low bandwidth net-

work or a cohesive-high bandwidth network more attractive in terms of access to novel information. The key intuition is conveyed by representing "bias" as the tendency of cohesive ties to share the same redundant elements from a topic vector. When the disadvantage of bias swamps the advantage of bandwidth, the diverse-low bandwidth tie provides greater chance of encountering novel information. But, when the advantage of bandwidth swamps the disadvantage of bias, the constrained-high bandwidth tie is preferable. While a range of intermediate cases span these extremes, conditions exist (depending on bias, bandwidth and the number of links already present) in which a person will always prefer one or the other type of tie.

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

The diversity-bandwidth trade-off implies that vision advantages are contingent on the different social settings and information environments in which brokers are situated. In turbulent social settings or intellectual domains where conditions change rapidly and news, ideas and methods are frequently updated, greater channel bandwidth is more useful for delivering novel information. On the other hand, if information possessed by alters is relatively static, structural diversity becomes the more important factor. In highly heterogeneous information environments in which local network neighborhoods possess distinct, non-overlapping information, bandwidth is less beneficial than structural diversity. But, when the overlap of information among alters is more pronounced the opposite is true. In environments with multiple complex ideas, bandwidth delivers greater novelty, but when the topic space is limited, structural diversity trumps bandwidth. These contingencies are critical to understanding brokerage because the configurations that produce them are among the most prevalent in human social networks. Since structurally diverse strong ties and cohesive embedded weak ties are both relatively rare (Granovetter 1973, Burt 1992, Watts and Strogatz 1998, Watts 1999, Centola and Macy 2007), the contingent scenarios are the most useful for explaining relationships between networks, information flow and performance outcomes in a variety of social contexts.

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

Unfortunately, the vast majority of empirical work on networks and information advantage is "content agnostic" (Hansen 1999: 83). While there is abundant evidence linking social structure to performance (e.g. Burt 1992, 2004a, 2007, Reagans and Zuckerman 2001, Sparrowe et al. 2001, Cummings and Cross 2003, Cummings 2004, Aral et. al. 2006, 2007), empirical data on information flowing through networked relationships is rarely used to validate information-based theories of brokerage and the strength of weak ties. As Burt (2008: 253) notes: "Empirical success in predicting performance with network models has far outstripped our understanding of the way information flow in networks is responsible for network effects. A cluster of network concepts emerged in the 1970s on the idea that advantage results from connections with multiple, otherwise disconnected, groups and individuals. The hubs in a social network were argued to have advantaged access to information and control over its distribution… However, the substance of advantage, information, is almost never observed." "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 …." (Burt 2005: 60)[11] We therefore test our arguments by combining social network and performance data with direct observation of the information content flowing through e-mail communication.

DATA AND METHODS

Research Setting

We collected e-mail messages exchanged by employees of an executive recruiting firm with fourteen offices across the United States, analyzing their topical content to determine the relative heterogeneity and novelty of the information passed between the employees. Previous research by Wu et. al. (2004) and Kossinets and Watts (2006, 2009) validates the usefulness of e-mail data in characterizing and analyzing social networks in firms and academic institutions. We extend that research by combining analysis of the social structure of e-mail communication with an evaluation of the information content of messages. We argue that combining analysis of message content and communication topology will open new avenues for answering questions at the heart of the sociology of information. Although information flow

can be documented in a limited way with ethnographic and survey data (Baker 1984, Obsfeldt 2005, Reagans and McEvily 2003), direct observation of information content and its variation across and movement through networks is critical to accurately testing information-based theories of social capital (Burt 2008).

By analyzing e-mail communication patterns and message content, we are not only able to match network structures to the subject matter of the content flowing through them, but also to avoid inaccuracy in respondents' recall of their social networks and communication. Most prior research elicits network data from respondents who have difficulty recalling their networks (e.g. Bernard et. al 1981), particularly when contacts are socially distant (Krackhardt and Kilduff 1999). The inaccuracy of respondent recall and the bias associated with recall at social distance creates inaccurate estimates of network variables (Kumbasar, Romney and Batchelder 1994), forcing most empirical studies to artificially limit the boundary of estimated networks to local areas around respondents (e.g. Reagans and McEvily 2003). Such artificial boundaries 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. We therefore took several steps to ensure a high level of participation in the study (described below). As 87% of eligible employees agreed to participate we collected email network and content data with nearly full coverage of the firm. There are no statistical differences between participants and those who opted out of the study on dimensions of relevance to the analysis.[12]

As the company's work was geographically dispersed and instant messaging was rarely used, recruiters relied on e-mail as their primary means of communication.[13] As one recruiter put it "[s]taff spend an enormous amount of time coordinating. We are *big* users of e-mail." The e-mail 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. This structure offers a unique perspective on the value of network and information diversity as measured in e-mail data for two reasons. First, since geographic dispersion makes face-to-face meetings difficult it establishes e-mail as an even more important source of information (Hinds and Keisler 2002). Second, redundant information and ex-

pertise tend to pool in each dispersed geographic location, enabling recruiters with diverse networks to reach across structural holes into distinct pools of information, making this setting particularly well suited to analyzing the information benefits of brokerage.

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

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. Recruiters report being more effective when they receive rich information from their colleagues about candidate qualifications, client idiosyncrasies, team coordination and methods for circumventing secretarial screens or handling difficult placements.[14]

Executive recruiters are quintessential brokers. Access to diverse and novel information is a critical component of their business. 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 and Coverdill 2000). Information about a diverse pool of candidates, diverse markets and diverse client firms reduces the time a recruiter wastes interviewing unsuitable candidates and improves the quality of placements (Aral et. al. 2006). 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." Having different information on how to 'penetrate' different private companies can make recruiters more effective at gathering the information and contacts they need to match candidates to clients. Information sharing also enables coordination, reducing total work among teams of recruiters searching for similar candidates or clients. As one recruiter told us, "Communication within and across teams is a big success factor. It eliminates double work."

In these ways, recruiters' access to diverse information is critical for filling different types of positions and performing complex matching of candidate strengths and weaknesses to client needs. Recruit-

ers emphasize the need for diverse contacts, reporting that "[d]iversity means more and better contacts" because "[s]kill sets are complementary and not perfectly overlapping." Our interviews included several executive recruiter trainers. One trainer, who describes her job as "helping recruiters learn to be better recruiters," told us,"[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." For those reasons we expect diverse and novel information is particularly important for explaining variance in recruiter performance.

Data

Our data come from four sources: (i) detailed accounting records of individual project assignments and performance, (ii) e-mail 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 complete access to their internal accounting and project databases for records spanning 2000 to 2005. Those databases describe revenues generated by individual recruiters, contract start and stop dates, projects handled by each recruiter, project team composition and job levels of recruiters and placed candidates. From that data we were able to mine excellent performance measures that could be normalized for quality. E-mail data includes all messages sent through the firm for a period of ten months, 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 e-mail messages captured.[15] Details of e-mail data collection are described by (Aral et al 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 data set is made up 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 C). Descriptive statistics and correlations of all variables are provided in Tables 3 and 4 (we detail construction of each variable in the next section). An observation is one person-month.

----- TABLES 3 & 4 -----

Variable Construction

Dependent Variables

Recruiters in this firm measure success by the number of job openings filled and the amount of revenue generated per unit time. We therefore assess a recruiter's performance by measuring the number of *projects completed* per month and *revenues generated* per month as recorded in the firm's accounting records. In addition to revenues and project completions, the speed with which vacancies are filled is also an important intermediate measure of workers' productivity. Contract completion implies that recruiters have met a client's minimum thresholds of candidate fit and quality. Project completion can be interpreted as a quality controlled measure of productivity—a faster rate implies that a recruiter is creating high quality matches in a shorter period of time. As one recruiter told us: "[t]he longer a client delays, the lower the probability of job acceptance." We therefore also measure *average project duration*.

Network Variables

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, including degree centrality, betweenness centrality and network reach, which describes the breadth and range of actors' networks (see Burt 1992: 12). Network size is significantly correlated with degree centrality ($\rho = .70$; $p < .001$), betweenness centrality ($\rho = .77$; $p < .001$) and reach ($\rho = .56$; $p < .001$) among employees in this organization, 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. 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.[16] We define constraint C_i (Burt 1992: 55)[17] as the lack of structural holes in an actor's network using bidirectional e-mail traffic to construct ego networks, 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 structural equivalence of two actors, measured as the Euclidean distance of their contact vectors.[18] By measuring both network diversity and the structural equivalence of alters we account for the possibility that small-world networks, or cohesive cliques linked by infrequent weak ties, could bring novel information into a clique (Watts and Strogatz 1998).

Channel Bandwidth. Bandwidth measures the volume of communication over a given channel. As our unit of analysis is the monthly ego network and performance variables are computed monthly, we measure bandwidth by recording average monthly message traffic over communication channels or ties, operationalized as the amount of incoming e-mail over the total number of contacts at time *t*, providing a measure of the average channel bandwidth of actors' ties:

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

Information Diversity and Novelty: A Vector Space Model of Communication Content

We model and measure the diversity and total novelty of information in individuals' e-mail using a Vector Space Model of the topics present in e-mail content (e.g. Salton et. al. 1975).[19] Vector Space Models represent textual content as vectors of topics in multidimensional space based on the relative prevalence of topic keywords. They are widely used in information retrieval and search query optimization algorithms to identify similar documents or to find topics identified by search terms. In our model, each e-mail is represented as a multidimensional topic vector in which elements are the frequencies of

keywords in the e-mail. The prevalence of certain keywords indicates that a topic that corresponds to those keywords is being discussed. For example, an e-mail about pets might include frequent mentions of the words "dog," "cat" and "veterinarian;" while an e-mail about statistics might mention the words "variance," "specification" and "heteroskedasticity." We evaluated the relative topical similarity of two e-mails by topic vector convergence or divergence – the degree to which their vectors point in the same or orthogonal directions in multidimensional topic space.[20] E-mails about similar topics are more likely to contain similar language, so the vectors used to represent them are closer in multidimensional space, reducing their collective variance, or spread. We therefore measured e-mail content diversity by characterizing all e-mails as topic vectors and measuring the spread of topic vectors in individuals' inboxes and outboxes as described below.

Construction of Topic Vectors and Keyword Selection. Our Vector Space Model represents each e-mail D_{il} (where *i* indexes e-mails and *l* indexes recruiters) as a vector of keyword frequencies k_{in}. Each e-mail is therefore represented as an n-dimensional vector of keyword frequencies in topic space,

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

where k_{in} represents the frequency of the *n*th keyword that appears in the *i*th e-mail. As terms that appear frequently in an e-mail are more likely to be thematic and to relate to the e-mail's subject matter, we used the 'term frequency' of keywords in e-mail as weights to construct topic vectors. An example of the vector construction process is shown in Figure 5.

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

The choice of keywords is an important step in the process. Rather than imposing exogenous keywords on the topic space based on our own thinking, we chose keywords likely to characterize useful, representative topics based on the following procedures.[21] First, we initialized our data by removing common "stop words," such as "a, "the," and "and" and other words that appear with high frequency across all e-mails, which are likely to create noise in content measures. We then ran an iterative, k-means clustering algorithm to group e-mails into clusters based on the co-occurrence of words in e-mails across

the entire corpus.[22] The result of iterative k-means clustering is a series of assignments of e-mails to clusters based on their language similarity. These clusters represent "topics" in that they group e-mails with similar topical language.

Second, in order to identify distinct topics in our corpus, keywords should *distinguish* topics from one another. We therefore chose keywords that maximized the mean frequency variation across k-means clusters, choosing words that tend to appear in the same topic clusters often and in other clusters relatively infrequently. This refinement favors words with widely differing mean frequencies across clusters, retaining words with an ability to distinguish between topics. In our data, we found the coefficient of variation of the mean frequencies of keyword *i* across topics (C^i) to be a good indicator of this dispersion.[23]

$$C^i = \frac{\sqrt{\frac{1}{n}\sum_{c=1}^{n}(m_c^i - \overline{M^i})^2}}{\overline{M^i}}$$

Third, keywords should *represent* the topics they are intended to identify. To achieve that goal we chose keywords that minimize the mean frequency variance within k-means clusters, favoring words that are consistently used across a large number of the e-mails in a given topic cluster. The Intra-Topic Frequency of keyword *i* (ITF^i) is therefore defined as follows:[24]

$$ITF^i = \frac{\sqrt{\sum_e \sum_c \left(f_{ec}^i - m_c^i\right)^2}}{\overline{M^i}}$$

Fourth, 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 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.[25] We then populated topic vectors representing the subject matter of each email (shown in Figure 4) and meas-

ured the diversity and novelty of the streams of email flowing to recruiters over time using the methods described below.

Measures of Information Diversity and Total Non-Redundant Information. Current literature remains vague in defining the dimensions of novelty or novel information that should matter for vision advantages. We believe two distinct aspects of novelty are important – the diversity of the information received, which can be thought of as the variance of the topics being discussed, and the total volume of novel information received. We developed two distinct empirical measures of novelty, one that captures variance (which we term "information diversity") and one that captures volume (which we call "total non-redundant information").

We measured the degree to which the e-mails in an individual employee's inbox or outbox are 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 purpose of all five measures is to characterize the degree to which e-mails are about a set of either focused or 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 e-mails based on their "information content."[26] All five diversity measures are highly correlated (~ corr = .98; see Appendix B), so our specifications use one of the most common measures, the average cosine distance of employees' incoming e-mail topic vectors d_{ij}^{I} from the mean vector of their topic space M_i^{I}, to represent incoming information diversity (ID_i^{I}):

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

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

This measure aggregates the cosine distance of e-mail vectors in an inbox from the mean topic vector of that inbox, approximating the spread or variance of topics in incoming e-mail for a given individual. We measure the total amount of *i*'s incoming e-mail communication as a count of incoming e-mail 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* as diversity (ID_i^I) times total incoming e-mail: $NRI_i^I = (E_i^I * ID_i^I)$. We performed extensive validation tests of our diversity measures by creating simulated e-mail inboxes using an independent data set from Wikipedia.com. These simulated inboxes ranged from sets containing highly diverse e-mails about different topics to sets containing highly focused e-mails about a limited number of similar topics. Our measures performed very well in accurately labeling the diverse sets as containing diverse information and vice versa (see Appendix C). A three-dimensional Vector Space Model of five e-mail vectors and their mean vector is shown in Figure 6.

