What (Exactly) is Novelty in Networks? Unpacking the Vision Advantages of Brokers, Bridges, and Weak Ties.

Sinan Aral
Massachusetts Institute of Technology
sinan@mit.edu

Paramveer S. Dhillon
University of Michigan, Ann Arbor
dhillonp@umich.edu

(Forthcoming at Management Science)

Abstract

The Strength of Weak Ties and Brokerage Theory both rely on the argument that weak bridging ties deliver novel information to create "vision advantages" for actors in brokerage positions. But our conceptualization of novelty is itself fundamentally underdeveloped. We, therefore, develop a theory of how three distinct types of novelty —diversity, non-redundancy, and uniqueness —combine with network structure to create vision advantages in social networks. We test this theory using panel data on an evolving corporate email network. Three main results emerge from our analysis. First, we confirm the Diversity-Bandwidth Tradeoff (DBT) at the heart of the vision advantage. As brokers' networks become more diverse, their channel bandwidth contracts, creating countervailing effects on access to novel information. Second, we uncover a mechanism driving the DBT, which helps explain differences in vision advantages across strong and weak ties. Strong, cohesive ties deliver greater information diversity and non-redundancy, while weak bridging ties contribute the most unique information (the information that is most different from what other contacts deliver). Third, we find network diversity (in contrast to network constraint) to be positively associated with longitudinal entropy, a measure of the accumulation of novel information over time. This suggests that weak bridging ties, which provide the most unique information through low bandwidth, structurally diverse channels, contribute the most to one's aggregation of novel information over time. Collectively, these results take a step towards resolving a long-standing debate in network theory about whether strong, cohesive networks or weak bridging networks contribute more to vision advantages. This work establishes firmly that it depends.

Keywords: *Networks, Information Flow, Knowledge Transfer, Content Analysis, Text Mining*

1 Introduction

For the last fifty years, researchers in disciplines as diverse as economics, sociology, management, marketing, and information systems have been developing an important line of social theory linking network structure to the distribution of information and knowledge in social groups. The Strength of Weak Ties (SoWT) and Brokerage Theory (BT) are two key theories at the heart of this research. They underpin tens of thousands of empirical investigations linking network structure to outcomes including wages, job placement, promotion, creativity, innovation, political success, social support, productivity, and performance (Aral et al., 2012, 2007; Baker, 1990; Bulkley and Van Alstyne, 2004; Burt, 2009, 2004; Granovetter, 1973; Hansen, 1999, 2002; Padgett and Ansell, 1993; Podolny, 2001; Reagans and Zuckerman, 2001; Uzzi and Spiro, 2005; Uzzi, 1997).

Both theories rely on the argument that networks low in cohesion and rich in structural holes provide diverse and novel information to actors enabling them to improve their innovation, productivity and career outcomes. These theories argue that weak bridging ties deliver novel information to actors in brokerage positions, providing them a "vision advantage" (Burt, 2005). But debates continue about whether strong cohesive ties or weak bridging ties are more valuable and there is a lack of empirical evidence validating the existence of vision advantages or how they work. While these two theories (and the various studies that link network structure to individual and group outcomes) rely on a notion of "access to novel information" as the key driver of performance, the literature is still vague about what novelty is and how it enables vision advantages. Granovetter (1973) notes that novelty is information that is "dissimilar," but does not elaborate on what that means. Without firm definitions on which to base empirical investigations, novelty and its role in the Weak Tie and Brokerage theories has, unfortunately, remained elusive.

To understand the vagueness of current conceptualizations of novelty, consider an employee working on a team that translates movies from one language to many others for the purpose of international distribution. Each project requires knowledge of current idioms and modern language use in each language, an understanding of workflow and thematic content, as well as other team members' expertise. Compare the situation in which this employee receives a message containing three bits of novel information from one of their team members regarding ways to improve their project workflow—three new ideas on how they could slightly alter their process to become more efficient. Then, consider the same employee receiving a message with just one bit of novel information, but about a new term in a modern language they are translating. Which of these two messages contains "more novel information?" A message with three bits of new information that the employee has never seen before, but that is closely related to information they are

already familiar with (information about efficient project workflows), or alternatively, a message with a single bit of novel information that is vastly different than any information they have ever seen before (a new term in a language spoken in a different country which they have never heard before)?

The sheer volume of new information contained in the first message is greater—it has three bits of novel information compared to the second message, which has only one. But, the relative distance of the information contained in the second message is, in some sense, greater or "more distinct or different" than what the employee has seen before. In this regard, the second message contains information that is more novel. We could imagine both types of novelty contributing to productivity. A vastly different idea can be thought of as being more akin to "out of the box" thinking. But, a greater volume of novelty with connections to what we already know could also contribute to productivity as team members could establish "common knowledge" around it and assimilate it more easily. Although this is a stylized example, it elucidates a weakness of current social theory in distinguishing between different facets of novelty. The theories are too imprecise to quantify the difference between novel information characterized by volume, distance from other information or cumulative heterogeneity, just to name three dimensions of what could be a large theoretical space.

This ambiguity in the conceptualization of novelty impedes our ability to explain important individual and group-level outcomes in network studies. Hence, the pivotal next step lies in theorizing about, observing, and measuring the novelty of the information content delivered by networked actors. As noted by Burt (2008) "[T]he 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," (Burt, 2005). Crystallizing the multiple facets of novelty by distilling the information flowing among network actors and quantifying the disparity in access to various types of novel information is essential for a holistic understanding of the mechanisms underlying the vision advantage.

The concept of novelty pervades the social sciences. The creativity literature theorizes that atypical thinking can lead to innovation. Atypicality or "novel combinations of prior work" (Uzzi et al., 2013), for example, is a related concept developed to explain the popularity of cultural products (Askin and Mauskapf, 2017; Goldberg et al., 2016). More recently, a "novelty hypothesis" was theorized to explain the spread of false news online (Vosoughi et al., 2018). In all of this literature, from weak ties, to brokerage, and from creativity to the spread of falsity, novelty and its measurement remains vague.

In this paper, we investigate the underlying dynamic mechanisms that enable vision advantages by unpacking, theorizing and more precisely measuring the concept of novelty. We first define and develop three new empirical measures of information novelty that provide specificity to our theoretical conceptualization: Information Diversity, Information Uniqueness, and Non-redundant Information. We then theorize how network structure affects access to these conceptually distinct dimensions of novelty by analyzing how much novel information of different types each actor in a broker's network should deliver to

the broker over time. Finally, we use vector-space and information-theoretic measures to operationalize novelty in an empirical organizational context and test our theory using the structure and content of an evolving corporate email network measured over twelve months. Temporal information structure also plays a critical role in our theoretical conceptualization of novelty and our empirical measures. We therefore analyze how network dynamics affect the amount of novel information brokers receive, allowing us to extend theory about the types of "information environments" in which brokers receive more or less novel information.

Three key findings emerge from our analysis. First, we confirm the Diversity-Bandwidth Tradeoff (DBT) at the heart of the vision advantage (Aral and Van Alstyne, 2011): As a broker's network becomes more diverse, the bandwidth of their communication channels contract, creating countervailing effects on access to novel information. Second, our analysis uncovers the mechanism driving the DBT and highlights differences between the vision advantages offered by strong cohesive ties and weak bridging ties. As our theory predicts, strong cohesive ties deliver greater information diversity and more non-redundant information. In contrast, weak bridging ties contribute greater uniqueness—information, which is the most different from what other contacts are delivering. Finally, weak bridging connections lead to greater aggregation of novel information over time compared to strong, high-bandwidth ties. Our conceptualization of novelty and the results of our empirical analysis together contribute to a dynamic ego- and dyad-level model of "vision advantages" that have for several decades been hypothesized to explain the strength of weak ties and brokerage.

Our work, therefore, makes three key contributions to these important lines of argument. First, we propose a theoretical explanation for how vision advantages work: weak bridging ties provide brokers with more unique information. In contrast, strong and cohesive connections provide more information diversity and more non-redundant information. Second, we show that the novel information provided by weak ties qualitatively and quantitatively differs from the novel information supplied by strong ties. Third, we find that network diversity (or, inversely, network constraint) is the dominant factor in the relationship between network structure and longitudinal entropy (a dynamic measure of non-redundant information). Taken together, these results suggest that weak bridging ties, which provide unique information through low bandwidth, structurally diverse channels, contribute the most to the aggregation of novel information over time. These contributions validate the information-based mechanisms theorized to drive the strength of weak ties and brokerage theory and serve to advance our understanding of the anatomy and dynamics of vision advantages.

The rest of the paper is organized as follows. In the next section, we position our work in the literature and develop our hypotheses regarding how network structure could affect access to the distinct dimensions of novelty. In Section 3, we theorize and develop the three dimensions of novelty corresponding to diversity, uniqueness, and non-redundancy of information. In Section 4, we describe our empirical setup, data, and model specifications. Section 5 shows the empirical results that unpack the effect of network structure on

access to different types of novelty. We conclude by summarizing our findings and their contributions in Section 6.

2 Theory

Human social networks tend to cluster due to triadic closure (Newman and Park, 2003). In his seminal work, Granovetter (1973) proposed the forbidden triad, i.e., a triad (a group of three networked actors) in which two strong ties are present, and one is not. He posited that forbidden triads are less likely to exist as two strong ties connected to a third connection are also themselves likely to be connected by a strong (or weak) link. The rationale for the lower likelihood of a forbidden triad is because the three networked actors are more likely to meet, more likely to have similar preferences and because their discord would create cognitive dissonance in the original strong tie friendship. The significant clustering that develops in human social networks as a result of triadic closure gives rise to small-world networks with short global path lengths (Watts and Strogatz, 1998) and heavy-tailed degree distributions (Barabási and Albert, 1999; Saramäki and Kaski, 2004).

Such structure—densely connected cliques connected by infrequent weak bridging ties—gives rise to opportunity. Brokers with structurally diverse networks, which lack cohesion and structural equivalence but are rich in structural holes, have privileged access to diverse and 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). These ties are the conduits through which ideas, influence, or information from a socially distant networked actor might reach an individual. Burt (2009) argues that "everything else constant, a large, diverse network is the best guarantee of having a contact present where useful information is aired." Since the information in local network neighborhoods tends to be redundant, structurally diverse contacts that reach across structural holes should intuitively provide channels through which novel information flows.

Novel information is thought to be valuable because of its local scarcity. Actors with scarce information in a given network neighborhood are better positioned to broker opportunities, make better decisions, and gain insights into problems that are intractable given local knowledge (Van Alstyne and Brynjolfsson, 2005; Burt, 2004; Hargadon and Sutton, 1997; Lazer and Friedman, 2007; Reagans and Zuckerman, 2001; Rodan and Galunic, 2004). In addition, access to novel information increases the breadth of individuals' absorptive capacity and strengthens their ability to communicate ideas across a wide range of topics to a broad audience. It also improves their persuasive ability to garner wider support from subject matter experts (Cohen and Levinthal, 1990; Reagans and Zuckerman, 2001; 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, 2009).

Several researchers have noted the difference in the amount and novelty of information flowing over strong and weak ties, e.g., Granovetter (1973) mentions that more total information passes over strong ties and that such relations convey conventional information, unlike weak links, which transfer risky information. Hansen (1999) also showed empirically that weak ties carry more unique information but less of it because it is challenging to transfer complex knowledge due to a lack of shared language and experience. More recently, Aral and Van Alstyne (2011) have explored the information benefits to structural diversity in networks and have uncovered a tradeoff between network diversity and channel bandwidth. Their main finding was that as an individual's ego network diversity increased, the bandwidth of their communication channels contracted, creating countervailing effects on access to novel information. The theoretical arguments underpinning this Diversity-Bandwidth Tradeoff (DBT) highlighted unexplored aspects of the SoWT and BT. In particular, if bridging ties are weak by nature and infrequent, they are also likely to deliver less novel information per unit time on a smaller number of topical dimensions. These results raised questions about exactly how novelty flows through network structure. Though Aral and Van Alstyne (2011) uncovered the countervailing effects of novel information in an organizational setup, they stopped short of theorizing about the multiple dimensions of novel information and exploring how the different facets of novelty can provide strategic advantages to networked actors.

Despite taking a first step towards measuring how vision advantages operate and how network structure and information flow are related, the DBT theory did not unpack the mechanism behind that tradeoff. The classic academic theories linking network diversity to novel information focused almost exclusively on the relative diversity of the information received across different alters in a network (Granovetter, 1973; Burt, 2009). They overlooked the diversity and volume of novel information flowing within each individual tie or channel over time. Aral and Van Alstyne (2011) argued that although dense, cohesive networks tend to deliver information that is redundant across channels (with each alter providing overlapping information), relationships in such networks are also typically stronger. These stronger relationships imply greater frequency of interaction, richer information flows, and thus access to more diversity and total novelty within each channel over time. This evidence raised new questions about whether vision advantages operated the way Granovetter and Burt had theorized. Unfortunately, Aral and Van Alstyne (2011) did not fully settle this debate since they employed ego-network level proxies for the information delivered by network ties. Specifically, they used averages of information across all the ties in a network to make statements about the general tendencies of various network ties to provide specific types of information. This lack of focus on information flow at the level of individual ties and information flow over time has paved the way for follow-up research on this topic.

We seek to reconcile the apparent tension in these theories with a simple unifying claim: strong embedded ties deliver greater information diversity and more non-redundant information, but they do so at the expense of information uniqueness, i.e., information that is

more topically distant. Thus, we argue that the information benefit provided by bridging ties is not in delivering greater information diversity or more non-redundant information. Instead, they deliver information that is unique —information that the ego is unlikely to get from anyone else in their contact network. The only way to test this unifying claim is to examine tie-level data that distinguishes the types of information delivered by strong embedded ties compared to weak bridging ties. In this regard, our empirical analysis extends the literature on vision advantages and brokerage theory by examining information diversity, total non-redundant information, and information uniqueness at the dyadic level.

We propose three metrics of information novelty that characterize three different aspects of novelty. Information diversity measures the topical variance of a set of information. Non-redundant information is a measure of the novel (or unrepeated) bits in a set of information. Information uniqueness, on the other hand, measures the distance between two information sets. We argue strong cohesive ties are likely to provide a broker with higher information diversity and more non-redundant information. This is because strong tie interactions occur through rich high-bandwidth channels that involve more detailed conversations, covering more topics, and addressing more complex, interdependent concepts over time. In contrast, weak bridging ties are likely to provide more unique information, which is more distant in information space, from the information ego receives from other contacts. This is because weak bridging ties communicate in social circles that are distant from the ego's other contacts. We therefore hypothesize:

- Hypothesis 1a: Strong cohesive ties deliver more information diversity and more non-redundant information than weak bridging ties.
- Hypothesis 1b: Weak bridging ties deliver more information uniqueness than strong cohesive ties.