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

Refresh Rate of Alters' Information (Refresh Rate). The information refresh rate of an alter *j* in month *t* (RR_{jt}) is defined as the cosine distance between every pair of *j*'s daily mean e-mail vectors in that month, including both incoming and outgoing e-mail.[27] In other words, to calculate the degree to which *j*'s information changed from day 1 to day 2 in month *t*, we calculated the mean vector of *j*'s e-mails on day 1 and the mean vector of *j*'s e-mails on day 2 then computed the cosine distance between them: $1 - Cos(M_{j\tau_1}, M_{j\tau_2})$. We then repeated this procedure for the mean vectors between day 1 and day 3, day 1 and day 4 and so on until we had dyadic comparisons between each pair of days in month *t*. We considered only measuring the cosine distance between contiguous days (day 1 and day 2, day 2 and day 3, etc.), but rejected this approach because topics of conversation may simply alternate over days in the week or longer periods. For example two contacts may email about topic 1 on Monday, topic 2 on Tuesday, go back to topic 1 on Wednesday, and again talk about topic 2 on Thursday. Topics might be repeated every third day, fourth day or every seventh day if there are recurring weekly meetings that inspire e-mail exchanges about those topics. Measuring information dissimilarity only among contiguous

days would not capture this potential topic switching and would incorrectly measure these patterns as being very diverse even though a limited number of topics are being repeatedly discussed. We therefore measure the information refresh rate of *i*'s local network P_{it} as a sum of information refresh rates (RR_{jt}) of *i*'s immediate neighbors *j* in month *t*, weighted by the strength of ties between *i* and *j*. We use the number of messages m_{jit} sent from *j* to *i* during month *t* as a proxy for the strength of incoming ties. Formally, we define the information refresh rate of a node *j* in month *t* as: $RR_{jt} = \sum_{\tau_1 < \tau_2} 1 - Cos(M_{j\tau_1}, M_{j\tau_2})$, where $M_{j\tau_1}$ is the mean vector of *j*'s e-mails on day τ_1, and where $M_{j\tau_2}$ is the mean vector of *j*'s e-mails on day τ_2. The information refresh rates of *i*'s contacts are then aggregated by summing the refresh rates of *i*'s alters *j* weighted by the strength of *i*'s incoming tie from each alter: $P_{it} = \sum_j RR_{jt} * m_{jit}$.

Topic Space of Alters (Topic Space). We measure the overall size of the topic space in ego's local network by measuring the total amount of non-redundant information *i*'s alters *j* exchange with their respective contacts $NRI_j = \sum_k (E_{jk} * ID_{jk})$. If the amount of total non-redundant information *i*'s alters receive and distribute is high, we expect *i* to be able to sample from a larger topic space. We therefore define the overall Topic Space of *i*'s network in month *t* (TS_{it}) as the sum the of non-redundant information of *i*'s contacts in month *t* weighted by number of messages sent from *j* to *i* during month *t* (m_{jit}): $TS_{it} = \sum_j NRI_{jt} * m_{jit}$.

Information Overlap of Alters (Information Overlap). Excessive similarity among alters' topic vectors signals that the information available to ego through different channels may be redundant. The extent to which the information of *i*'s neighbors is redundant depends on the dyadic overlap of all of *i*'s pairs of alters. We therefore calculate the information overlap of each pair of *i*'s alters in month *t* and average that result over the number of *i*'s contacts in month *t*: $IO_{jkt} = \sum_{k=1}^{N} Cos(M_{jt}, M_{kt}) / N$. We take the average information overlap between pairs of *i*'s alters so that the overlap proxy is independent of the number of alters in the network. We then simply sum the average overlap of the information of *i*'s con-

tacts in month t weighted by the number of messages sent from j to i during month t: $IO_{it} = \sum_j IO_{jkt} * m_{jit}$.[28]

Control Variables

Several additional factors could affect access to diverse novel information and individual performance. We therefore examine six possible alternative explanations for information advantage as control variables: expertise heterogeneity, demography, human capital, total communication volume, unobservable individual characteristics and temporal shocks to the flow of information in the firm.

Expertise Heterogeneity of Alters (Expertise Heterogeneity). A basic premise of brokerage theory is that disconnected network neighborhoods house dissimilar expertise, which brokers tap by reaching across structural holes. If that is true we would expect individuals with structurally diverse networks to be connected to alters with heterogeneous expertise and that this heterogeneity enables access to novel information. We measure the expertise heterogeneity of an employee's contacts by evaluating the diversity of their expertise accumulated through the projects they have completed in the past. 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 use the distributions of recruiters' prior project experience over project types rather than educational background to measure expertise heterogeneity. The firm categorizes projects into the following categories: CEO, COO, CIO, Medical Executive, Human Resources Executive, Business Development Executive, Nurse and 'Other.' We use these categories as the relevant areas of recruiters' expertise.[29] The *Expertise 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 she recorded against any given project in the accounting data. The measure is constructed as follows: $EH_{it} = 1 - \sum_{k=1}^{8} \left(\frac{q_{ik}}{q_i} \right)^2$.

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 a count of the number of projects of class *k*, weighted by effort share, that *j* has completed. The denominator, $q_i = \sum_{k=1}^{8} q_{ik}$ 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 that measure from one. As the expertise in *i*'s network becomes more concentrated in a few project classes the knowledge heterogeneity measure decreases.[30] Reagans and McEvily (2003) construct a similar measure of '*expertise overlap,*' but our measure differs by using accounting records to record project experience (rather than self reports of expertise) and weights the expertise in an employee's network by tie strength and the effort share of each alter on each project. Our measure of experience heterogeneity also changes over time as recruiters complete more projects of different types.

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 and Caldwell 1992, Reagans and Zuckerman 2001). Older employees may have related knowledge on a wider variety of topics or may be more aware of experts in the organization. Employment discrimination and interpersonal differences 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 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 individuals specialize as they gain experience, focusing their work and 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 differences 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 and Cross 2003). We therefore control for total e-mail communication.

Individual Characteristics and 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. Temporal shocks could also affect demand for the firm's services, with additional work stimulating information-seeking activities. 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 and create 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. We control for seasonal and transitory variation in our data by using dummy variables for each month and year. Figure 7 visualizes the expertise heterogeneity and information diversity variables by showing how project experience in different job classes and topics discussed in e-mails were distributed across a group of five recruiters.[31]

----- FIGURE 7 -----

Model Specification

We used panel data to estimate relationships between network structure and information access and between information access and performance. We are interested in how variations in network structure explain performance differentials between individuals, as well as how changes in actors' networks explain variation in their access to information and performance. If network structure generates social capital by influencing information access, actors that possess larger, more diverse networks with higher channel bandwidth should receive more novel information and perform better than their counterparts. However, unobserved heterogeneity in employees' personal characteristics, such as ambition, gregariousness or social intelligence, could simultaneously drive variation in network structure and performance. If unobserved characteristics of individuals are correlated with the error terms in our models, pooled OLS estimation will produce biased parameter estimates. To control for bias created by unobserved heterogeneity we examine variation within and across individuals over time using both fixed effects and random effects models. As observations in network data are not independent, we estimate a model of network autocorrelation of disturbances that provides consistent estimates of coefficients and standard errors that are robust to both network and temporal autocorrelation in panel data. Full details of our model specifications and estimation procedures are provided in Appendix D.

RESULTS

THE DIVERSITY-BANDWIDTH TRADE-OFF

If the diversity-bandwidth trade-off regulates the receipt of novel information we should observe two phenomena in our data. First, as recruiters' networks become more diverse, we should see the bandwidth of their communication channels contract. Second, they should receive more novel information as their networks become more structurally diverse and as channel bandwidth expands. If those conditions hold then a trade-off between network diversity and channel bandwidth is creating countervailing effects on the receipt of novel information.

We found strong evidence confirming the diversity-bandwidth trade-off. As recruiters communicated with contacts who were less well connected to each other and who occupied less structurally

equivalent positions in the network, the bandwidth of their communication channels to those contacts contracted quite rapidly. For instance, we estimated that a one standard deviation increase in the structural diversity of a recruiter's network over time was associated on average with a 21% reduction in the bandwidth of their communication channels (Models 1-3, Table 5, $p < .01$). As recruiters communicated more with contacts who were themselves densely connected and structurally equivalent the bandwidth of their communication channels expanded. There was a strong negative relationship between network diversity and channel bandwidth (Table 5, Model 1: $\beta = -.314$, $p < .01$) and a strong positive relationship between structural equivalence and channel bandwidth (Table 5, Model 1: $\beta = .107$, $p < .05$), indicating that as networks became more diverse the thickness of communication channels narrowed. These results held even when we controlled for network size and expertise heterogeneity in fixed effects models that also hold unobserved time-invariant heterogeneity constant across recruiters (Table 5, Models 1-3). We also found that both greater network diversity and greater channel bandwidth were strongly associated with the receipt of more diverse information and more total non-redundant information (Network Diversity: Table 6, Models 2 and 9, $p <.01$; Channel Bandwidth: Table 6, Models 4, $p < .05$ and 10, $p < .01$). Having established that the diversity-bandwidth trade-off regulates access to novel information, we then examined the conditions under which this tradeoff affects vision advantages.

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

In the organization we studied, work was organized by geographic regions and knowledge domains. Recruiters with diverse networks communicated with contacts whose prior experience and knowledge were heterogeneous, providing evidence of one way that diverse networks deliver diverse information – by providing access to pools of heterogeneous expertise. This mechanism is reflected in the strong positive association between expertise heterogeneity and network diversity in Models 6-10 in Table 5. In order to contact peers with varied expertise, recruiters diversified their communication networks (communicated with structurally distant alters) to reach across structural holes into local network neighborhoods less well connected with their own. This confirms earlier findings on the diversity of expertise

in networks rich in structural holes (Reagans and McEvily 2003, Rodan and Gallunic 2004) and supports a basic premise of brokerage theory - that disconnected network neighborhoods house dissimilar expertise and knowledge, which brokers tap by reaching across structural holes. However, as recruiters began to reach into diverse, unconnected network neighborhoods seeking advice, information or support, the bandwidth of their communication channels decreased (Table 5, Models 1-5 and 8-10). The negative associations between expertise heterogeneity and channels bandwidth in in pair wise correlations (ρ = - .25 $p < .05$), random effects models (β = -.21 $p < .01$, Model 5) and more conservative fixed effects models (β = -.095 $p < .10$, Model 3) provide corroborating evidence for the diversity-bandwidth trade-off. Individuals whose contacts had diverse knowledge and experience communicated more infrequently and with lower volume per channel, which is consistent with prior characterizations of the nature of weak-tie relationships (Granovetter 1973, Uzzi 1996) and provides new empirical evidence about how information tends to flow through them.

To create more diverse networks, recruiters must cultivate new structurally distant contacts, which increases their network size. Limited time, energy and attention could necessitate weaker, more infrequent and therefore lower bandwidth communication with those contacts, an argument consistent with the notion of network maintenance costs (Burt 1992). Interestingly however, our findings show that the reductions in channel bandwidth associated with greater network diversity do not seem to be driven only by the time and effort costs of network maintenance, but also by the nature of the relationships in sparse networks. The positive parameter estimate on the network size variable in bandwidth regressions (Table 5, Models 4-5) indicates that as recruiters cultivated more contacts the bandwidth of their communication channels widened rather than narrowing. If constraints on time and effort devoted to relationship maintenance alone were driving channel bandwidth we would expect bandwidth to decrease as network size increased. On the contrary, as recruiters communicated with more people, they also exchanged more messages per contact.

As network size continued to increase, time, energy and attention constraints eventually had their expected effect. The network size squared estimate on channel bandwidth is negative and significant in random effects specifications, indicating declining marginal increases in channel bandwidth as networks grew. The nonlinear relationship between network size and channel bandwidth suggests that there are simultaneous increases in network size and channel bandwidth in smaller networks, but that as network size exceeds the normalized population mean, time and effort costs and the nature of weak-tie relationships necessitate reductions in channel bandwidth (see Figure 8). Evidence of this maintenance-cost mechanism was only seen in random effects models that consider variation between recruiters and not in fixed effects models which analyze variation within observations of recruiters over time. This suggests unobserved heterogeneity between recruiters explains this variation. For instance, more gregarious recruiters could have larger networks and could communicate more with each contact on average up to a certain network size.

Models 7-10 in Table 5 also show a strong positive, but nonlinear, relationship between network size and network diversity. These results suggest that information benefits to larger networks are constrained in bounded organizational networks, and that marginal benefits to structural diversity decrease as a network grows in size. As recruiters contacted more colleagues, each new contact contributed a diminishing amount of structural diversity to the focal actor's network. The implications of this trade-off between size and structural diversity complement Burt's (1992: 167) concepts of "effective size" and "efficiency."[32] Figure 8 graphs the relationships among network size, network diversity and information diversity, clearly showing the positive, nonlinear relationships.

----- FIGURE 8 -----

Demographic variables have no effect on channel bandwidth in Models 4-5, while education has a consistently negative relationship, perhaps indicating that more educated employees are able to communicate more efficiently with fewer messages per channel. Fixed and random effects models are relatively consistent, except that network size and expertise heterogeneity variables are only correlated with channel bandwidth in random effects models, indicating that persistent variation in network size and ex-

pertise heterogeneity between individuals explained variation in channel bandwidth, while changes in individuals' network size and network expertise heterogeneity over time did not. On the other hand, changes in network diversity do explain changes in bandwidth over time. As recruiters' networks became more structurally diverse, the bandwidth of their communication channels contracted. Taken together, these results again confirm the trade-off between diversity and bandwidth.

THE DIVERSITY-BANDWIDTH TRADE-OFF AND ACCESS TO NOVEL INFORMATION

If vision advantages exist and are regulated by the diversity-bandwidth trade-off, we should observe positive effects from network diversity and channel bandwidth on the receipt of diverse novel information. Analyses estimating whether network diversity and channel bandwidth predict incoming information diversity (ID_{it}^{I}) and total non-redundant information received (NRI_{it}^{I}) are shown in Table 6.[33]

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

We found strong support for the basic argument that information benefits explain returns to structural diversity and brokerage. Network diversity was positively and significantly associated with greater information diversity in incoming e-mail. The first order diversity variable, which measures the lack of constraint in recruiters' networks, was highly significant in all specifications, while the average structural equivalence of recruiters' contacts did not influence access to diverse information (controlling for network size and first order structural diversity). A one standard deviation increase in network diversity was associated with ~ .15 standard deviation increase in the diversity of incoming information, demonstrating that large diverse networks provide access to diverse information. The expertise heterogeneity of recruiters' contacts was positively correlated with the diversity of the information recruiters received in both pair wise correlations (.23 $p < .05$, Table 4) and regression results (Table 6 Model 1). Controlling for total communication volume, a one standard deviation increase in the expertise heterogeneity of recruiters' contacts was associated with a .28 standard deviation increase in incoming information diversity (Model 1, $p < .01$). When the network diversity and structural equivalence terms were added to the estimation

(Model 2), the positive contribution of expertise heterogeneity to incoming information diversity was reduced by 75%, implying that network diversity and expertise heterogeneity are positively correlated and that network diversity is a stronger predictor of access to diverse information than the expertise heterogeneity of recruiters' contacts. As recruiters reached across structural holes they were not only communicating with those who had more diverse sets of expertise, they were also receiving more diverse information from their contacts as a result. This corroborates the theory that network diversity provides diverse information in part by providing access to diverse pools of expertise, but it also confirms that in our setting network structure is a stronger predictor of access to diverse information than the expertise heterogeneity of ego's contacts.

As recruiters added network contacts the contribution to information diversity lessened with each additional contact, implying diminishing marginal information benefits to larger networks. A one standard deviation increase in the size of recruiters' networks (approximately 8 additional contacts) was associated with a .5 standard deviation increase in information diversity (Models 3-7, $p < .01$); while the coefficient on network size squared was negative and significant, indicating diminishing marginal information benefits to network size (Models 3-7, $p < .01$).[34]

Finally, channel bandwidth was also associated with access to more diverse information, confirming that the diversity-bandwidth trade-off was regulating access to diverse information. A one standard deviation increase in channel bandwidth was associated with a .085 standard deviation increase in information diversity (Model 4, $p < .05$). When channel bandwidth was added to the specification, the magnitude of the estimated relationship between network diversity and information diversity increased. This implies a negative correlation between network diversity and channel bandwidth, providing additional corroborating evidence of the trade-off between the two.