Unpacking differences between diversity, non-redundancy, and uniqueness adds a theoretical subtlety to the information advantage argument, which could help reconcile conflicting evidence that has accumulated both for and against the brokerage theory over the years. One strand of research has found that diverse networks are associated with innovation (Burt, 2005), while another body of work has reached the opposite conclusion—that cohesion promotes innovation (Obstfeld, 2005; Uzzi and Spiro, 2005). We posit that one possible explanation for these contradictory results is that in situations where uniqueness matters, structural diversity is more valuable, while in cases where diversity or total non-redundancy matters, cohesion is more beneficial.

3 Conceptualizing Novelty

We theorize and measure three distinct aspects of novelty received from a particular contact: (i) the diversity of information received, which can be thought of as the dispersion of the topics discussed with that contact, (ii) the uniqueness of information received, i.e., the

distance between the topics discussed with a particular contact and the topics discussed with all of one's other contacts, and (iii) the incremental non-redundant information received from one's contacts. The distinctions among these three measures of novelty have clear implications for our theory, and we develop three different empirical measures that correspond to these concepts below.

For illustrative purposes, consider the communication network shown in Figure 1. Each node in the network is an individual, and the edge directionality denotes a message exchange over that dyad. We assume that each exchanged message is summarized by the set of topics that it discusses.[1] For example, the ego u receives three messages (summarized by the set of topics discussed) $m_{uv}^{\{1\}}$, $m_{uv}^{\{2\}}$, and $m_{uv}^{\{3\}}$ from the alter v.

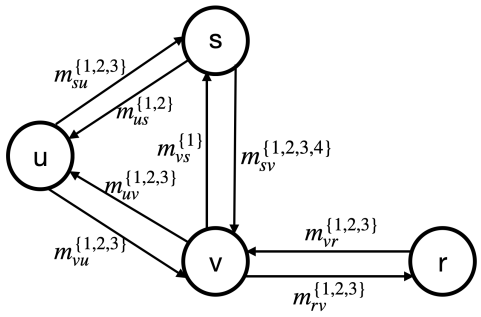

Figure 1: A toy communication network showing the messages exchanged between the different members of the network. *Note:* The superscript denotes the index of the message, since there can be multiple messages communicated over a tie.

3.1 Information Diversity

Information Diversity measures the degree to which a specific stream of information is focused or diverse based on the set of topics that it discusses. It quantifies the spread or variance of the topic distribution of the communication messages. Hence, a richer, more diverse set of communications will result in higher information diversity, while very specific communications focusing on a small set of topics will lower the information diversity. Information Diversity was first conceptualized and described as a measure of novelty by Aral and Van Alstyne (2011), who only considered ego-level information diversity.

We measure information diversity at both the ego as well as the dyad-level. Ego-level information diversity is calculated as shown in Equation 1. This measure computes the topical dispersion of all the messages that an ego u receives from all of their alters. Along similar lines, dyadic information diversity (Equation 2) captures the variety in topics discussed over a given tie $\langle s, v \rangle$.

[1]This is to simplify the discussion; our analysis is valid for other message summaries also.

Since the variance of a random variable is calculated as its deviation from the average value, both our information diversity measures compute the deviation from the average of the ego-level ($\overline{m}_{u\cdot}^{\{\cdot\}}$) or dyad-level ($\overline{m}_{sv}^{\{\cdot\}}$) topic distribution.

$$\text{InformationDiversity}_u = \textbf{Variance}[m_{uv}^{\{1,2,3\}}, m_{us}^{\{1,2\}}] \tag{1}$$

$$\text{InformationDiversity}_{sv} = \textbf{Variance}[m_{sv}^{\{1,2,3,4\}}] \tag{2}$$

3.2 Information Uniqueness

Information uniqueness is a dyadic level variable and measures the relative distance of topic distributions of the messages communicated between the ties. It quantifies how similar the information conveyed to an ego v by one contact r is to the information received from all their other contacts $r'(\forall r' \neq r)$. A greater distance between the information content that a particular contact provides relative to what other contacts offer indicates the uniqueness of the information conveyed over that specific dyad compared to the information the broker receives from everyone else.

Equation 3 describes the measurement of tie-level information uniqueness. Variables $\overline{m}_{vr}^{\{\cdot\}}$ and $\overline{m}_{v\setminus r}^{\{\cdot\}}$ represent the average set of topics communicated over the dyad $\langle v, r\rangle$ and all other dyads that v is a part of excluding r respectively. **Distance(·)** denotes any flexible distance metric that can be used to quantify the distance between two sets of messages in information space.

$$\text{InformationUniqueness}_{vr} = \textbf{Distance}[\overline{m}_{vr}^{\{\cdot\}}, \overline{m}_{v\setminus r}^{\{\cdot\}}] \tag{3}$$

3.3 Non-Redundant Information

Till now, we have proposed two measures to characterize the novelty of information; however, none of them capture the inherent "informativeness" or "surprise" contained in the communicated messages. If an ego learns something they already know, then the novel information they receive is minimal. Hence, a message mainly containing information already known to the ego will have very low entropy.

Non-redundant information captures precisely this intuition and is an information-theoretic measure of the information contained in the messages. As we will see later in the paper, we operationalize non-redundant information via Shannon Entropy (Cover, 1999) since entropy quantifies how much information or surprise there is in an event. Non-redundant Information is defined at both the ego and at the dyad level. We determine the amount of non-redundant information conveyed through messages along a tie $\langle v, r\rangle$ by controlling for all other information that the ego v receives through all other ties $r'(\forall r' \neq r)$ as shown

in Equation 4. To compute the non-redundant information for a given ego, we sum up the non-redundant information conveyed over all the ties (Equation 5).

$$\text{NonRedundantInformation}_{vr} = \textbf{Entropy}[\overline{m}_{vr}^{\{\cdot\}} | \overline{m}_{vu}^{\{\cdot\}}, \overline{m}_{vs}^{\{\cdot\}}] \tag{4}$$

$$\text{NonRedundantInformation}_{v} = \sum_{r} \textbf{Entropy}[\overline{m}_{vr}^{\{\cdot\}} | \overline{m}_{vu}^{\{\cdot\}}, \overline{m}_{vs}^{\{\cdot\}}] \tag{5}$$

To illustrate the differences between these three measures of information novelty, consider the following scenarios. A set of messages received by an ego will have high information diversity if they cover many different topics such as accounting, projects, IT, social gatherings, and news. In contrast, if most messages are about one or two topics only, then the corresponding information diversity will be low. Next, if a networked actor gets information about a certain topic, e.g., sports, from only one of their contacts, then the dyadic information uniqueness will be high over that dyad; otherwise, the tie-level uniqueness of information will be low. Finally, non-redundant information quantifies the amount of additional information a given source contributes (as measured in an information-theoretic sense). Hence, if only a single contact talks about a given topic (i.e., sports), the amount of non-redundant information is identical to the total volume of novel information from that source. However, suppose at least one other contact mentions that topic. In that case, the amount of non-redundant information provided by the first source is reduced by a measure proportional to the amount that others discuss that same topic.

3.4 Longitudinal Novelty

Up to this point, we have discussed the connection between network diversity and informational novelty in a static sense, considering information sent and received in a single (or pooled, cross-sectional) period. However, the value of the information that we receive depends, among other factors, on prior knowledge. Hence, a complete characterization of information novelty should also consider how these factors depend on prior knowledge or what one knows or has learned in the past. To incorporate prior knowledge, we use the information entropy framework that was just described but in a longitudinal setting. In particular, we ask: "*What is the amount of non-redundant information received during time period t given prior knowledge received in $t - n$ prior time periods?*"

In extending the framework we developed for the static setting to a dynamic one, a couple of nuances are worth noting. First, when studying information novelty across dyads in a communication network, we quantify the information received across a dyad relative to all other dyads. In a static setting, this conceptualization is symmetric across the reference dyad. However, there is an aggregation of knowledge in the dynamic setting, and hence the point of reference (the prior information of the receiving node in the dyad) is not interchangeable due to this accumulation of information over time.

Second, in a dynamic setting, we must account for memory loss and/or decay of the value of information over time. To comprehensively characterize information decay or memory loss, we consider two extreme cases of the degree to which information decays. First, we assume that information aggregates in time-periods $\{1, \ldots, t-1\}$, relative to time-period t, without any decay. We call this the memory (mem) model, as illustrated in Figure 2. Our second model which is memoryless (ml), on the other hand, considers only the information aggregated in time-period $t-1$ relative to time-period t; that is, it assumes decay of all the information before time-period $t-1$.

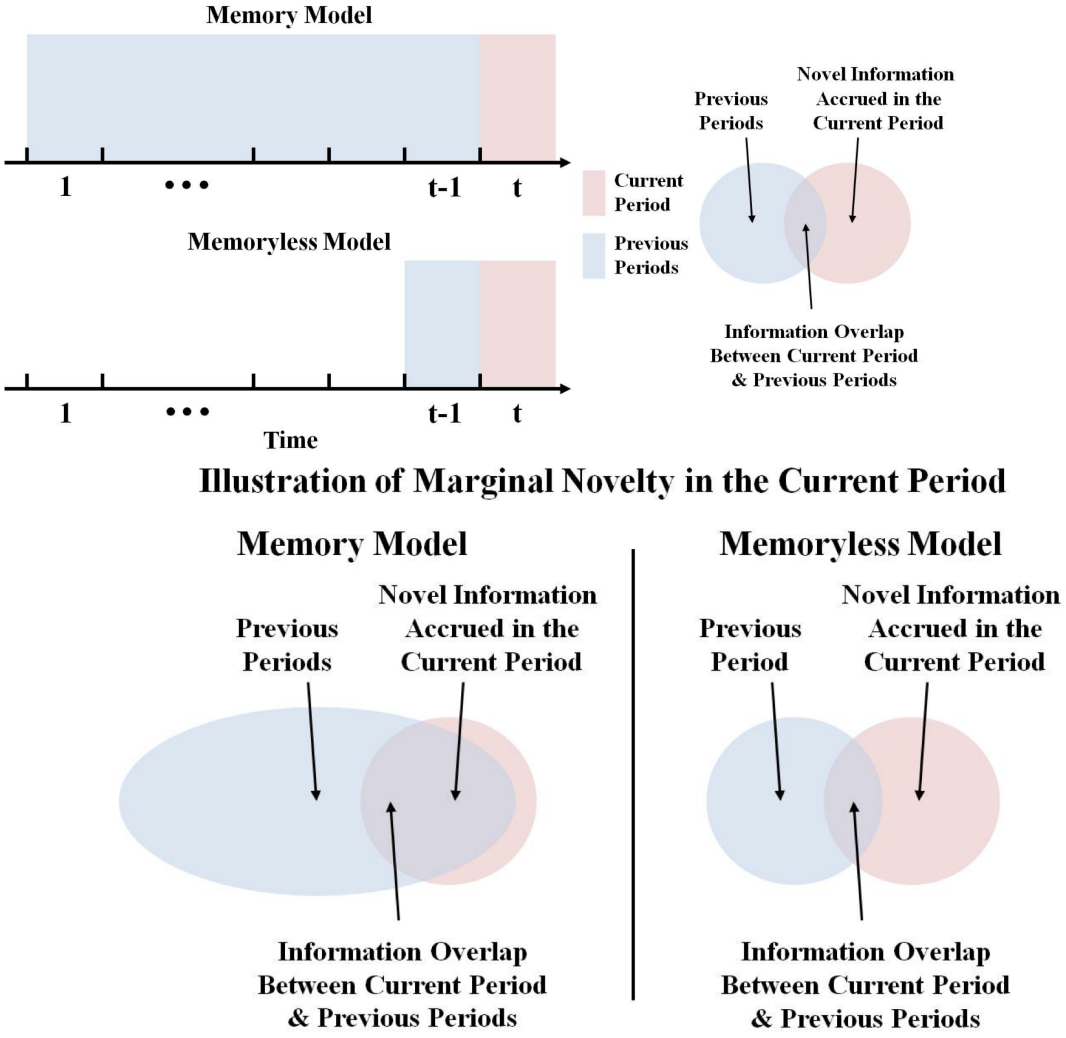

Figure 2: Illustration of the Memory & Memoryless Models of Longitudinal Entropy. The amount of novel (non-redundant) information accrued in the current panel is the set of information in the "Current Period" in the Venn diagram that does not overlap with the information in "Previous Periods." The memory model considers information accumulated in all previous periods, while the memoryless model only considers information accumulated in the last period (t-1) before the current period (t).

Both cases describe different aspects of information aggregation and capture different novelty dimensions relevant to the theory we develop and our modeling approach. In the first case, we retain long-term memory of information received over time. The amount of prior information that each ego is aware of grows over time and the new novel information obtained per unit time decreases systematically as a consequence, since new information is compared to a larger body of potentially redundant information already known to ego (See Figure 2)

If we assume that knowledge in a topic area is finite and does not increase over time, as one gains knowledge of that topic area over time, if they retain knowledge with no memory loss, then new information obtained on that topic in each period is likely to be less and less novel to them. Hence, the total amount of novel information that an ego receives about that topic will decrease over time as they learn all there is to know about that topic.

In the second case (memoryless), we only consider information aggregated in the prior period as our reference point. In essence, this model assumes a complete decay of information from one time period to the next. The corresponding hypothetical scenario is of a memoryless Markov process, i.e., the amount of additional non-redundant information is only a function of the new information received and information known in the prior period, but not information obtained in periods 1 to $t-2$. We operationalize both these models of information accumulation and decay by extending the entropy-based measures of information redundancy that were discussed earlier to a longitudinal setting.

Understanding the relationship between network structure and novel information a) at the level of individual dyadic ties, b) aggregated over the entire ego network, and c) longitudinally over time allows us to paint a nuanced picture of the mechanics of the broker's vision advantage and reveals new insights on how the information mechanism underpinning the strength of weak ties and brokerage operates.

4 Empirical Setting

We explore the anatomy and dynamics of vision advantages by analyzing the content and structure of an evolving corporate email network over twelve months. Working with email data allows us to measure network structure and topical discussion content accurately and further alleviates the bias involved in respondent self-reports. Several previous studies have validated the efficacy of using email data in characterizing and analyzing social networks (Wu et al., 2004; Kossinets and Watts, 2006; Aral and Van Alstyne, 2011).

We study the email network of a medium-sized, global digital media firm with offices across North America, Europe, and Asia. It has more than 1000 employees worldwide and several thousand freelance workers. This firm delivers language and localization services such as translation, dubbing, and subtitling for film, digital gaming, and web content for clients worldwide. The services provided by this firm's employees require constant information

seeking and communication to solve highly localized problems. For example, translating a movie from English into thirty other languages requires translators to inquire about current local language use and modern-day idioms from regional experts in the firm. In interviews and during participant observation, employees frequently reported and were observed seeking information from people in disparate parts of the firm's communication network to solve these highly idiosyncratic problems. Our interviews revealed that timely access to such novel information from disparate parts of the network were important drivers of project completion rates and error rates.