While Models 1-7 in Table 6 estimate correlates of information diversity, Models 9-13 show that the total volume of novel information flowing to recruiters increased with their network size, network diversity and channel bandwidth. Expertise heterogeneity had a strong positive relationship with total non-redundant information received (Model 8, $p < .01$), until the network diversity and structural equiva-

lence variables were added to the specification (Model 9), again demonstrating that recruiters accessed novel information by reaching across structural holes into diverse pools of expertise. Network diversity and channel bandwidth both had strong positive relationships with the total amount of novel information flowing into actors' inboxes (Model 10, $p < .01$), with a one standard deviation increase in bandwidth associated with a .35 standard deviation increase in total novel information received ($p < .01$). As network size and the thickness of channels increased, the total volume of novel information received also increased. These results demonstrate the importance of considering channel bandwidth, as well as the diversity-bandwidth trade-off, when estimating relationships between network structure and access to diverse novel information. Bandwidth trades off with network diversity and has a strong positive relationship with incoming information diversity and total non-redundant information, creating countervailing effects on the information benefits to brokerage.

Although network diversity predicts both the diversity and the total amount of novel information actors receive, the coefficient on network diversity drops by 66% when network size and channel bandwidth are added to the specification. A one standard deviation increase in channel bandwidth was associated with a .35 standard deviation increase in total non-redundant information received, while a one standard deviation increase in network diversity was only associated with a .07 standard deviation increase. These results imply that while structural diversity and channel bandwidth both have 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 the size of actors' contact networks and their channel bandwidth, drawing attention to the importance of the thickness of communication channels and the number of contacts in providing larger total volumes of novel information.

To investigate how the diversity-bandwidth trade-off behaved in different information environments, we examined the effects of the refresh rate, the size of the topic space and information overlap on relationships between network diversity, channel bandwidth and access to novel information. Implications of variation in the refresh rate are shown in Table 6, Models 5 and 11. When the refresh rate of alters' information increased, recruiters received more novel information and channel bandwidth had a stronger

effect on the volume of novel information received (Model 11). In other words, as alters' information changed more from day to day, higher bandwidth ties to those alters delivered more total non-redundant information. Interestingly, the refresh rate did not have the same effect on the average diversity of information received – the variance of topics (Model 5).

As the topic space of alters' information increased, recruiters received more total non-redundant information from their contacts and greater channel bandwidth provided even more total non-redundant information than when the alters' topic space was smaller (Model 12). Communicating through thicker channels with those who know about many topics affords an ability to sample more information on distinct topics. As these models are estimated using fixed effects specifications the variation comes from changes in the topic space of a recruiters' alters over time. As the topic space of recruiters' contacts increased, they received more novel information and their high bandwidth ties were even more valuable in delivering more novel information.

These two results highlight why the distinction between information diversity (as a measure of variance) and total non-redundant information (as a measure of volume) is important. Although having more samples of alters' topic space per period increased the number of novel topics sampled and the total volume of novel information received, it did not change the variance of the distribution of topics from which recruiters were sampling. Recruiters who increased the bandwidth of their communication channels saw increases in the total amount of novel information they received, but not necessarily in diversity per unit information. When maintenance costs are considered, that implies actors must weigh the benefits of additional novel information against the costs of obtaining that information, which makes the functional form of the relationship between novel information and performance particularly salient – a relationship we consider in more detail below.

Finally, as the overlap of alters' information topic spaces increased network diversity was less useful for delivering more total non-redundant information (Table 6, Model 13). Perhaps surprisingly, greater information overlap in an ego network was associated with greater access to non-redundant information. Upon reflection it is clear why this relationship is positive. As the topic spaces of alters grew

larger they were more likely to overlap, but they were also more likely to contain more total novel information and to thus offer more novel topics to ego. This is confirmed by the fact that when topic space was added to the specification in Model 13, the information overlap variable did not significantly predict total non-redundant information.

In summary, network diversity and channel bandwidth both predict access to more diverse information and more total non-redundant information, although bandwidth is a more powerful predictor of the total volume of novel information received. As alters' topic spaces grew larger and changed more rapidly, bandwidth became more important for delivering novel information. Finally, the more alters' information overlapped, the less important network diversity became to delivering novel information.

PERFORMANCE EFFECTS

Table 6 displays strong evidence of a positive relationship between access to non-redundant information and performance, as measured by revenues generated per month, projects completed per month and average project duration.[35] We estimated both random effects and fixed effects specifications, but in the interests of space only provide the more conservative fixed effects results in the text. Random effects estimates were all in the same direction and stronger than the fixed effects results.

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

As recruiters' structural diversity and channel bandwidth increased, they fulfilled contracts more quickly, fulfilled more contracts per unit time and generated more revenue.[36] A one standard deviation increase in the bandwidth of communication channels was associated with just over $1,500 more revenues generated (per person per month) (Model 7, $p < .01$) and an additional two-tenths of a project completed (Model 4, $p < .05$). The performance effects of network structure were enabled in large part by the provision of non-redundant information. When non-redundant information was added to the specifications, the performance effects of network structure were reduced and non-redundant information strongly predicted performance across all dimensions. A one standard deviation increase in the amount of non-redundant information flowing to individuals was associated on average with just over $2,900 more in

revenues generated (Model 8, $p < .01$), an extra one-tenth of one project completed (Model 5, $p < .01$), and an average project duration that is 12 days shorter, per person per month (Model 2, $p < .01$). These results offer evidence that diverse networks provide access to diverse, non-redundant information, which in turn predicts performance. As a robustness check, we estimated the relationships between information diversity (the variance measure) and performance with very similar results. A one standard deviation increase in information diversity was associated with increases in revenues (β_{FE} = 1322.97, N.S.; β_{RE} = 2254.75, $p < .01$) and project completions (β_{FE} = .036, $p < .05$; β_{RE} = .049, $p < .01$), and with reductions in average project duration (β_{FE} = - 16.04, $p < .01$; β_{RE} = - 15.78, $p < .01$).

We also uncovered evidence of alternative mechanisms linking network structure to performance. Holding access to novel information constant, network diversity was associated with more completed projects (Model 5, $p < .05$) and faster project completion (Models 2-3, $p < .01$). These results leave open the possibility that some benefits to network diversity come not from access to novel, non-redundant information, but rather from other mechanisms, such as access to job support, power or organizational influence (Burt 1992). It is interesting that network diversity seems to affect project duration more than revenues or the amount of projects completed per unit time. We suspect that network diversity enables recruiters to get the diverse information and resources they need to finish a given project faster, but that without a greater volume of novel information, to support needs for more information relevant to more projects, a greater number of projects are not completed per unit time.

Across the board, access to non-redundant information had diminishing marginal performance returns for each of our performance measures (Models 3, 6 and 9). 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 9 demonstrate, there seem to be negative returns to more novel information beyond the normalized mean.[37] These nonlinearities in the value of novel information likely arise for at least two reasons. First, beyond the threshold for decision relevance, new information adds no value. Second, employees' capacity to process new information can

be constrained as excess novel information becomes burdensome or distracting. These explanations are consistent with theories of bounded rationality, limited cognitive capacity and information overload.

----- FIGURE 9 -----

DISCUSSION

Structural theories of social capital and brokerage have developed to a significant extent around intuitions and anecdotal evidence about how information is likely to be distributed in networks and how different types of information are likely to accrue to individuals in different structural positions (Simmel 1922 (1955), Moreno 1940, Granovetter 1973, Baker 1990, Burt 1992, Padgett and Ansell 1993, Uzzi 1996, 1997, Podolny 2001, Reagans and Zuckerman 2001, Hansen 1999, 2002, Zuckerman and Reagans 2008a, b, Burt 2008). However, the actual information flowing between individuals is rarely observed (Burt 2008), and we lack detailed dynamic theories of how social groups access, share and distribute information under different network and environmental conditions.

This article develops a theory of how social actors gain access to novel information that accounts for how stocks of information are distributed in a network as well as how information flows between contacts. Specifically, we propose that a trade-off exists between gathering novel information through more diverse network structure and gathering it through higher bandwidth communication channels. As diversity and bandwidth counterbalance one another, it is difficult to increase both simultaneously. Structurally diverse networks tend to deliver information that exhibits more variation across channels because there tends to be information homogeneity within connected social groups (Simmel 1922 (1955), Granovetter 1973, Blau 1986, Burt 1992). However, diverse networks also tend to include weaker ties (Granovetter 1973) that lack cooperative norms (Granovetter 1985, Coleman 1988) and display less multiplexity and dimensionality (Hansen 1999, Uzzi 1997), making them likely to deliver less information diversity and less total non-redundant information through each channel over time. We show—intuitively, analytically and empirically—that this trade-off creates countervailing effects on access to diverse novel information.

Statistical analyses that combine social network and performance data with direct observation of the information content flowing through e-mail at a medium-sized executive recruiting firm provide

strong evidence in support of the diversity-bandwidth trade-off. As network diversity increased channel bandwidth fell, and both diversity and bandwidth delivered novel information. In accordance with existing theory, reaching across structural holes provides access to novel information. As recruiters communicated across structural holes, they tended to tap contacts with varied expertise and to receive more diverse information from them. However, they paid for this diversification by foregoing communication bandwidth which, all else equal, reduced the total volume of novel information they received through thicker bandwidth channels.

We also found support for each of the three main environmental conditions hypothesized to moderate the effects of the diversity-bandwidth trade-off. When the information of recruiters' contacts changed more rapidly from day to day and when they were aware of a larger number of topics, bandwidth was even more influential in providing access to novel information. On the other hand, when the information overlap between recruiters' contacts was higher, network diversity had a greater impact on access to novel information. These findings suggest that information benefits, vision advantages and returns to brokerage are contingent on the information environments in which brokers find themselves. The prevailing wisdom among sociologists for the last forty years has been that the strength of weak ties and information advantages to brokerage operate with a fair degree of regularity across contexts (Centola and Macy 2007). In contrast, our analysis shows that context matters. In certain information environments, brokers with many bridging ties to disparate parts of a social network can have *disadvantaged* access to novel information because their lower bandwidth communication constrains the volume of novelty they receive.

High bandwidth channels are more important in turbulent environments where information changes rapidly. Several implications follow from that result. First, the prevailing view that information redundancy exists in dense cohesive networks ignores the fact that the information each actor has may be changing rapidly at the same time, even when holding constant changes created by information coming to them from social contacts. A densely connected group of arbitrageurs in New York might all know each other well but may also constantly get new information from one another because what each person

knows changes moment by moment. Second, weak, structurally diverse ties provide information at a lower rate, with less frequency, less complexity and more delay. Weak ties are more likely to deliver obsolete information. In the classic case of job market opportunities or other time-sensitive settings such as stock traders exchanging tips, fewer relevant and useful opportunities are likely to be delivered by diverse-low bandwidth ties. That is compounded by the fact that our cohesive-high bandwidth ties are more likely to know what we need and are therefore more likely to volunteer relevant information in a timely manner. Such thinking highlights the importance of timely access to novel information (rather than access alone) as a factor in brokerage theory.

The dependence of vision advantages on information turbulence suggests two important questions: which social environments are more turbulent, and is society moving toward greater overall information turbulence? The implications for brokerage are clear – if turbulence makes high bandwidth channels more important for access to novel information, then vision advantages from brokerage positions are less likely in social and economic sectors where the general stock of knowledge changes rapidly. If turbulence increases population heterogeneity, then diverse structure can remain salient. But if that is not the case and society is moving toward greater information turbulence, then over time brokerage positions may become less useful than leadership positions in cohesive cliques. Turbulent environments in which key environmental variables change quickly or a large number of new events occur within a given period of time have been described as post-industrial (Bell 1973, Huber 1984), high-velocity (Eisenhardt 1989) and time sensitive (Glazer and Weiss 1993) and are typically associated with markets where information technology plays a critical role (Glazer 1991). Incorporating the rate of environmental change and information turbulence into brokerage theory could explain why brokerage is salient in some industries but not others.

High bandwidth channels also deliver more non-redundant information in high dimensional information environments in which knowledge is complex and comprised of many distinct topics. It is not surprising that evidence contradicting the predictions of brokerage theory typically emerges in R&D (Reagans and McEvily 2003), innovation (Obsfeld 2005) and the creative arts (Uzzi & Spiro 2005). In

those settings new novelty is produced by exploiting interactions between complex complementary ideas. In high dimensional information environments, innovation is born of union strategies that connect alters. Prior work describes those effects as resulting from complex interactions, collaborations and brainstorming, which are all more likely to occur in dense cohesive networks in which strong ties are prevalent. Our work provides an additional underlying mechanism supporting this argument: increasing the volume of novel information flowing between collaborators provides even greater support for innovation in high dimensional information environments. In contrast, in environments where efficiency is more important than innovation, weaker ties are sufficient. Having been asked to provide "the one thing you would change to improve [the company's] supply chain management" in 2000 characters or less, supply chain managers possessing networks rich in structural holes provided answers that were scored higher in peer evaluations (Burt 2004). We speculate that these contrasting results can be explained by the complexity of the innovation and ideas being solicited in the different contexts. Simple good ideas come more easily to brokers, but complex innovation that requires coordinating high dimensional interdependent information requires high bandwidth communication.

An important question raised by the benefit of bandwidth in high dimensional information environments is whether it is more important to develop thick bridges or wide bridges, where a 'thick bridge' refers to a high bandwidth tie to a socially distant community and a 'wide bridge' refers to several reinforcing weak ties to a socially distant community. Centola and Macy (2007) contend that, because adoption of complex behaviors requires social affirmation and reinforcement, exposure from multiple different contacts is the key structural characteristic of bridges across structural holes that enables diffusion of complex contagions. But, our results show that thick, high bandwidth bridges are critical to the amount of complex novel information that traverses a tie. The open question is whether a bundle of multiple weak ties is the same as one strong tie of equal channel bandwidth, both in the types of information they deliver and their role in social reinforcement and affirmation? Is the width or rather the thickness of a bridge more important for the movement of complex, high dimensional information or the diffusion of complex contagions via social reinforcement? To answer these questions, the importance of social reinforcement

through multiple weak ties and rich interactions through high bandwidth ties must be considered simultaneously. It could be that social reinforcement not only depends on multiple exposures but also on the transfer of rich information from trusted sources, which happens less often over low bandwidth channels. Social reinforcement from multiple casual acquaintances may be less important than social reinforcement from one trusted peer. High bandwidth ties could therefore also explain the tendency of social movements to diffuse spatially (Centola and Macy 2007). Our results imply that, however rare, the most important tie for access to novel information in high dimensional information environments is the thick bridge – a high bandwidth tie to a distant network neighborhood.

Finally, information-based mechanisms do in fact explain performance benefits to brokerage. Network structure explains access to novel information which in turn explains variation in performance. These results confirm prior theory and represent some of the first quantitative evidence of an information-based mechanism explaining returns to brokerage. As recruiters accessed more diverse information (variance) and more total non-redundant information (volume), they generated more revenue, completed more projects per unit time, and completed projects faster. These results held even in conservative fixed effects specifications, and were stronger in random effects models that also evaluated variation across recruiters. An important limitation is that we cannot make causal claims about the relationship between access to information and performance (Aral et al 2009, Aral 2010, Aral and Walker Forthcoming). In order to identify these relationships, future work could exploit random exogenous variation in the receipt of novel information to examine whether access to information actually causes performance increases, or if top performers are simply magnets for information. More detailed theoretical development and new empirical inquiry in different contexts will no doubt shed further light on these and other tradeoffs. Toward this end, our methods for analyzing network structure and information content in e-mail data are replicable, opening a new line of inquiry into the information mechanisms that make social networks valuable.

CONCLUSION

The importance of weak ties is that they connect individuals to socially distant ideas and novel information. Access to novel information provides vision advantages to individuals that connect socially distant network neighborhoods. These two inferences have for decades guided sociologists' thinking on information flow in networks. However, our research shows that as networks become more structurally diverse their communication channel bandwidth contracts, and that this trade-off regulates the degree to which structurally diverse networks deliver non-redundant information to actors in brokerage positions. As individuals communicate across structural holes they tend to tap contacts with varied expertise and to receive more diverse information from them. However, they pay for that diversification by foregoing communication bandwidth, which on balance reduces the total volume of novel information they receive. In turbulent and high dimensional information environments, the diversity-bandwidth trade-off implies that brokers with bridging ties to disparate parts of a social network may actually have *disadvantaged* access to novel information because their lower bandwidth communication curbs the total volume of received novelty. Our findings therefore suggest that information advantages to brokerage are contingent on the information environments in which brokers find themselves.

References

Allen, T.J. 1977. *Managing the Flow of Technology*. MIT Press, Cambridge, MA.

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

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

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

Aral, S. & Walker, D. Forthcoming. "Creating Social Contagion through Viral Product Design: A Randomized Trial of Peer Influence in Networks." *Management Science*.

Aral, S. 2010. "Identifying Social Influence: A Comment on Opinion Leadership and Social Contagion in New Product Diffusion." *Marketing Science*. In Press.

Aral, S., Muchnik, L., & Sundararajan, A. 2009. "Distinguishing Influence Based Contagion from Homophily Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences*, Dec. 22, 2009, vol. 106, no.51.

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

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

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

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

Bell, D. 1973. The Coming of Post-Industrial Society. Basic Books, New York.

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

Blau, PM. 1986. Exchange and Power in Social Life. Transaction Publishers. London, UK.

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

Burt, R. 1987. "Social Contagion and 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, and Sutton, R. (Ed.), *Research in organizational behavior* (Vol. 22). New York, NY, JAI Press.

Burt, R. 2004a. "Structural Holes and 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 and Closure: An Introduction to Social Capital. Oxford University Press. New York, NY.

Burt, R. 2008. "Information and structural holes: comment on Reagans and Zuckerman." *Industrial and Corporate Change*, 17(5): 953-969.

Centola, D., and Macy, M. 2007. "Complex Contagions and the Weakness of Long Ties." *American Journal of Sociology.* 113 (3) (November): 702–34.

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

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

Cramton, C.D. 2001. "The Mutual Knowledge Problem and its Consequences for Dispersed Collaboration." *Organization Science*, 12(3), 346-371.

Cummings, J., and 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.

Currarini, S., Jackson, M.O., and Pin, P. 2009. "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica*, 77(4): 1003-1045.

Currarini, S., Jackson, M.O., and Pin, P. 2010. "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences*, 107(11): 4857-4861.

Drew, P. and J. Heritage. 1992. Talk at Work: Interaction in Institutional Settings. Paul Drew and John Heritage editors. Cambridge University Press. Cambridge, UK.

Driscoll, J. C., and A. C. Kraay. 1998. Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data. *Review of Economics and Statistics*, 80: 549–560.

Eliasoph, N. 1996. "Making a Fragile Public: A Talk-Centered Study of Citizenship and Power." Sociological Theory, (14): 262-289.

Eisenhardt, K. 1989. “Making Fast Strategic Decisions in High Velocity Environments,” *Academy of Management Journal*, 32: 543-76.

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

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

Gibson, D.R. 2005. “Taking Turns and Talking Ties: Networks and Conversational Interaction.” *American Journal of Sociology*, 110(6): 1561-1597.

Glazer, R. 1991. “Marketing in Information-Intensive Environments: Strategic Implications of Knowledge as an Asset,” *Journal of Marketing*, 55: 1-19.

Glazer, R. and Weiss, A.M. 1993. “Marketing in Turbulent Environments: Decision Processes and the Time-Sensitivity of Information,” *Journal of Marketing Research,* 30(4): 509-521.

Goffman, E. 1961. Encounters. Bobbs-Merrill, New York.

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 and 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 and R.G. Eccles (eds.), *Networks and 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. and R, Sutton. 1997. “Technology brokering and innovation in a product development firm.” *Administrative Science Quarterly*, (42): 716-49.

Heider, F. 1958. The Psychology of Interpersonal Relations. Wiley. New York.

Helper, S. MacDuffie, J.P., and C. Sabel. (2000). “Pragmatic Collaborations: Advancing Knowledge While Controlling Opportunism.” *Industrial and Corporate Change*, 9(3): 443-488.

Hinds, P. and Kiesler, S. 2002. Distributed Work. MIT Press. Cambridge, MA.

Huber, G. 1984. “The Nature and Design of Post-Industrial Organizations,” *Management Science*, 30: 928-51.

Lave, J. and Wenger, E. 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, New York.

Lazer, D. and Friedman, A. 2007. The Network Structure of Exploration and Exploitation. *Administrative Sciences Quarterly*, 52: 667-694.

Liang, D.W., Moreland, R. and Argote, L. 1995. “Group Versus Individual Training and Group Performance: The Mediating Role of Transactive Memory.” *Personality and Social Psychology Bulletin*, 21(4): 384-393.

Lichterman, P. 1996. The Search for Political Community: American Activists Reinventing Commitment. University of Cambridge Press, Cambridge, UK.

Lin, N. 2002. Social Capital: A Theory of Social Structure and Action. Cambridge University Press, Cambridge.

Kempe, D., Kleinberg, J., and 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. 1987. “Cognitive Social Structures.” *Social Networks*, 9(2): 109-134.

Krackhardt, D. 1990. “Assessing the Political Landscape: Structure, Cognition and Power in Organizations.” *Administrative Science Quarterly*, 35: 342-369.

Krackhardt, D. and 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. and D. Watts. 2006. “Empirical Analysis of an Evolving Social Network.” *Science* (311:5757): 88-90.

Kossinets, G. and D. Watts. 2009. “Origins of Homophily in an Evolving Social Network.” *American Journal of Sociology*, 115(2): 405-50.

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

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

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

McLean, P. 1998. “A Frame Analysis of Favor Seeking in the Renaissance: Agency, Networks, and Political Culture.” *American Journal of Sociology*, (104): 51-91.

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

Mische, A. and H.C. White. 1998. “Between Conversation and Situation: Public Switching Dynamics Across Network Domains.” *Social Research*, 65(3): 695-724.

Mische, A. 2000. “Cross-Talk in Movements: Reconceiving the Culture-Network Link.” *Conference on Social Movement Analysis: The Network Perspective*, Ross Priory, Loch Lomond, Scotland.

Montgomery, J.D. 1991. “Social Networks and Labor-Market Outcomes: Towards and Economic Analysis.” *American Economic Review*. 81(5): 1408-1418.

Newcomb, T.M. 1961. The Acquaintance Process. Holt, Rinehart and Winston. New York.

Obstfeld, D. 2005. “Social networks, the tertius iungens orientation, and involvement in innovation.” *Administrative Science Quarterly*, 50, 100–130.

Padgett, J.F., and C.K. Ansell. 1993. “Robust Action and 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.

Putnam, R.D. 1995. “Bowling Alone: America’s Declining Social Capital,” Journal of Democracy: 65-78.

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

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

Reagans, R. and Zuckerman, E. 2008a. "Why Knowledge Does Not Equal Power: The Network Redundancy Trade-off." *Industrial and Corporate Change*, 17(5): 903-944.

Reagans, R. and Zuckerman, E. 2008b. "All in the family: reply to Burt, Podolny, van de Rijt, Ban, and Sarkar." *Industrial and Corporate Change*, 17(5): 979-999.

Reynolds, M., Van Alstyne, M. Aral, S. 2009. “Privacy Preservation of Measurement Functions on Hashed Text” Annual Security Conference. Dhillon, G. *“Discourses in Security Assurance & Privacy*," Las Vegas, NV. *April 15-16, 2009*: Information Institute Publishing. p 41-45.

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

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

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

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

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

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

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

Steinberg, M.W. 1999. "The talk and Back Talk of Collective Action: A Dialogic Analysis of Repertoires of Discourse among Nineteenth-Century English Cotton Spinners." *American Journal of Sociology*, (105): 736-780.

Swedberg, R. 1990. Economics and Sociology. Princeton, NJ: Princeton University Press.

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

Tsai, W. and Ghoshal, S. 1998. "Social Capital and Value Creation: The Role of Intrafirm Networks," Academy of Management Journal, 41(4): 464-476.

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.

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

Van Alstyne, M. and 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. and Zhang, J. 2003. "E-mailNet: A System for Automatically Mining Social Networks from Organizational E-mail Communication," NAACSOS.

Walker, G., Kogut, B. and Shan, W. "Social Capital, Structural Holes and the Formation of an Industry Network." *Organization Science*; 8(2): 109-125.

Watts, D. and 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.

Weick, K. 1979. *The social psychology of organizing*. Addison-Wesley, Reading, MA.

Weitzman, M. 1998 "Recombinant Growth." *Quarterly Journal of Economics*, 113(2): 331-360.

Wenger, E. 1987. "Transactive Memory: A Contemporary Analysis of the Group Mind" Chapter 9 in Theories of Group Behavior, B Mullen & G Goethals (eds): 185-208.

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

White, H. C. 1995. "Network Switchings and Bayesian Forks: Reconstructing the Social and Behavioral Sciences." *Social Research*. (62): 1035-1063.

White, H. C. 2000. "Modeling Discourse in and around Markets." *Poetics*. (27): 117-135.

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

Tables

Table 1. Implications of the Diversity-Bandwidth Trade-off for Access to Novel Information		
Theory	**Ego's Perspective**	**Alter's Perspective**
Social Capital (e.g. Putnam 1995, Burt 1992, Tsai & Ghoshal 1998, Lin 2002)	Greater intimacy, trust, reciprocity and cooperation in cohesive-high bandwidth networks makes ego more willing to request novel information from alter.	Greater intimacy, trust, reciprocity and cooperation in cohesive-high bandwidth networks makes alter more willing to share novel information with ego.
Transactive Memory (e.g. Wegner 1987, Liang et. al. 1995)	Awareness of whom to ask and what to ask for in cohesive-high bandwidth networks enables ego to request novel information more effectively from alter.	Awareness of what to volunteer in cohesive-high bandwidth networks enables alter to volunteer relevant novel information more effectively to ego.
Search-Transfer (e.g. Hansen 1999)	Close, frequent interaction and tight coupling in cohesive-high bandwidth networks makes ego better able to comprehend and thus receive novel information from alter.	Close, frequent interaction and tight coupling in cohesive-high bandwidth networks makes alter able to express and thus transfer novel information to ego.
Knowledge Creation (e.g. Uzzi 1996, 1997, Uzzi & Spiro 2005, Obstfeld 2005)	Embeddedness and cohesion enable ego to find synergies and connections between her information and alter's information in order to generate new ideas and new novel information.	Embeddedness and cohesion enable alter to find synergies and connections between her information and ego's information in order to generate new ideas and new novel information.
Homophily (e.g. McPherson et. al. 2001, Blau 1986, Uzzi 1997, Helper et. al. 2000)	Alters are more likely to have mutual interests with ego across a wider variety of topics inspiring multifaceted communication and access to more of the different dimensions of alters' information.	Ego is more likely to have mutual interests with alters across a wider variety of topics inspiring alter to communicate more of the different dimensions of their information.

Table 2. Summary of Hypotheses		
Domain	**Hypothesis**	**Hypothesized Relationship**
The Diversity-Bandwidth Trade-off	**H1a**	*Network diversity is positively associated with receiving more diverse information and more total non-redundant information.*
	H1b	*Network diversity is associated with lower channel bandwidth.*
	H1c	*Channel bandwidth is positively associated with receiving more diverse information and more total non-redundant information.*
	H2a	*The greater the information overlap among alters, the less valuable structural diversity will be in providing access to novel information.*
	H2b	*The broader the topic space, the more valuable channel bandwidth will be in providing access to novel information.*
	H2c	*The higher the information refresh rate, the more valuable channel bandwidth will be in providing access to novel information.*
Performance Effects	**H3**	*Access to non-redundant and diverse information is positively associated with individual performance.*

Table 3: 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 E-mails	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
Expertise Heterogeneity	560	.86	.07	.51	.97
Channel Bandwidth	555	5.87	4.13	0	51
Alters' Information Refresh Rate	564	34.24	25.97	0	178.84
Alters' Topic Space	564	46.59	35.06	0	214.67
Information Overlap of Alters	564	310.11	362.35	0	3292.83
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 4: Pairwise correlations Between Variables

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
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 E-mail	-.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. Expertise 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. Refresh Rate	-.33*	-.11*	-.26*	-.13*	.95*	.29*	.94*	.61*	.34*	.22*	.09*	.47*	1.00				
14. Topic Space	-.34*	-.11*	-.30*	-.15*	.97*	.30*	.97*	.62*	.33*	.23*	.03	.50*	.97*	1.00			
15. Information Overlap	-.25*	-.07	-.19*	-.07	.85*	.20*	.85*	.71*	.30*	.09*	.15*	.27*	.85*	.85*	1.00		
16. Revenue	.44*	-.02	.33*	.15*	-.09*	.23*	-.12*	-.12*	.27*	-.16*	.12*	-.05	-.11*	-.14*	-13*	1.00	
17. Completed Projects	.41*	-.01	.29*	.11*	-.09*	.23*	-.11*	-.09*	.25*	-.14*	.10*	-.07	-.11*	-.13*	-13*	.92*	1.00
18. Average Project Duration	.50*	.12*	.49*	.21*	-.30*	.14*	-.31*	-.07	.18*	-.21*	.07	-.14*	-.32*	-35*	-28*	.54*	.47*

* $p < .05$

Table 5. The Network Diversity – Channel Bandwidth Trade-off

Dependent Variable:	***Channel Bandwidth***					***Network Diversity***				
Model:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Specification:	*FE*	*FE*	*FE*	*RE*	*RE*	*FE*	*FE*	*FE*	*RE*	*RE*
Age					-.006					-.007
					(.011)					(.006)
Gender					-.113					-.129*
					(.136)					(.070)
Education					-.123**					-.012
					(.059)					(.030)
Industry Experience					-.014					.012*
					(.012)					(.006)
Partner					.145					.155
					(.284)					(.145)
Consultant					-.231					.174
					(.217)					(.110)
Network Diversity	-.314***	-.288***	-.335**	-.286***	-.190*					
	(.078)	(.095)	(..136)	(.089)	(.111)					
Structural Equivalence	.107**	.101*	.105*	.167***	.149**	-.055	-.033	-.021	-.034	-.073**
	(.042)	(.052)	(.055)	(.054)	(.068)	(.038)	(.031)	(.030)	(.0267)	(.030)
Expertise Heterogeneity		-.074	-.095*	-.209***	-.141**	.254***	.132***	.127***	.132***	.122***
		(.056)	(.049)	(.058)	(.068)	(.019)	(.019)	(.025)	(.029)	(.031)
Network Size			.213	.476***	.398**		.747***	.753***	.965***	.911***
			(.231)	(.171)	(.199)		(.102)	(.102)	(.074)	(.078)
Network Size Squared			-.123	-.345**	-.333**		-.445***	-.446***	-.612***	-.543***
			(.149)	(.142)	(.161)		(.073)	(.073)	(.065)	(.066)
Channel Bandwidth								-.076***	-.069***	-.037*
								(.028)	(.021)	(.021)
Constant	.181***	.192***	.164***	.105	2.793***	.342***	.199***	.202***	-.034	.530
	(.046)	(.047)	(.047)	(.124)	(.972)	(.036)	(.040)	(.037)	(.062)	(.503)
Temporal Controls	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	45938***	125.91***	788.77***	74.63***	76.63***	112301***	2479.08***	1820***	425.64***	408.09***
(d.f.)	(11)	(12)	(14)	(13)	(19)	(11)	(13)	(13)	(13)	(19)
R^2	.09	.09	.10	.20	.24	.23	.35	.35	.55	.59
Observations	536	535	535	535	429	538	538	535	535	429