Before starting the quantitative data collection, we collected ten weeks of participant observation data over six months. This initial data collection included data from interviews of the senior executive team and key informants from sales, technology, and operations. We also conducted interviews with employees in each of the major language teams that produce the localization work. These areas represent a comprehensive set of all the employees in the firm. In addition to these interviews, we observed the employees from each of these divisions performing their work and took detailed notes of our observations. This initial data collection helped us understand the setting, the work that was being done, the role of novel information in the work, and the nature of the social network dynamics at play in the firm's communications.

Following the qualitative data collection, we collected comprehensive data on the content and structure of the firm's evolving corporate email network. Figure 3 displays the largest connected component of the email network using data aggregated over the entire observation period. It is easy to see the distinct communication clusters in the firm's email network in the figure. The distinct communities that have developed within the firm's communication structure fulfill different roles in the company's workflow. They accumulate distinct pools of knowledge and information and create a setting in which employees must reach outside of their local networks, through weak bridging ties, to gain access to novel information they need to complete their work. The variegated nature of the community communications in the firm provides a perfect opportunity to study the role of structural diversity in giving privileged access to novel information and, therefore, to investigate the dynamics of vision advantages.

4.1 Data

Our data consists of three sources: (i) human resources information such as employees' gender and date of hire, (ii) all internal employee email communications, (iii) survey data on information-seeking behaviors. Our analysis aims to test the theoretical mechanisms that establish and enable vision advantages from structural holes, SoWT, and DBT. To do so, we unpack and operationalize the concept of novel information to reflect the theoretical distinctions between information diversity, total non-redundant information, and information uniqueness. We then measure information diversity, non-redundant information, and information uniqueness in the content of the emails exchanged between employees and statistically relate variance in these measures to the dynamic structural characteristics of

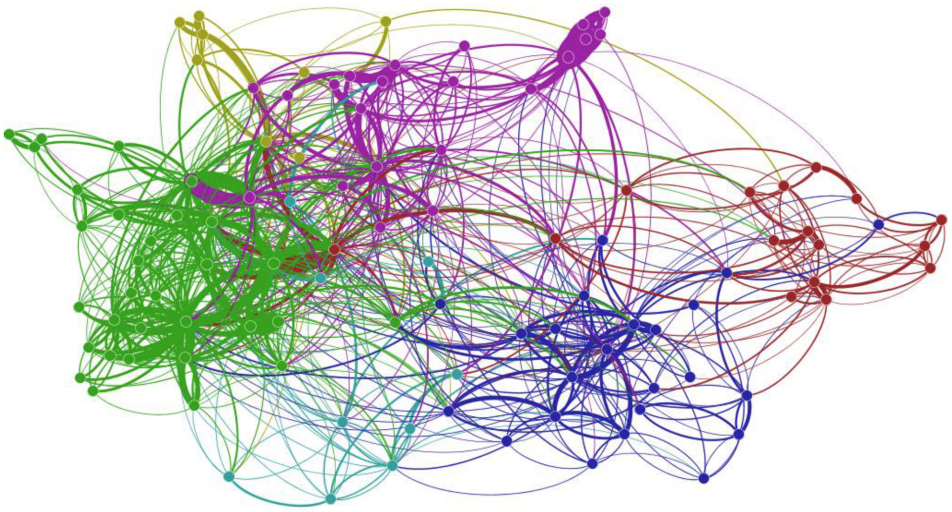

Figure 3: Communities in Network Structure of the email network of the media firm. Note: The communities (color coded) were discovered using the algorithm by Blondel et al. (2008).

the evolving corporate email network over time (specifically, Burt's constraint measure, Aral and Van Alstyne (2011)'s bandwidth measure, and relevant control variables). The result provides empirical evidence of how dynamic network structure is related to the flow of novel information in a firm; specifically, the diversity of the received information, the ebb and flow of novel information exchanged, and the variation in the uniqueness of the communicated information. The results provide intriguing evidence of how vision advantages, the strength of weak ties, and the diversity bandwidth tradeoff all operate in practice.

Overall we collected two million emails[2] exchanged among 232 employees over twelve months during 2010.[3] Next, we preprocessed the email content using a standard text processing pipeline (Loper and Bird, 2002). We tokenized the text, removed the stop words such as (*"a," "an," "the," "and"*), and stemmed the roots of words such as *multitasking* and *multi-task* to make them *multitask*. Finally, the email content was anonymized using a hashing algorithm. This hashing was done primarily due to privacy concerns to protect potentially sensitive information shared over emails. It is not ideal to use hashed email data, but there is a clear tradeoff. We can either use a self-selected subsample of unhashed emails exchanged between employees (to which employees have selectively provided access) or use the complete set of hashed emails. We chose the latter option (as is common in this literature) because the internal and external validity concerns associated with self-selected data were much more serious, in our assessment, than not being able to read the actual text of email communications. We used the same hashing algorithm used by Aral

[2]All email aliases were associated with a single user.

[3]Based on the survey regarding employees' information-seeking behavior, 88% of the respondents mentioned using email as their primary form of communication for work-related information.

and Van Alstyne (2011) due to its remarkable strength in permitting text analyses while preserving privacy (for more information about these and other properties of the hashing algorithm see Reynolds et al. (2009)).

Next, we derive our novelty measures of information diversity, non-redundancy, and information uniqueness from the email texts. Based on a popular paradigm in text mining and natural language processing, we represent each email by the topical distribution of the concepts that it discusses (Manning and Schütze, 1999). In the parlance of natural language processing, our "corpus" consists of many "documents" (emails). Each email is further comprised of a sequence of words $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$. Our goal is to summarize each email's content as a probability distribution over a set of k (typically 50-100) latent topics (or concepts). Then, it is straightforward to compute the various novelty measures by simply computing the degree of similarity or overlap between these topics.

Although this text representation task can be accomplished using several state-of-the-art natural language processing methods, we chose Latent Dirichlet Allocation (LDA) (Blei et al., 2003) due to its simplicity and popularity. LDA is a probabilistic model of text documents, that models each document as a stochastic mixture of k (typically 50-100) latent topics. It takes as input a collection of documents and outputs the topical distribution for each one of them. Note that in our case, the content is anonymized; hence, the semantic meaning of the topics is unclear. However, this is not a problem as LDA only models the co-occurrence patterns of words in documents and not the words' actual semantic identity or meaning.

We use LDA to estimate topics from all our hashed email data. We chose the total number of topics to be 50, a typical number used in many text modeling studies. We also varied the number of topics to 25 and 100 and re-estimated the LDA model. The corresponding results are strikingly consistent and are shown in the Appendix.

Though LDA is a popular choice for modeling text data, recently, there has been a surge in the use of vector space models for text representation (Mikolov et al., 2013; Dhillon et al., 2011, 2012, 2015; Pennington et al., 2014). These models "embed" each document into a real-valued k dimensional (typically 50-300) vector. These document embeddings can then be used similarly to the topical distribution estimated by the LDA. Like LDA, the vector space models also harness the co-occurrence patterns of text in estimating document embeddings. One such popular method is doc2vec (Le and Mikolov, 2014). We replicate all our analyses with doc2vec instead of LDA for the sake of completeness. The results are shown in the Appendix, and they are broadly similar to those based on LDA.