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

Table 6. Effects of Network Diversity and Channel Bandwidth on Access to Diverse, Novel Information

Dependent Variable:	***Information Diversity***							***Non-Redundant Information***					
Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Specification	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*
Total E-mail Incoming	.002**	.000	-.001	-.002*	-.002**	-.005***	-.002						
	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)						
Expertise Heterogeneity	.281***	.073***	.028	.037*	.035*	.032	.037*	.181***	.099***	-.011	-.010	.0002	-.007
	(.054)	(.023)	(.019)	(.021)	(.019)	(.020)	(.021)	(.026)	(.030)	(.016)	(.019)	(.018)	(.023)
Network Diversity		.232***	.145*	.155*	.161**	.165**	.139**		.206***	.066**	.016	.045**	-.036
		(.062)	(.083)	(.083)	(.075)	(.071)	(.095)		(.072)	(.032)	(.016)	(.016)	(.043)
Structural Equivalence		.012	.019	.021	.024	.027	.017		-.032	-.007	.012	.007	.008
		(.042)	(.045)	(.036)	(.040)	(.040)	(.035)		(.041)	(.049)	(.007)	(.004)	(.033)
Network Size			.439***	.505***	.485***	.488***	.468***			.668***	.231***	.115***	.161**
			(.102)	(.105)	(.069)	(.069)	(.091)			(.102)	(.030)	(.019)	(.050)
Network Size-Squared			-.250***	-.259***	-.257***	-.256***	-.216***						
			(.065)	(.054)	(.051)	(.053)	(.067)						
Channel Bandwidth				.085**	.081**	.084**	.080**			.352***	.120***	.060***	.201***
				(.041)	(.033)	(.032)	(.040)			(.067)	(.030)	(.011)	(.051)
Refresh Rate					.028						.642***		
					(.070)						(.040)		
Channel Bandwidth x Refresh Rate					-.016						.077***		
					(.037)						(.009)		
Topic Space						.210						.832***	
						(.146)						(.044)	
Channel Bandwidth x Topic Space						-.009						.037**	
						(.036)						(.017)	
Information Overlap							-.003						.638***
							(.037)						(.079)
Network Diversity x Information Overlap							-.038						-.230**
							(.043)						(.076)
Constant	-.456***	.004	.064	.165	.248**	.408***	.152	-.172***	.371***	-.131***	-.012	-.045***	.215***
	(.068)	(.088)	(.077)	(.111)	(.106)	(.093)	(.107)	(.021)	(.041)	(.037)	(.017)	(.014)	(.033)
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	Month / Year
F-Value / Wald χ^2	6.6e4***	6994***	492.8***	97.32***	80.06***	374.1***	269.9***	4.6e5***	1.0e5***	333.5***	9845***	2976***	27960***
(d.f.)	(11)	(13)	(15)	(16)	(18)	(18)	(18)	(10)	(12)	(14)	(16)	(16)	(16)
R^2	.14	.11	.14	.16	.16	.16	.16	.18	.19	.56	.86	.91	.75
Observations	556	538	538	535	535	535	535	556	538	535	535	535	535

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

Table 7. Performance Effects of Network Diversity, Channel Bandwidth and Non-Redundant Information

Dependent Variable:	***Project Duration***			***Projects Completed***			***Revenues Generated***		
Model:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Specification	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*	*FE*
Network Diversity	-16.02***	-14.15***	-12.28***	.037***	.02**	.01	802.2	365.44	-420.4
	(3.90)	(3.39)	(3.53)	(.010)	(.01)	(.01)	(520.15)	(544.69)	(479.47)
Bandwidth	-5.35	-2.41	-1.57	.023**	.002	-.01	1585.3***	899.25*	544.47
	(2.92)	(4.22)	(4.05)	(.009)	(.01)	(.01)	(493.79)	(467.81)	(517.18)
NRI		-12.64***	-19.00***		.10***	.16***		2947.46***	5626.2***
		(4.72)	(5.19)		(.02)	(.02)		(617.80)	(933.27)
NRI Squared			7.70***			-.07***			-6858.8***
			(1.91)			(.01)			(1545.01)
Constant	307.96***	305.04***	302.62***	.64***	.66***	.60***	27865***	28545***	27447***
	(.287)	(1.17)	(2.74)	(.004)	(.004)	(.01)	(276.15)	(227.28)	(438.83)
Temporal Controls	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month/ Year	Month / Year
F-Value	1.8e6***	4.6e3***	4.3e3***	1.29e7***	1.4e5***	5.2e4***	1.78e6*	2.6e4**	3.2e4***
(d.f.)	(11)	(12)	(13)	(11)	(12)	(14)	(11)	(12)	(13)
R^2	.08	.09	.10	.05	.08	.14	.05	.06	.10
Obs.	416	416	416	416	416	416	416	416	416

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

Figures

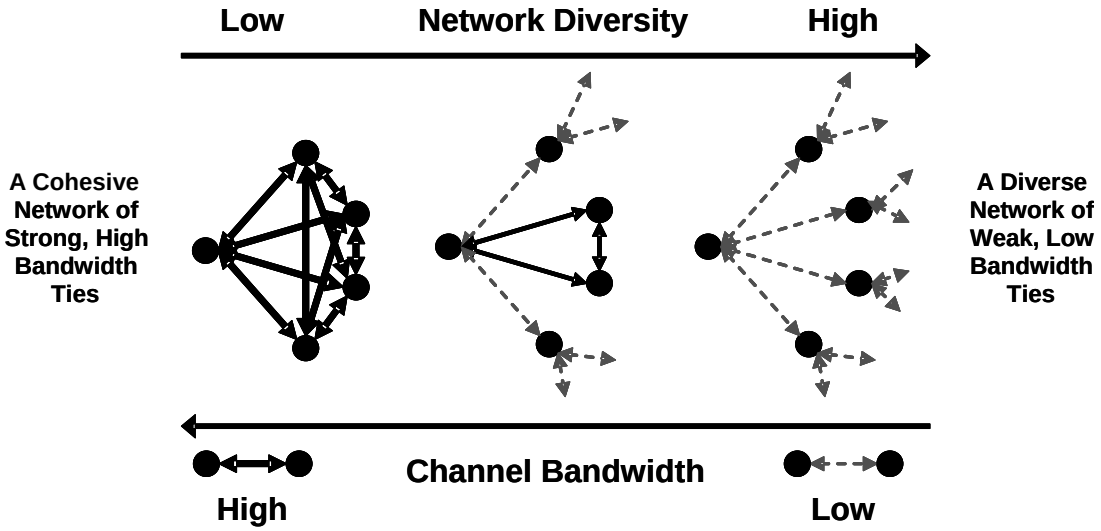


Figure 1. The Diversity – Bandwidth Trade-off: As structural diversity increases, channel bandwidth decreases.

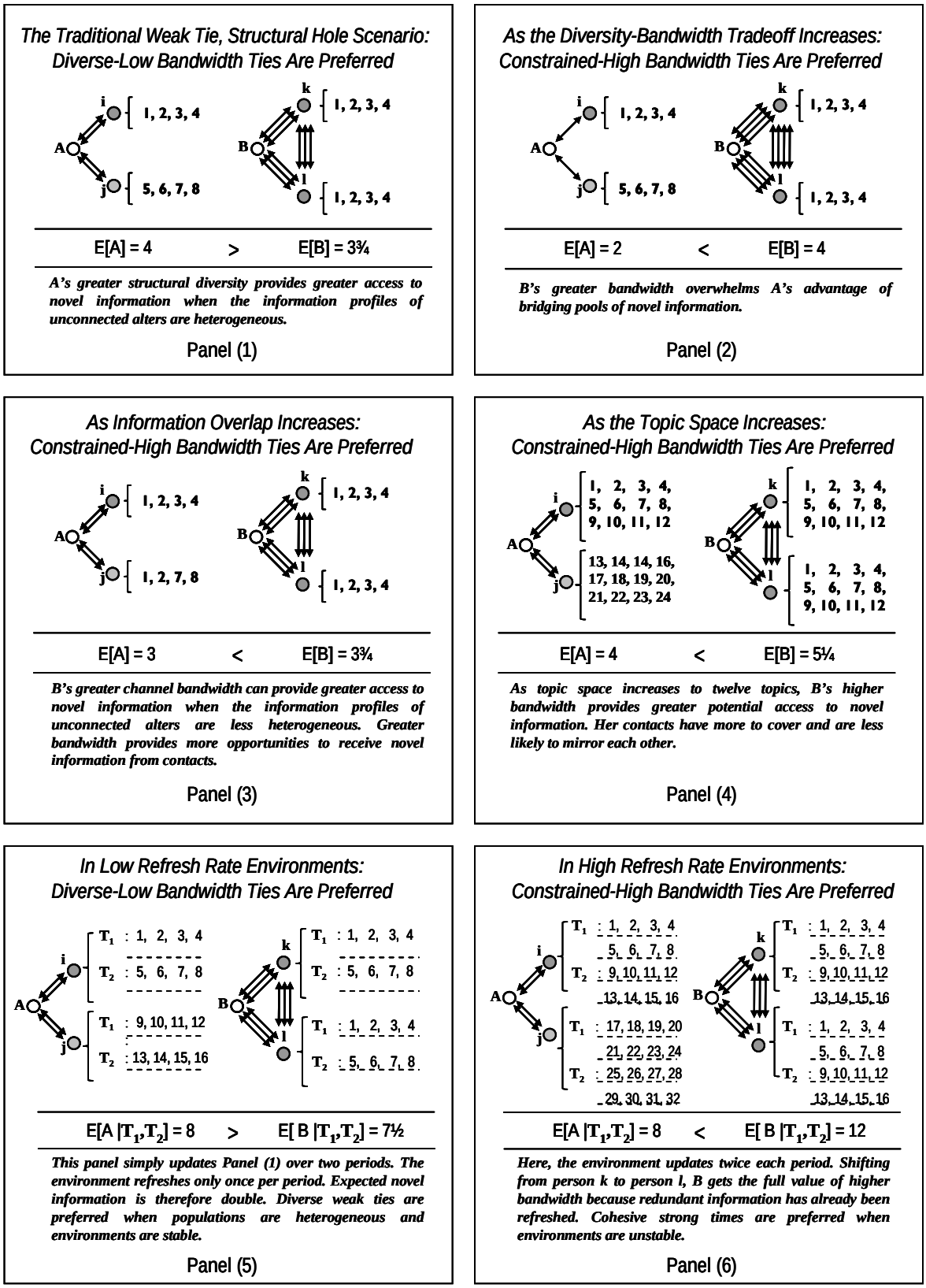


Figure 2. The Diversity – Bandwidth Trade-off under varying information environments: 1) a base case, 2) as the strength of the trade-off increases, 3) as the information overlap of alters increases, 4) as the topic space increases and in panels 5) and 6) as the refresh rate of alters' information increases.

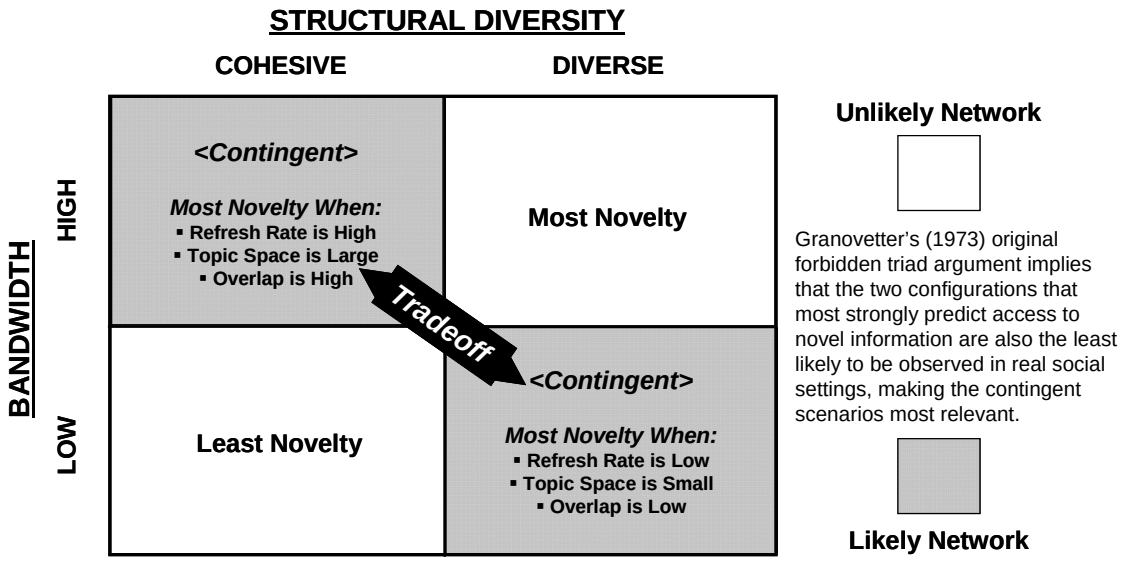
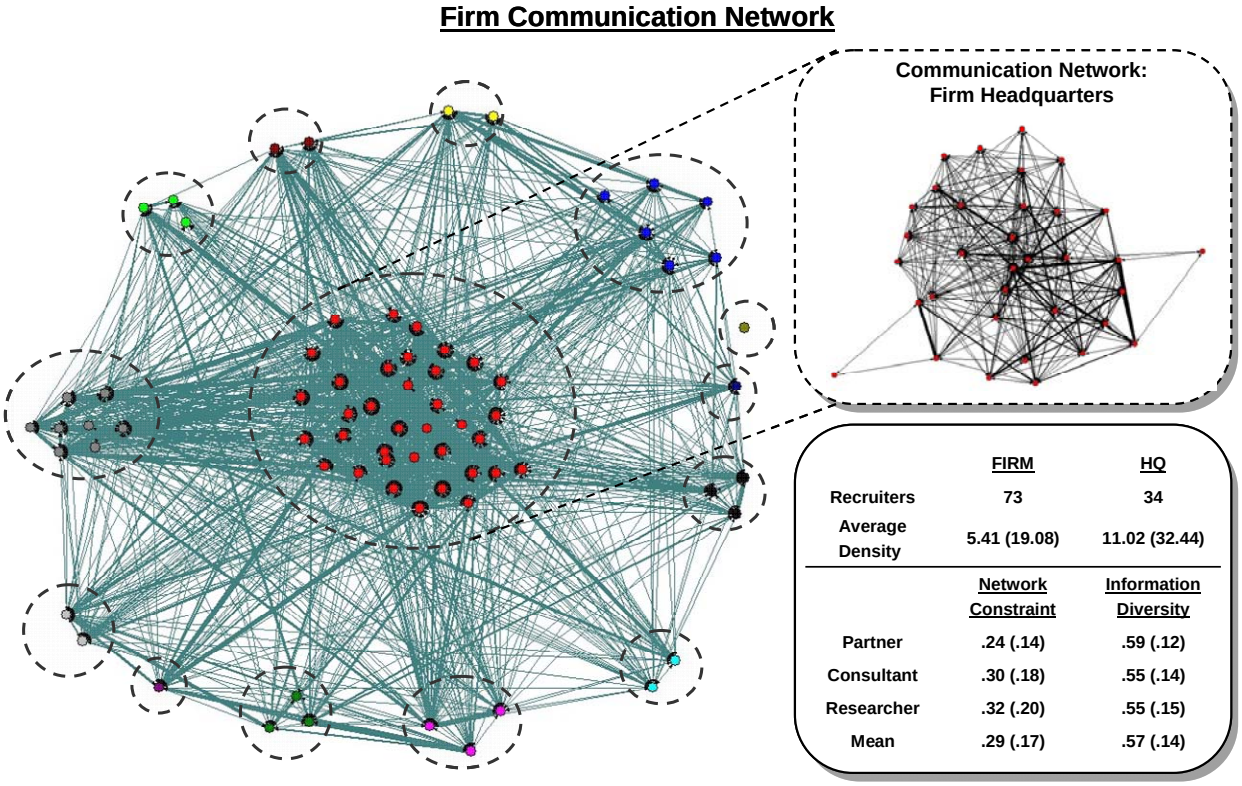


Figure 3. Granovetter's (1973) original forbidden triad argument implies that the two configurations that most strongly predict access to novel information (Diverse High Bandwidth ties and Cohesive Low Bandwidth Ties) are also the least likely to be observed in real social settings, making the contingent scenarios the most relevant. Cohesive-High Bandwidth networks deliver the most novel information when the refresh rate of information is high, when the topic space is large and when information overlap between alters is low.



	FIRM	HQ
Recruiters	73	34
Average Density	5.41 (19.08)	11.02 (32.44)

	Network Constraint	Information Diversity
Partner	.24 (.14)	.59 (.12)
Consultant	.30 (.18)	.55 (.14)
Researcher	.32 (.20)	.55 (.15)
Mean	.29 (.17)	.57 (.14)

Figure 4. The e-mail 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.

Date: Sun, 01 Feb 2009 10:02:10 -0500
From: xxx@yyy
To: abc@123
CC: zzz@yyy,
abc@zzz
Subject: ~~Re:~~ IWP 1 Extended Thoughts

Actually,

~~To be~~ even more succinct about 3 main take aways:

1. Given ~~our~~ ability ~~to~~ make connections between abstract concepts, ~~our~~ productivity ~~is~~ determined more ~~by our~~ ability ~~to~~ multitask, than ~~by our~~ ability ~~to~~ conduct sequential work faster.

~~So~~, lets explore ~~the~~ mechanisms behind multitasking ~~a~~ bit more:

2. ~~The~~ relationship between output ~~and~~ multitasking ~~is~~ convex ~~at~~ low levels ~~of~~ multitasking ~~and~~ concave ~~at~~ high levels ~~of~~ multitasking (~~Because~~ information inputs ~~are~~ non-rival ~~and~~ complementary, unlike physical inputs, ~~their~~ use enables convexity ~~in the~~ relationship between multitasking ~~and~~ output ~~at~~ low levels ~~of~~ multitasking. Because human information processing ~~is~~ constrained ~~by~~ bounded rationality, ~~and~~ limited cognitive capacity ~~the~~ relationship between multitasking ~~and~~ output ~~is~~ concave ~~at~~ higher levels ~~of~~ multitasking).

~~So~~, how ~~do we~~ acquire ~~the~~ inputs ~~we~~ use?... Socially:

3. Efficient positioning ~~in the~~ social network creates efficient means ~~to~~ gather ~~and~~ use information ~~and is~~ correlated ~~with~~ higher productivity. [~~Because we~~ require ~~a~~ social support system ~~of~~ information acquisition (embodied ~~in our~~ social networks) which ~~we~~ rely ~~on to~~ extend ~~our own~~ individual mental capacity. ~~We~~ gather information inputs socially (~~and~~ through IT ~~which we use as a~~ control variable)]

Dr. XXX

Construction of E-mail Vectors

1. Header information is extracted to create the social network. Names are matched and identities are validated by hand.
2. The subject and body of the e-mail message are analyzed to extract frequencies of use of keywords (Steps 3-6).
3. Stop words (e.g. "a, "an," "the," "and," and other common words) with high frequency across all e-mails are removed as shown by words that have been struck.
4. Keywords are extracted based on the three principles outlined on pages 30-31 of the manuscript.
5. Keywords are root-stemmed, such that for example "multitask," "multitasking," become "multitask*."
6. The frequency of each key word is counted and recorded.
7. A vector representing the e-mail is created which logs the e-mail ID, the ID number of each keyword used and the frequency of use of each keyword noted inside brackets as follows:
   <E-mailID7842B|748821<9>; … ; 849247<2>|>
   A vector representing the example e-mail to the left is shown in truncated form below.
8. The content similarity of e-mail vectors is then compared using several standard distance metrics such as the Cosine distance.

$$\overrightarrow{d_i} = (IWP < 1 >; connection* < 1 >; productivity < 2 >; multitask* < 9 >; sequential < 1 >; output < 3 >; ...; input* < 4 >; social* < 5 >; IT < 1 >; control < 1 >; \text{variable} < 1 >)$$

Figure 5. An example e-mail is shown on the left. The step by step method used to construct e-mail vectors is described on the right.

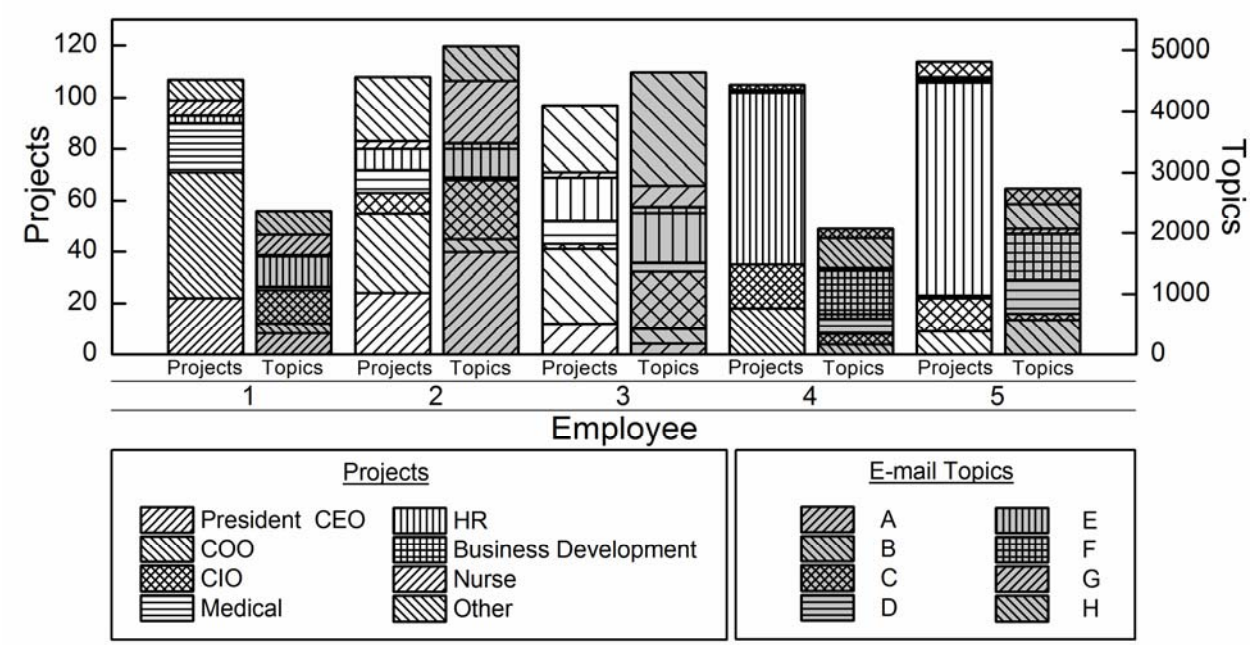


Figure 7. Distributions of Project Expertise and Topics in Incoming E-mail. Panel (a) displays histograms of (a) the number of projects each of five recruiters has worked on over the eight project categories recognized by the firm and (b) the number of incoming e-mails that discuss each of eight topics A though H (an e-mail can contain reference to more than one topic) for each of five recruiters.

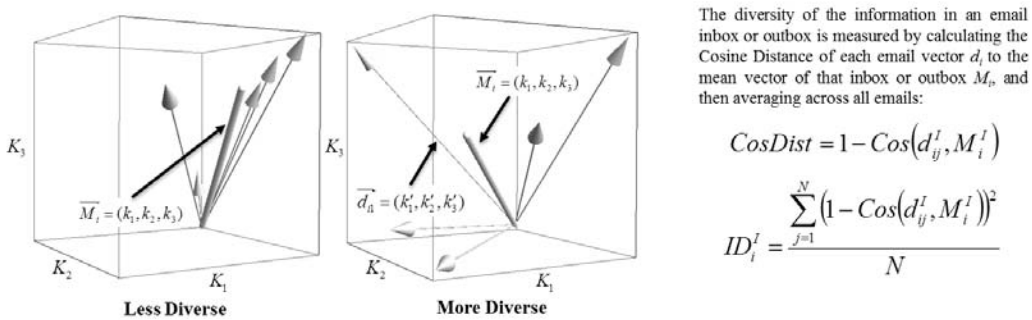


Figure 6. A three dimensional Vector Space Model of five e-mail vectors and their mean vector for inboxes containing relatively more diverse and less diverse information is shown on the left. A summary of how information diversity was calculated is shown on the right.

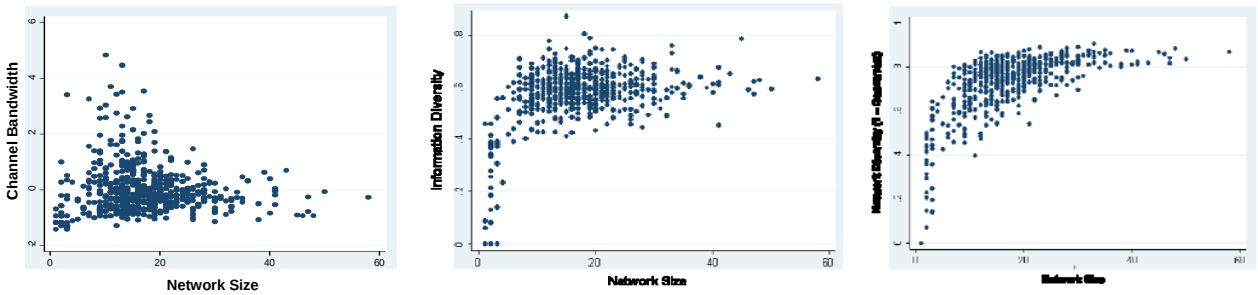


Figure 8. Relationships between network size, bandwidth, network diversity and information diversity.

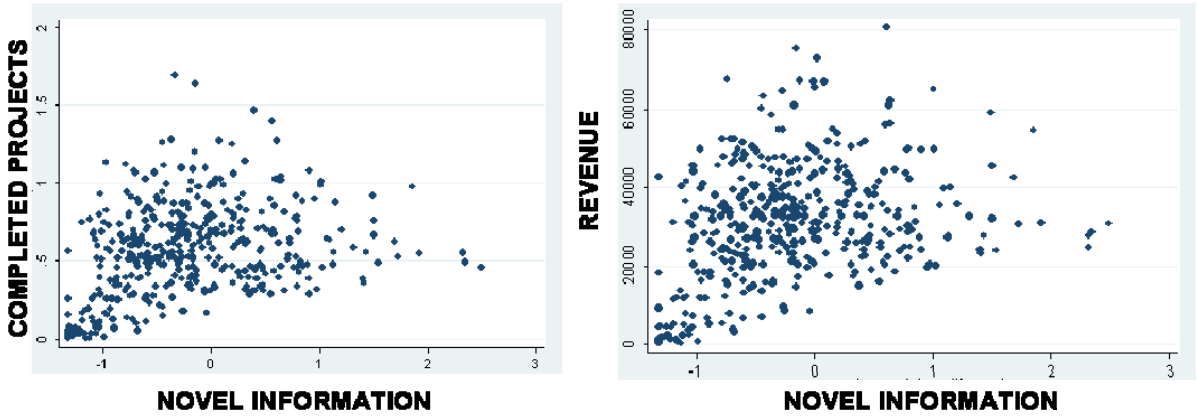


Figure 9. Graphs of the relationships between novel information, completed projects and revenue.

Appendix A.

Models of the Diversity-Bandwidth Trade-off: Proofs of Consequences for Information Acquisition

To make our claims precise regarding the Diversity-Bandwidth Trade-off, this section provides probabilistic models of information acquisition. Such models also address the "prior knowledge" problem. In a deterministic model, a person needs to know who knows what exactly in order to ask for it. In a probabilistic model, a person only needs to know the best expected contact policy over the population without having to know which person knows what. This represents an even softer constraint than one imposed by a "transactive memory" model where ego needs to know which alter to ask. Yet, still bandwidth to a cohesive tie can be favored over weak access to a diverse tie. More generally, these models demonstrate when a constrained-high bandwidth tie can be expected to provide greater novelty than a diverse-low bandwidth tie, and vice versa. Obviously, a diverse (unconstrained) high bandwidth tie is best but we wish to show how various degrees of constraint affect the diversity-bandwidth trade-off. Analysis proceeds in three parts, one each for topic space, information overlap, and refresh rate as shown in Figure 2. Topic space is simplest so we first generalize Panel 4. "Biasing information overlap" generalizes Panels 1-3. Finally, temporal analysis generalizes Panels 5 and 6.

Without loss of generality, normalize the capacity of the low-bandwidth channel to 1 and that of the high-bandwidth channel to $B>1$. In Panel 1, this means the weak tie has bandwidth 1 and the strong tie has bandwidth 1.5.

1. Topic Space

Consider two diverse low-bandwidth ties, with completely non-overlapping information, and two constrained high-bandwidth ties, with complete overlap.

Proposition 1.1:

As the number of topics T grows without bound, constrained high-bandwidth ties provide strictly greater expected novelty than diverse low-bandwidth ties.

Proof:

Two non-overlapping low-bandwidth ties provide 2 normalized units of novel information. Novelty on the first high-bandwidth channel is B. Complete overlap on the second high-bandwidth channel implies only T-B previously unshared items remain. Without repeating herself (i.e. without replacement), the second high-bandwidth contact has the potential to reveal new items according to a hypergeometric distribution with draw capacity B. From standard probability, the mean of a hypergeometric distribution is $\frac{B(T-B)}{T}$ or simply B-B^2/T. Total expected novelty across two high-bandwidth channels is thus $2B$-B^2/T, which has limit $2B$ as T grows without bound. Since $2<2B$ for $B>1$, this proves the claim. ■

As an aside, equation $2B$-B^2/T implies that optimal bandwidth across both ties is T.