4.2 Definition of Variables

The variables that we use for our modeling consist of both the employee human-resource information and the hashed email communications between the employees. The focal

point of our analyses is both the ego-level (employee i) and dyad-level (employees i and j) information shared during time-period t.

Let $\text{NumberMessages}_{it}$ denote the number of messages received by an ego i during the time-period t such that $\sum_i \sum_t \text{NumberMessages}_{it} = \text{N}$, where N is the total number of emails exchanged among all the employees during our observation period. Further, let NetworkSize_{it} represent the number of contacts from which an ego i received at least one message during time-period t. Next, let Γ_{itm} be the k dimensional topic distribution vector for the m^{th} email message received by ego i in the time period t. Our operationalization of "topics" here is highly general and denotes a k-dimensional clustering of the underlying hashed email text. The latent clusters or topics can be estimated using any Natural Language Processing (NLP) method, e.g., Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Probabilistic LSA (p-LSA), or doc2vec (Le and Mikolov, 2014). For the reasons outlined in the previous section, we use LDA (Blei et al., 2003) to estimate the topics in our empirical analyses. Based on these core definitions, next we describe the operationalization of various variables used in our analyses.

4.2.1 Channel Bandwidth

$\text{ChannelBandwidth}_{itj}$, i.e., the channel bandwidth per tie $\langle i, j \rangle$, is simply the number of messages received by ego i from alter j in the time-period t. Further, an ego i's average channel bandwidth can be defined as the total number of messages they receive over all the incoming ties during that time-period. In other words:

$$\text{ChannelBandwidth}_{it} = \frac{\text{NumberMessages}_{it}}{\text{NetworkSize}_{it}} \tag{6}$$

4.2.2 Network Constraint

We use Burt's network constraint metric to measure brokerage. Specifically, we break this measure down into its individual components to apply it to the constraint of each broker's ego network and measure ego's investment in specific ties. Derived by using the bidirectional traffic of messages between any two brokers, we denote the proportion of time and effort invested by ego i in a specific alter j as p_{ij}. We represent this as his direct investment. Further, we consider secondary or redundant investments via mutual relationships (indexed by j) in the communications network. This measure of redundant investment is defined as:

$$\text{RedundantInvestment}_{ij} = \sum_{j' \neq i \neq j} p_{ij'} p_{j'j} \tag{7}$$

We expect both direct and redundant investment to influence the diversity and amount of non-redundant information the ego receives from a specific contact in their network. To

quantify the amount of network constraint an ego experiences in their personal network, we sum over all investments (direct and redundant) for all of their peers:

$$\begin{aligned}\text{NetworkConstraint}_{it} &= \sum_{j=1}^{\text{NetworkSize}_{it}} (\text{DirectInvestment}_{itj} + \text{RedundantInvestment}_{itj})^2 \\ &= \sum_{j=1}^{\text{NetworkSize}_{it}} (p_{itj} + \sum_{j' \neq i \neq j} p_{itj'} p_{j'tj})^2 \end{aligned} \tag{8}$$

The decomposition of network constraint into its direct and redundant investment components is especially useful for dyad-level analyses since these measures are defined over individual ties and allow us to isolate dyadic information flows.

4.2.3 Information Diversity

Information diversity measures the degree to which a specific stream of information is focused or diverse. We quantify it by measuring the dissimilarity of topic distributions of the received messages. We used the most common measure of document similarity, cosine similarity, to operationalize the dissimilarity in topic distributions[4]. The Appendix shows robustness of our results to using Hellinger Distance and KL-divergence as alternate distance measures. Information diversity is measured at both the ego network level and the dyad level as described below:

1. **Ego-level Information Diversity:** The information diversity of all the messages ($\text{NumberMessages}_{it}$) that an ego i receives from all their peers in time-period t is the variance of the topic distribution of the messages received by that ego in that time period.

   Let $\overline{\Gamma}_{it}$ be the average topic distribution vector of the messages received by the ego i during time-period t. In other words,

   $$\overline{\Gamma}_{it} = \frac{1}{\text{NumberMessages}_{it}} \sum_{m=1}^{\text{NumberMessages}_{it}} \Gamma_{itm} \tag{9}$$

   where, Γ_{itm} is the topic distribution of the m^{th} message received by ego i in time period t.

[4]Note that cosine similarity is not a "distance metric" in the strict mathematical sense https://en.wikipedia.org/wiki/Cosine_similarity. However, it is still employed for similarity calculations in the vector-space modeling literature (Turney and Pantel, 2010).

Then, the information diversity of all the messages received by ego i in time-period t can be computed as:

$$\text{InformationDiversity}_{it} = \frac{1}{\text{NumberMessages}_{it}} \sum_{m=1}^{\text{NumberMessages}_{it}} \left[1 - \cos(\Gamma_{itm}, \overline{\Gamma}_{it})\right]^2 \tag{10}$$

It is worth noting that since information diversity is a variance-based measure, it involves a squared deviation from the mean[5]. Based on this definition of information diversity, it is easy to see that, a richer, more diverse set of communications will result in a higher information diversity, while very specific communications, focusing on a small set of topics, will result in lower information diversity.

2. **Dyad-level Information Diversity:** We also measured information diversity within a specific dyad. Dyadic information diversity is defined as the information diversity of all the messages between a specific sending alter (say) j and ego i in time-period t. The definition of the dyad-level information diversity is the same as described above, with the exception that now we only sum over the $\text{NumberMessages}_{itj}$ messages exchanged between i and j in time-period t.

$$\text{InformationDiversity}_{itj} = \frac{1}{\text{NumberMessages}_{itj}} \sum_{m=1}^{\text{NumberMessages}_{itj}} \left[1 - \cos(\Gamma_{itjm}, \overline{\Gamma}_{itj})\right]^2 \tag{11}$$

4.2.4 Information Uniqueness

Information uniqueness is a dyadic level variable and measures the distance of topic distributions between the ties in a given time-period. It quantifies how similar the information conveyed to an ego i by one contact j is to the information conveyed to that ego by all their other contacts $j'(\forall\ j' \neq j)$, from whom the ego received at least one message in time-period t. Let $\overline{\Gamma}_{itj}$ be the average topic distribution of the messages received by ego i from alter j in time-period and further let NetworkSize_{it} be the number of contacts from whom the ego i received at least one message in time-period t. Operationalizing distance between topic distributions via cosine similarity as earlier[6], the information uniqueness variable is defined as:

$$\text{InformationUniqueness}_{itj} = \frac{1}{\text{NetworkSize}_{it} - 1} \sum_{j'=1}^{\text{NetworkSize}_{it}} \left[1 - \cos(\overline{\Gamma}_{itj}, \overline{\Gamma}_{itj'})\right] \tag{12}$$

[5]This measure was first introduced by Aral and Van Alstyne (2011) to model the topical dispersion of information flow in a network.

[6]Note that cosine similarity is not a "distance metric" in the strict mathematical sense https://en.wikipedia.org/wiki/Cosine_similarity. However, it is still employed for similarity calculations in the vector-space modeling literature (Turney and Pantel, 2010).

A greater distance between the information content a particular contact provides and what all other contacts provide indicates that the information conveyed over that specific dyad is unique compared to the information the broker receives from everyone else.