2. Information Overlap

To generalize insights concerning information overlap from Panels 1-3, we introduce more flexible notation for information sharing "bias." Let there be 1 … n_i topics in in-group topic set n_i and 1 … n_o topics in out-group topic set n_o for a total of $n_i+n_o = T$. Define the likelihoods of encountering n_i and n_o topics as p_i and p_o respectively. It follows that $n_ip_i + n_op_o = 1$. An actor receives "biased" content if she is more likely to receive news on one set of topics than another ($p_i > p_o > 0$), which we use to characterize the increased likelihood that cohesive high bandwidth ties discuss the same things. She can also receive *un*biased content ($p_i = p_o = 1/T$), or completely biased content ($p_i = 0 < p_o = 1/n_o$ and $p_i = 1/n_i > p_o = 0$). If ide-

as in n_i become x times more likely to appear among in-group communications, then $p_i=x/T$ (which implies that $p_o=\frac{1-n_1 x/T}{T-n_1}$ with $n_i < T$, $x<T$, and $xn_i \leq T$).

With this terminology, we can derive $P(\Psi_{biased})$, the probability of encountering a new idea given that there are k ideas remaining to be seen, allowing differences in p_i and p_o. Let E represent the event that a person encounters new information through a new link. Since novelty depends on what one has learned from prior links, let L represent links. Then, define the following:

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

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

$$\Psi = \{\text{Event that link } L\text{+1 connects to a new idea}\}$$

Here, J_k indicates whether idea k has failed to appear among the information provided by any of the social links 1 … L. $P(\Psi)$ can then be constructed as follows.

$$\begin{aligned} P(\Psi) &= E[P(\Psi \mid J_1 ... J_k)] \\ &= E[\sum_{l=1}^{n_i} J_l p_i + \sum_{h=n_1+1}^{T} J_h p_o] \\ &= n_i p_i E[J_l] + n_o p_o E[J_h] \\ &= n_i p_i (1-p_i)^L + n_o p_o (1-p_o)^L \end{aligned} \qquad [1]$$

The last step arises because an idea that occurs with probability p must not have occurred in any of the previous L draws. It is useful to note three properties of $P(\Psi_{biased})$. First, unbiased information implies $p_i = p_o = 1/T$. Because unbiased ties provide equal access across all topics, unbiased chances of encountering a new idea simplify to $P(\Psi_{unbiased})=(1-1/T)^L$. Second, having no prior links L=0 implies that a new idea is encountered with certainty. Third, increasing links without bound $L\rightarrow\infty$ implies the chances of encountering a new idea approach 0. The likelihood of encountering novel information (for both biased and unbiased ties) decreases strictly and asymptotically toward 0 with each additional tie L. This theoretical model exactly mirrors the pattern we observe empirically as shown in Figure 8.

Proposition 2.1:
When the advantage of bandwidth swamps the disadvantage of bias, an ego prefers the constrained-high bandwidth tie to the diverse-low bandwidth tie to increase the chances of encountering novel information.

Proposition 2.2:
When the disadvantage of bias swamps the advantage of bandwidth, an ego prefers the diverse-low bandwidth tie to the constrained-high bandwidth tie to increase the chances of encountering novel information.

Proof:
Let $P[E^c]= P(\Psi_{biased})$ and $P[E^D]= P(\Psi_{unbiased})$, where E^c and E^D represent the events of forging a constrained and a diverse link and getting new information with a single unit of bandwidth. 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 a given interval. Simplifying with $n_o=T-n_i$ gives total accumulated probability of:

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

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:

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

This demonstrates the first claim that a constrained-high bandwidth tie can supply a greater volume of novel information than a diverse-low bandwidth tie provides. To see when a diverse-low bandwidth tie could be preferred, consider when extreme bias results in topic heterogeneity. The subset of n_i topics occurs with probability $p_i = B/T$ (such bias necessarily constrains $p_o \approx \varepsilon$). For ease of simplification, let $n_i = T/B$. Then algebra reduces relative probabilities to:

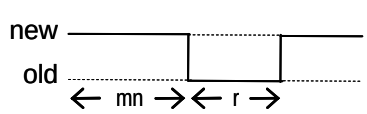
$$P[E_L^c] = \left(1 - \frac{B}{T}\right)^L < \left(1 - \frac{1}{T}\right)^L = P[E_L^D] \quad [4]$$

This alternative case demonstrates the counterclaim, that a diverse-low bandwidth tie can supply a greater volume of novel information than a constrained high bandwidth tie provides. Although $P[E_L^C] = P[E_L^c] + \ldots P[E_{L+B}^c]$ 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 of constrained ties, there always exists a fixed number of links L such that link $L+1$ should be an unconstrained tie. This establishes the second claim. ■

While a range of intermediate cases span these two extremes, conditions exist when a person will always prefer one or the other type of link depending on bias, bandwidth, and the number of links already present.

3. Refresh Rate

To model information renewal, we consider a standard Poisson process. Let mn be the average time between samples in a subset of n topics and r be the average time until news is refreshed.



The chance a given sample produces new information is the ratio of average time between samples over the total time until information renews $\frac{mn}{mn + r}$. Note also, if m is average time between samples, then $1/m$ gives the number of samples per unit time. Consider non-overlapping in-group and out-group topic subsets n_i and n_o with shorter sampling times for the higher bandwidth in-group.

Proposition 3.1:
Let high bandwidth ties provide more frequent updates $m_i < m_o$ (so sample times are shorter) but let low bandwidth ties provide access to distinct topics n_o not included in n_i. Then among ties balanced to provide optimal access to news, an increase in the refresh rate favors an increase in high bandwidth ties.

Proof:
We first find the optimal balance between in-group and out-group ties then show the proportion of in-group ties grows in refresh rate. Since the number of ties is finite, the sum of samples per unit time is bounded by the number of social links $1/m_i + 1/m_o \le L$. To get the most news, a person chooses:

$$\max \pi = (inGroup\,Samples) * P(inGroup\,Success) + (outGroup\,Samples) * P(outGroup\,Success)$$

$$\max_{m_i, m_o} \pi = \frac{1}{m_i}\left(\frac{m_i n_i}{m_i n_i + r_i}\right) + \frac{1}{m_o}\left(\frac{m_o n_o}{m_o n_o + r_o}\right)$$

$$\text{such that: } 1/m_i + 1/m_o \le L$$

Use boundary condition $1/m_o = L - 1/m_i$ to substitute for m_o. Since refresh occurs sooner in higher bandwidth ties, substitute for out-group refresh rate using $r_o = \delta r_i$, $\delta \ge 1$. Solving $\frac{\partial \pi}{\partial m_i} = 0$ produces a quadratic equation with two roots, of which the second is out of bounds.

$$\frac{1}{m_i} = \left\{\frac{L\delta n_i}{n_o + \delta n_i}, -\frac{n_i(2n_o + \delta L r_i)}{(n_o - \delta n_i) r_i}\right\}$$

Based on the first root, the absolute time to spend on in-group ties rises in the number of ties L, the relative refresh delay on out-group topics δ, and the number of in-group topics n_i. It falls in the number of out-group topics n_o. The *proportion* of time to spend on high bandwidth ties is

$$\frac{1/m_i}{1/m_i + 1/m_o} = \frac{\delta n_i}{n_o + \delta n_i}$$

which increases strictly toward 1 as the refresh delay of out-group topics increases. Higher refresh rates strictly favor a higher proportion of high bandwidth ties. ■

Appendix B.

Descriptions and Correlations of Information Diversity Metrics

1. Information Diversity (ID)

Variance based on cosine distance (cosine similarity):

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

We measure the variance of deviation of e-mail topic vectors from the mean topics vector and average the deviation across e-mails 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

$$DistDice(d) = DiceDist(d, M) = 1 - Dice(d, M), \text{ and where}$$

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

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 e-mail 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 e-mails were clustered together divided by the number of clustering iterations. Therefore, every two e-mails 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 e-mails reside. AvgCommonIC extends the AvgCommon concept by compensating for the different amount of information provided in the fact that an e-mail resides in the same bucket for either highly diverse or tightly clustered clusters. For example, the fact that two e-mails are both in a cluster with low intra-cluster diversity is likely to imply more similarity between the two e-mails than the fact that two e-mails 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 e-mails:

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

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

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

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

Correlations Between the Five Measures of Information Diversity

Measure	1	2	3	4	5
1. InfoDiversity	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 C1.[38] 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.[39] After creating this independent data set, 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 C1 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.[40] Overall, these results give us confidence in the ability of our diversity measurement to characterize the subject diversity of groups of text documents of varying sizes.

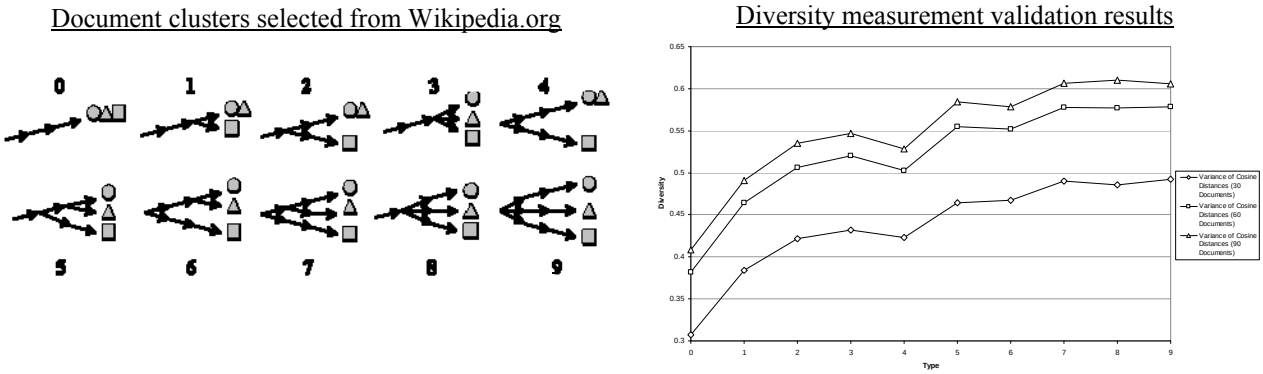


Figure C1. Wikipedia.org Document Clusters and Diversity Measurement Validation Results.

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: Model Specifications and Estimation Procedures

Model Specifications

To explore the mechanisms driving the creation and appropriation of information advantages from network structure we first explicitly considered the trade-off between network diversity and channel bandwidth by estimating the two following specifications:

$$B_{it} = \gamma_i + \beta_1 ND_{it} + \beta_2 SE_{it} + \beta_3 EH_{it} + \beta_4 NS_{it} + \beta_5 NS_{it}^2 + \sum_j B_j HC_{ji} + \sum_m B_m M_{it} + \varepsilon_{it} \qquad [1]$$

$$ND_{it} = \gamma_i + \beta_1 SE_{it} + \beta_2 EH_{it} + \beta_3 NS_{it} + \beta_4 NS_{it}^2 + \beta_5 B_{it} + \sum_j B_j HC_{ji} + \sum_m B_m M_{it} + \varepsilon_{it} \qquad [2]$$

where B_{it} represents channel bandwidth, ND_{it} represents network diversity (measured by one minus constraint), SE_{it} represents average structural equivalence, EH_{it} represents the knowledge heterogeneity of i's contacts, NS_{it} represents the size of i's network, NS_{it}^2 represents network size squared, $\sum_j \beta_j HC_{ji}$ represents controls for human capital and demographic variables (Age, Gender, Education, Industry Experience, and Managerial Level), and $\sum_m \beta_m M_{it}$ represents temporal controls for each month/year. If network diversity and channel bandwidth trade-off, network diversity should be associated with lower channel bandwidth, and we would expect to observe parameter estimates such that $\beta_1 < 0$ and $\beta_2 > 0$ in equation 1 and $\beta_5 < 0$ in equation 2.

We then examined the structural correlates of access to diverse and novel information. We first estimated an equation relating network structure to the diversity of information flowing into actors' e-mail inboxes.[1] The estimating equation is specified as follows:

$$ID_{it}^I = \gamma_i + \beta_1 E_{it}^I + \beta_2 EH_{it} + \beta_3 ND_{it} + \beta_4 SE_{it} + \beta_5 NS_{it} + \beta_6 NS_{it}^2 + \beta_7 B_{it} + \sum_j \beta_j HC_{ji} + \sum_m \beta_m M_{it} + \varepsilon_{it} \quad [3]$$

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*. We then examined the relationship between network structure and the total amount of novel information flowing into actors' e-mail inboxes (NRI_{it}^I) using the following model:[2]

$$NRI_{it}^I = \gamma_i + \beta_1 EH_{it} + \beta_2 ND_{it} + \beta_3 SE_{it} + \beta_4 NS_{it} + \beta_5 B_{it} + \sum_j \beta_j HC_{ji} + \sum_m \beta_m M_{it} + \varepsilon_{it} \quad [4]$$

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

$$P_{it} = \gamma_i + \beta_1 ND_{it} + \beta_2 B_{it} + \beta_3 NRI_{it}^I + \beta_4 \left(NRI_{it}^I\right)^2 + \sum_j B_j HC_{ji} + \sum_m B_m Month + \varepsilon_{it} \qquad [5]$$

If information benefits to network diversity and channel bandwidth exist, they should be positively associated with access to diverse and non-redundant information, and non-redundant information should be positively associated with performance. If network structure confers additional benefits beyond information advantage (such as power or favorable trading conditions) network diversity and channel bandwidth should contribute to performance beyond their contribution through information diversity.[3] Finally, if there are diminishing returns to novel infor-

1 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.

2 We did not include the network sized squared term because it had no explanatory power. The relationship between network size and total non-redundant information is linear and positive.

3 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

mation, we should see a concave relationship between novel information and productivity. As a robustness check we also estimated equation [5] replacing the non-redundant information variable (NRI_{it}^{I}) with incoming information diversity (ID_{it}^{I}) with similar results.

Estimation Procedures

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 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 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:

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it} \,. \qquad [6]$$

The fixed effects transformation is obtained by first averaging equation 10 over t = 1,…, T, to create the cross section equation or between estimator:

$$\overline{y_i} = \alpha_i + \overline{x_i}\beta + \overline{\varepsilon_i} \,, \qquad [7]$$

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

By subtracting equation 7 from equation 6, the fixed effects transformation removes unobserved time invariant individual specific heterogeneity embodied in α_i :

$$y_{it} - \overline{y_i} = (x_{it} - \overline{x_i})\beta + \varepsilon_{it} - \overline{\varepsilon_i} \,. \qquad [8]$$

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 as robustness checks with similar results.

---

errors clustered by individual. Clustered robust standard errors are robust to correlations within observations of each individual, but are never fully efficient. They are conservative estimates of standard errors.

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:

$$y_{it} - \overline{\lambda y_i} = (x_{it} - \overline{\lambda x_i})\beta + \varepsilon_{it} - \overline{\lambda \varepsilon_i} \,. \qquad [9]$$

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.