4.2.5 Non-Redundant Information

Non-Redundant Information quantifies the amount of novel information contained in a message and is defined both at the ego and dyadic level. It is simply a measure of knowledge that an ego (or a tie) receives that they did not already know. Hence, we need to operationalize it using a measure that quantifies the potential amount of information conveyed by a message given its topic distribution vector. An ideal measure for this purpose is the information entropy $\mathbb{H}(\mathbf{\Gamma}) = -\sum_{a=1}^{k} \Gamma_a \ln \Gamma_a$.[7]

Information entropy measures how many "bits" of information or "surprise" there is in an event (Cover, 1999). If an ego learns something they already know, then the novel information they receive is minimal. A message containing mostly information already known to the ego will have very low entropy. If we want to determine the amount of non-redundant information conveyed through messages along a tie $\langle i, j \rangle$, we want to account for all other pieces of information that i receives through other ties $j'(\forall j' \neq j)$ and control for redundancy in the information provided by those other ties.

1. **Dyad-level Non Redundant Information:** We operationalize the dyad-level non-redundant information via Conditional Entropy. It measures the average amount of non-redundant information an ego i receives from a specific alter j, given the topic distribution vectors of all other alters communicating with the ego. More precisely, it measures the amount of marginal information provided to the ego by a specific alter relative to the combined information provided by all other alters. Overlaps in information between the dyad $\langle i, j \rangle$, and any other dyad $\langle i, j' \rangle$ are discounted, and hence conditional entropy measures only the fraction of genuinely non-redundant information provided by alter j. Assuming, $\overline{\Gamma}_{itj}$ and $\overline{\Gamma}_{itj'}$ as the average topic distributions for the messages exchanged between the ties $\langle i, j \rangle$, and $\langle i, j' \rangle$ respectively, we can define the conditional entropy (or dyad-level non-redundant information) of tie $\langle i, j \rangle$ as:

$$\text{ConditionalEntropy}_{itj} = \mathbb{H}(\overline{\Gamma}_{itj} | \overline{\Gamma}_{it1}, \overline{\Gamma}_{it2}, \ldots, \overline{\Gamma}_{it(\text{NetworkSize}_{it}-1)}) \tag{13}$$

   where $\text{NetworkSize}_{it} - 1$ is the total number of ties (excluding j) from which i received at least one message in time-period t.

2. **Ego-level Non Redundant Information:** We measure the amount of non-redundant information that an ego i receives from all his contacts by summing the conditional entropies of the information that an ego i receives from all their

[7] Recall that the k-dimensional topic vector $\mathbf{\Gamma}(= [\Gamma_1, \Gamma_2, \ldots, \Gamma_k])$ provides a probability distribution over all the topics for a given message.

contacts. It turns out that this summation of all the conditional entropies is equal to the Joint Entropy (Cover, 1999). Hence, Joint Entropy measures the ego-level non-redundant information and is calculated as below:

$$\text{JointEntropy}_{it} = \mathbb{H}(\overline{\Gamma}_{it1}, \overline{\Gamma}_{it2}, \ldots, \overline{\Gamma}_{it\text{NetworkSize}_{it}}) \tag{14}$$

To summarize, conditional entropy and joint entropy operationalize the amount of non-redundant information provided to the ego by a given contact and by all of an ego's contacts, respectively. We use both these measures of non-redundant information in our regression models and denote them simply as non-redundant information.

4.2.6 Longitudinal Entropy

All the measures of information novelty we have operationalized to this point are static and do not account for the accumulation and decay of information and novelty over time. So, we propose two measures of information accumulation and decay. First, we assume that information aggregates in time-periods $\{1, \ldots, t-1\}$, relative to time-period t, without any decay. We call this the memory (mem) model. Our second model considers only the information aggregated in time-period $t-1$ relative to time-period t; that is, it assumes decay of all the information before time-period $t-1$. This second model is called the memoryless model (ml). We characterize this dynamic accumulation (or correspondingly decay) of information via longitudinal entropy, which is simply the difference in Joint Entropy accrued over time. The resulting definitions of longitudinal entropy in both cases are:

$$\text{LongitudinalEntropy}^{mem}_{it} = \text{JointEntropy}_{i(1:t)} - \text{JointEntropy}_{i(1:t-1)} \tag{15}$$

$$\text{LongitudinalEntropy}^{ml}_{it} = \text{JointEntropy}_{it} - \text{JointEntropy}_{i(t-1)} \tag{16}$$

where $\text{JointEntropy}_{i(1:t)}$ is the joint entropy of all the messages that an ego i received from all their contacts in time-periods 1 through t, that is, $\mathbb{H}(\overline{\Gamma}_{i(1:t)1}, \overline{\Gamma}_{i(1:t)2}, \ldots, \overline{\Gamma}_{i(1:t)\text{NetworkSize}_{it}})$. Note that $\overline{\Gamma}_{i(1:t)S_{it}}$ denotes topic averaging from time-periods 1 through t. Similarly, $\text{JointEntropy}_{i(1:t-1)}$ is the joint entropy of all the messages received by an ego, where the topics were averaged from time-periods 1 through $t-1$. In contrast, JointEntropy_{it} and $\text{JointEntropy}_{i(t-1)}$ represent joint entropy of messages received just in time-periods t and $t-1$ respectively, that is, the topics are averaged within those time-periods only.

4.2.7 Control Variables

In addition to the above variables, we also employ two important control variables. Specifically, we control for the gender difference between the ego and each of their contacts who

sent them an email message. The gender difference variable GenderDiff_i is operationalized as the average of the gender difference between the ego and each of their contacts from whom they received an email message. Along similar lines, we operationalize the HireDateDiff_i variable, which controls for the difference in the hiring date of ego and the alters. Both these variables control for the idiosyncrasies of the email content that could be attributed either to the difference in tenure at the organization (HireDateDiff_i) or to the gender differences in the role assignment on the project teams (GenderDiff_i).

The descriptive statistics of the key variables in the dataset are shown in Table 1. Table 2 shows the correlations between those variables.

Variable	Mean (μ)	SD (σ)	Minimum	Maximum
Gender Difference	0.50	0.20	0	1
Hire Date Difference	-0.11	3.94	-15.07	12.48
Total Incoming Emails	243.1	305.1	1	2566
Network Size	23.08	19.18	1	96
Channel Bandwidth	9.58	8.21	1	99
Network Constraint	0.34	0.27	0.07	1.30
Information Diversity	0.81	0.21	0	0.98
Non-Redundant Information (Joint Entropy)	49.1	44.19	0.25	236.4

Table 1: Descriptive Statistics (Panel of 232 employees (N=2300) who received atleast one email over a 12 month period from Jan-Dec. 2010). Hire date and gender differences are the differences (to-from) averaged over all the contacts who sent atleast one email during that panel period.

4.3 Model Specification

We use monthly panels (i.e., the time-period t is a month) to estimate the relationship between network structure and information novelty.[8] To understand the exact underlying principles, we investigate the relationship between ego network structure and the information that an ego receives. Further, we also zoom-in on the flow of information along individual dyadic ties.

In all analyses, we control for the effects of differences in demographic factors between senders and receivers in the email network (the difference in hire date and the difference in gender between senders and receivers). We compute demographic control variables at the level of individual ties and aggregate them by calculating their average values for the ego networks of information brokers across all incoming ties in a time-period t.

[8]This is the right level of granularity for our analysis as it is neither too fine-grained to run into the issue of sparsity, nor too coarse-grained to have too little temporal variation over our 12 month panel data. Further, the monthly granularity of analysis aligns with the firm's internal progress update deadlines for projects.