To adjust for non-independence of observations in network panel data, we employ a consistent covariance matrix estimator which is robust to very general forms of network, spatial and temporal dependence (Driscoll and Kraay 1998). This approach is similar to common network autocorrelation models considered in the literature (e.g. Ord 1975, Doreian 1980, 1989, Dow et al 1982, Loftin and Ward 1983, Marsden and Friedkin 1993) but also takes into account temporal dependence across panels in longitudinal data as well as both cross sectional and time dependent network autocorrelation. The estimator assumes a data generating process with both contemporaneous and lagged cross-sectional dependence across observations as follows:

$$y_{it} = x_{it}\beta + \varepsilon_{it} \text{; where} \qquad [10]$$

$$\varepsilon_{it} = \lambda_i f_t + v_{it} \text{; and}$$

$$f_t = \rho f_{t-1} + u_{it} \,;$$

where u_{it} and v_{it} are mutually independent normal random variables with mean zero and contemporaneous and lagged cross-sectional dependence in disturbances is modeled through the presence of the unobserved factor f_t. The extent of dependence between two observations *i* and *j* depends of the strength of the network autocorrelation terms λ_i and λ_j and the degree of temporal persistence in the factor ρ.

Endnotes

1 We define the "structural diversity" or "network diversity" of an ego network as the extent to which it is low in "constraint" as defined by Burt (1992: 55), low in the average structural equivalence of alters and rich in structural holes. We define the "structural cohesion" or "network cohesion" of an ego network as the extent to which it is high in "constraint" as defined by Burt (1992: 55), low in structural holes and high in the average structural equivalence of alters. Various phrases have been used in the literature to describe analogous concepts including ego density (or sparseness) and network embeddedness. These definitions and their measures are highly correlated with and change in proportion to network diversity and network cohesion. We chose to use the phrases network diversity and network cohesion because they are the ones most commonly used in the literatures to which we refer (e.g. Burt 1992, 2004, 2005, Reagans and McEvily 2003). At times we also use the terms embeddedness and constraint to highlight that our arguments draw from and contribute to literatures that also use those terms (e.g. Granovetter 1985, Uzzi 1996, 1997, Burt 1992).

2 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, individuals 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 and volume, observed over actual communications channels, in developing the theories that follow. 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 and 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.

3 As Granovetter (1973: 1362) notes, homophily could also explain closure without a causal relationship between the strength of ties and closure, breaking the causal relationship between structural diversity and the rate and volume of interaction (if individuals interact more with similar others because they are similar and not because they are connected in embedded relationships). However, prior empirical work on friendship formation demonstrates that exposure and preferences both play highly significant roles in tie formation (see Currarini et al 2009, 2010). The diversity bandwidth trade-off can therefore be viewed to a significant extent as a causal theory, with structure driving the rate and volume of interaction. Exposure and motivation are likely to play an even bigger role in our setting because we study work relationships in which, as we explain in our empirical analyses, recruiters seek diversity constrained by exposure in order to perform well at work.

4 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 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."

5 We are indebted to one of our reviewers for this helpful insight.

6 Since social people talk, it could be the case that Beth tells Lauren what she learned from Kim in order that Lauren shares her non-redundant information in a single draw. But then Beth would already have used her three units of bandwidth. Allowing more targeted requests in a transactive memory sense (Wegner 1987) complicates the analysis but in no way invalidates the basic bandwidth tradeoff.

7 The same insight follows exactly if we instead decrease the overlap of Kim and Lauren rather than increase that of Isaac and Jake.

8 In the appendix, we formally prove an even stronger claim. If a weak tie can access all topics in S and a strong time can only access in-group subset ni $\subseteq$ S, then the strong tie can still provide more access to novel information than the weak tie, provided bandwidth exceeds a specific threshold.

9 The complete chain is given as: (9/12)+(9/12)(8/11)+(9/12)(8/11)(7/10)+(9/12)(3/11)(8/10)+(3/12)(9/11)+(3/12)(9/11)(8/10) +(3/12)(2/11)(9/10) = (9/4). As shown in Appendix A, a more straightforward approach is to use the mean of the hypergeometric distribution which gives equivalently (3)(12-3)/12=(9/4).

10 Although this assumes sequential attention across alters in a given period, the main insights do not change assuming ego attends simultaneously to all alters. To model simultaneous draws without replacement on a given alter, use a hypergeometric distribution, then estimate expected non-overlap. Given complete non-overlap, the numbers for ego A are unchanged across all six panels. For B in panel 1, total expected novelty from each alter is 15/8 for a total of 15/4 over both alters. In the sequential draw, Lauren had much higher novelty than Kim, by virtue of getting attention first, namely 3 versus 3/4. In Panel 6, simultaneous draws over each alter provides 39/8 for a total of 9.75. This is lower than the sequential calculation but higher bandwidth still provides 1.75 more expected novel units of information than structural diversity. Thus, whether using simultaneous or sequential draws, primary intuitions do not change in these examples.

11 A significant body of related work in political sociology, research on social movements and on cognition and network structure has developed around networks and language. Some of this work examines discourse in markets (White 2000), dialogic processes (Steinberg 1999), framing practices (McLean 1998), civic talk (Eliasoph 1996), and commitment styles (Lichterman 1996) in social movements, as well as sociolinguistic approaches to conversational dynamics in social movements (see the work of Harrison White and Ann Mische, e.g. White 1995, Mische and White 1998, and Mische 2000). Work on cognition in networks (e.g. Krackhardt 1987, 1990) has examined content from the perspective of what is perceived in and through social networks, and conversation-analytic approaches have been used to examine the structure of interaction (e.g. Goffman 1961, Drew and Heritage 1992, Gibson 2005). We build on this related work by focusing specifically on the diversity and total novelty of information exchanged between actors in networks over time in order to examine the information mechanisms that explain returns to brokerage.

12 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). We also calculated the indegrees of missing nodes based on the choices of the non-missing nodes. We found that the indegrees (insize) of missing nodes were lower than those of non-missing nodes (average monthly mean indegree non-missing = 14.7; average monthly mean indegree missing = 10.7), however t-tests reveal no statistically significant differences between the two (t-statistic = -1.38; $p < .172$). Size is the raw number of contacts while degree is weighted by message counts. We thank an anonymous reviewer for this suggested robustness check.

13 Most employees we talked to reported that e-mail was their primary means of communication. Although we did not collect phone conversation data or face to face information exchanges, e-mail provides the best means of assessing codified communications between employees at this firm. That said, we took several steps to investigate whether use of the phone and use of e-mail were similar in the organization. First, our survey had asked employees to report "the number of people they communicated with on a typical day a) by phone and b) by e-mail." A Pearson correlation returned a .31 correlation which was significant at the $p < .001$ level, indicating the size of e-mail networks and phone networks was likely to be similar. However, this did not give us insight into network structure, so we went further. Second, we found three reasonable proxies for phone communication between two people. First, our interviews indicated that recruiters most often spoke with their project team members (more than other recruiters in the firm) both by e-mail and by phone. We therefore decided that if two people worked on the same project together, that it would be reasonable to expect they would talk on the phone. In fact, the more projects they worked on together, the more likely they would exchange a greater volume of phone traffic. We therefore constructed a network of "project co-work" which measured as the strength of a tie the number of projects two individuals in the firm had worked on together. Our interviews also indicated that work was frequently regionally clustered (in other words candidates typically looked for jobs in the same region that they were currently working). We therefore conjectured that if two recruiters worked in the same region they would be more likely to seek information from one another over the phone about candidates that might be interested in a specific job in that region. Similarly, if they worked in the same office, they may have reasons specific to the social workings of the office to exchange a higher volume of phone communication. We therefore also created two new matrices in which dyads shared a tie if they 'worked in the same region' or 'worked in the same office.' We took these three new matrices "Project Co-Work," "Same Region," and "Same Office" and used Quadratic Assignment Procedure (QAP) to assess QAP correlations and to analyze correlations via MRQAP with a pooled matrix of the total e-mail exchanged between these same individuals (a single pooled matrix of e-mail traffic over all 10 months of data). If these proxies for greater phone traffic (Project Co-Work, Same Region, and Same Office) were highly correlated with the e-mail adjacency matrix, then the e-mail network should approximate the phone network. The e-mail network was significantly correlated with the project co-work network (.426; $p < .001$) and with the same region network (.359, $p < .001$) which makes it likely that the e-mail network mirrors the phone network relatively well given that our interviews indicated recruiters talked more frequently via phone and e-mail to others on the same project or in the same region. Correlation with the 'same office' network was slightly lower (.148, $p < .001$) perhaps because it is less necessary to talk via phone with those in the same office, but also and perhaps most tellingly, because the co-work network and the same office network had the lowest correlation (.079, $p < .005$) reflecting the fact that project teams were typically geographically dispersed across different offices – again lending credibility to the argument that project co-work should be a better proxy for phone communication than simply being in the same office. These results mirror the MRQAP results which indicate that the project co-work network is the strongest predictor of the e-mail network (.339, $p < .01$) and the same region network is also a strong predictor (.225, $p < .01$), while the same office network was correlated but was not as strong a predictor (.084, $p < .05$). As our interviews revealed that recruiters talked on the phone most often with those who were on the same projects and in the same regions, the results of the QAP correlations and MRQAP analysis indicate that the e-mail network should mirror the phone network relatively well. A separate question is whether the same type of information is exchanged over the phone and over e-mail. However, the interview evidence that e-mail was the communication medium of choice in this setting gives us confidence that our results of e-mail analyses are the most important in this study with regard to access to information and the role of information in performance. Perhaps more importantly, phone communication data, if we had it, would likely only support our claims rather than detract from them. If the phone is a richer communication medium through which high bandwidth, high novelty information is likely to flow, then the social microprocesses arguments that predict high bandwidth communication in socially proximate relationships would simply be magnified in the telephone context. For example, we are less likely to have the social capital standing to 'cold call' a weak tie to ask for a significant amount of their time to give us detailed novel information, nor would such a tie likely call us out of the blue to volunteer such information. Several of the other social microprocesses operate in the same way in that they predict that social proximity enables high bandwidth exchanges which are likely to occur over the phone as well as over e-mail. However, future work should assess the differences between phone and e-mail networks.

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

15 We wrote and developed e-mail 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 and Zhang 2003 and Reynolds, Van Alstyne and Aral 2009 for more detail). Spam messages were excluded by eliminating external contacts who did not receive at least one message from someone inside the firm.

16 By measuring both first and second order network diversity we account for the possibility that small world networks (Watts and Strogatz 1998, Watts 1999), clustered cliques linked by infrequent weak ties, could bring novel information into a cohesive clique from contacts two steps removed from ego.

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

18 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.

19 While e-mail 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 e-mail is widely used. In our data, the average number of contacts by phone ($\rho = .31$, $p < .001$) are positively and significantly correlated with e-mail contacts. Our interviews indicate that in our firm, e-mail is a primary communication media.

20 Each e-mail may pertain to multiple topics based on keyword prevalence, and topic vectors representing e-mails 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 e-mails that may pertain to several topics of differing emphasis.

21 Another common weighting scheme is the 'term-frequency/inverse-document frequency.' However, we use a more sophisticated keyword selection refinement method described in detail in the text.

22 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.

23 The coefficient of variation is particularly useful due to its scale invariance, enabling comparisons of data sets, like ours, with heterogeneous mean values (Ancona and 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. C^i refers to the coefficient of variation of keyword *i*, m_c^i is the mean frequency of keyword *i* in e-mails in topic cluster *c*, and $\overline{M^i}$ is the mean frequency of keyword *i* across all topics.

24 *i* indexes keywords and *c* indexes *k*-means clusters of e-mails which represent topics. f_{ec}^i is the frequency of keyword *i* in e-mail *e* in topic cluster *c*, m_c^i is the mean frequency of keyword *i* in e-mails in topic cluster *c*, and $\overline{M^i}$ is the mean frequency of keyword *i* across all topics. We squared the variation to ease computation as in footnote 22.

25 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.

26 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*).

27 We measure the refresh rate of alters using both incoming and outgoing e-mail vectors to capture the degree to which information being received was changing and the degree to which alters changed the topics they sent information about over time. Both are likely to affect the effective refresh rate in ego's network. However, an argument could be made for only considering the refresh rate of information received by alters as a proxy for information they are privy to. We therefore created an alternative refresh rate measure that only considered alters' incoming e-mail. Use of that variable did not change the results significantly.

28 We exclude each alter's overlap with themselves, which would only add a constant to the measure as the cosine similarity of *j* to *j*, $Cos(M_{jt}, M_{jt})$ is always 1.

29 We also ran specifications 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.

30 To normalize the Expertise Heterogeneity measure so that its values range from zero to one, we scale the measure by multiplying the final metric by (8/7), creating this final metric: $KH_i = \frac{8}{7}\left[1 - \sum_{k=1}^{8}\left(\frac{q_{ik}}{q_i}\right)^2\right]$. 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 easing interpretation.

31 Multicollinearity is not a significant issue in our study for several reasons. First, we never include any of the variables with a high correlation in the same model making multicollinearity due to their simultaneous inclusion in an estimating equation unlikely. Second, we conducted Variance Inflation Factor (VIF) analysis, which provides a measure of the degree to which the variance of an estimated regression coefficient is increased due to multicollinearity, to quantify the severity of the effects of multicollinearity in our models. We examined the VIF for all coefficients in all of our specifications. None of the variables listed above ever generated a VIF greater than 5, which is well below the acceptable threshold of 10 noted by Kutner et al. (2004). The only two variables that ever recorded VIFs greater than 5 were the Network Size and Network Size Squared variables when they were simultaneously entered into regressions, which is common for nonlinear terms. We therefore estimated these regressions with the Network Size Squared term removed, which created no qualitative change in the parameter estimates or the significance of any of the variables in the models. We choose to include the Size Squared term however if it was significant in order to document the nonlinear effects of Size and to remove it in cases where it contributed no explanatory power. Third, the real danger of multicollinearity is to bias parameter estimates toward zero by inflating the variance of an estimated regression coefficient. As such, any variance inflation due to collinearity should only serve to make our estimates more conservative (by making confidence intervals wider) and to therefore make it more difficult to estimate statistically significant results. The parameters of interest are significant even if there is variance inflation due to collinearity. Fourth, aside from statistical concerns such as variance inflation due to multicolinearity, we also considered the theoretical implications of these constructs being highly correlated. The measures we have devised are theoretically and conceptually distinct and produce different results in our analyses which are theoretically interesting and make sense from the perspective of our theory. For example, in the case of Total Non-Redundant Information and Information Diversity, we discuss in the paper how these variables are conceptually related in that the first is a proxy for the "volume" of novel information while the second is a proxy for the "variance" of information or topics. We discuss in some depth why these distinctions are important, and why we believe they are significant in some models but not in others. We have also added to and clarified this discussion to bring it to the fore. Furthermore, the raw correlation is particularly uninformative summary statistic in this case. When we consider the correlation of these two variables over their joint distributions, we find that they are highly uncorrelated in particular regions and highly correlated in others. In particular, when the volume of novel information is low the diversity of information is low as well and they are highly correlated. However, they only display a correlation of .12 for values of information diversity greater than one standard deviation less than the

normalized mean. The zero order correlation is driven by their high correlation at low levels which makes sense. In the limit, when you receive no e-mail, you receive no novel information and the diversity of your incoming information is also zero. Epsilon perturbations from this limit exhibit similar high correlations. However, as we get into more interesting regions of the distributions, the correlation of these variables decreases significantly as expected and described above. The other variables you list exhibit similar behavior.

32 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."

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 correlated.

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

35 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.

36 Given the core-periphery structure of the e-mail network of this firm (displayed in Figure 4), 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 evidence that being in a peripheral office is associated with lower 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 geographic isolation is a time invariant binary characteristic. However, these results indicate that future work on the importance of network diversity for the geographically isolated may be fruitful.

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

38 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.

39 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.

40 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.