Variable	1	2	3	4	5	6	7
1. Gender Difference	-	-	-	-	-	-	-
2. Hire Date Difference	-0.05	-	-	-	-	-	-
3. Total Incoming Emails	-0.01	0.05	-	-	-	-	-
4. Network Size	-0.03	0.05	0.81	-	-	-	-
5. Channel Bandwidth	0.02	0.02	0.50	0.14	-	-	-
6. Network Constraint	0.07	-0.05	-0.39	-0.65	0.09	-	-
7. Information Diversity	-0.06	0.04	0.39	0.55	0.17	-0.77	-
8. Non-Redundant Information (Joint Entropy)	-0.02	0.04	0.84	0.99	0.16	-0.62	0.53

Table 2: Pairwise Correlations between different variables of the panel of 232 employees.

4.3.1 Ego-level Analysis

We first replicate the Diversity-Bandwidth tradeoff (DBT) analysis of Aral and Van Alstyne (2011) on our dataset with the specifications given in Equations 17 and 18.

$$\begin{aligned}\text{ChannelBandwidth}_{it} = \gamma_i + \delta_t + \text{NetworkConstraint}_{it} + \text{NetworkSize}_{it} \quad + \\ \text{NetworkSize}^2_{it} + \text{GenderDiff}_i + \text{HireDateDiff}_i + \epsilon_{it} \end{aligned} \tag{17}$$

$$\begin{aligned}\text{NetworkConstraint}_{it} = \gamma_i + \delta_t + \text{ChannelBandwidth}_{it} + \text{NetworkSize}_{it} \quad + \\ \text{NetworkSize}^2_{it} + \text{GenderDiff}_i + \text{HireDateDiff}_i + \epsilon_{it} \end{aligned} \tag{18}$$

Next, we examine the relationship between network structure and information diversity and the relationship between network structure and total non-redundant information, with non-redundant information operationalized as joint entropy. The specifications are given in Equations 19 and 20.

$$\begin{aligned}\text{InformationDiversity}_{it} \quad = \quad & \gamma_i + \delta_t + \text{ChannelBandwidth}_{it} + \text{NetworkConstraint}_{it} + \\ & \text{NetworkSize}_{it} + \text{NetworkSize}^2_{it} + \\ & \text{GenderDiff}_i + \text{HireDateDiff}_i + \epsilon_{it} \end{aligned} \tag{19}$$

$$\begin{aligned}\text{NonRedundantInformation}_{it} \quad = \quad & \gamma_i + \delta_t + \text{ChannelBandwidth}_{it} + \text{NetworkConstraint}_{it} + \\ & \text{NetworkSize}_{it} + \text{NetworkSize}^2_{it} + \\ & \text{GenderDiff}_i + \text{HireDateDiff}_i + \epsilon_{it} \end{aligned} \tag{20}$$

In the specifications above, i indexes ego/individual and t indexes time-period (month). γ_i denotes ego-level random effects, δ_t represents time controls, and ϵ_{it} is the idiosyncratic error term. We also flexibly control for network size by incorporating the linear (NetworkSize_{it}) and quadratic terms ($\text{NetworkSize}_{it}^2$), controlling for average gender differences and hire date differences between ego and the alters.

4.3.2 Dyad-level Analysis

Next, we turn to dyad-level analyses, which are specified in Equations 21, 22, and 23. These dyadic models examine the relationship between tie characteristics (e.g., constraint, tie strength, bandwidth, and network size) and the various dyadic novelty measures, information diversity, non-redundant information, and information uniqueness. We decompose the network constraint measures into its tie-level components—direct investment and redundant investment to isolate the informational advantages provided by different types of novelty. We again allow for non-linear dependence of network size on the various novelty characterizations.

$$\begin{aligned}\text{InformationDiversity}_{itj} &= \gamma_{ij} + \delta_t + \text{ChannelBandwidth}_{itj} + \text{NetworkConstraint}_{itj} + \\ &\quad \text{NetworkSize}_{it} + \text{NetworkSize}_{it}^2 + \\ &\quad \text{DirectInvestment}_{itj} + \text{RedundantInvestment}_{itj} + \\ &\quad \text{GenderDiff}_{ij} + \text{HireDateDiff}_{ij} + \epsilon_{itj} \end{aligned} \tag{21}$$

$$\begin{aligned}\text{NonRedundantInformation}_{itj} &= \gamma_{ij} + \delta_t + \text{ChannelBandwidth}_{itj} + \text{NetworkConstraint}_{itj} + \\ &\quad \text{NetworkSize}_{it} + \text{NetworkSize}_{it}^2 + \\ &\quad \text{DirectInvestment}_{itj} + \text{RedundantInvestment}_{itj} + \\ &\quad \text{GenderDiff}_{ij} + \text{HireDateDiff}_{ij} + \epsilon_{itj} \end{aligned} \tag{22}$$

$$\begin{aligned}\text{InformationUniqueness}_{itj} &= \gamma_{ij} + \delta_t + \text{ChannelBandwidth}_{itj} + \text{NetworkConstraint}_{itj} + \\ &\quad \text{NetworkSize}_{it} + \text{NetworkSize}_{it}^2 + \\ &\quad \text{DirectInvestment}_{itj} + \text{RedundantInvestment}_{itj} + \\ &\quad \text{GenderDiff}_{ij} + \text{HireDateDiff}_{ij} + \epsilon_{itj} \end{aligned} \tag{23}$$

In the specifications above, γ_{ij} represents dyad-specific random effects, δ_t is the monthly time control as earlier, and ϵ_{itj} denotes the idiosyncratic error term. We again control for gender and hire date differences, though here they are measured at the dyad-level rather than the ego-level.

4.3.3 Longitudinal Analysis

Finally, we perform longitudinal analysis of the temporal differences of variables. Here we want to assess the determinants of novel information aggregation over time. In particular, we assess whether channel bandwidth or network constraint (diversity) leads to more novel information aggregation over time and ask: Does this association vary based on whether the accumulation process is memoryless or not? The specification is given in Equation 24.

$$\text{LongitudinalEntropy}_{it} = \Delta[\text{ChannelBandwidth}_{it}] + \Delta[\text{NetworkConstraint}_{it}] + \Delta\epsilon_{it} \quad (24)$$

where, $\Delta[\text{ChannelBandwidth}_{it}] = \text{ChannelBandwidth}_{it} - \text{ChannelBandwidth}_{it-1}$ and $\Delta[\text{NetworkConstraint}_{it}] = \text{NetworkConstraint}_{it} - \text{NetworkConstraint}_{it-1}$

Here, we measure the longitudinal entropy in memory (mem) and memoryless (ml) models. The operator Δ denotes the change in the corresponding variable. As earlier, the subscript i denotes an ego, and t represents the time-period (month). The Appendix shows the robustness of our results to using a weekly time-window to measure the longitudinal novelty measures.

5 Results

5.1 The Diversity-Bandwidth Tradeoff (DBT)

By investigating the relationship between network constraint and bandwidth and their joint association with the novelty of incoming email, we can describe how changes in the communication network structure are associated with changes in the type of information received. If the DBT regulates the receipt of novel information, then we should observe two phenomena in our data. First, as employees' networks become more diverse (less constrained), the bandwidth of their communication channels should contract. Second, we should observe increases in the receipt of novel information both as networks become more structurally diverse and as channel bandwidth increases. If these conditions hold, then a tradeoff between network diversity and channel bandwidth exists, and it creates countervailing effects on the receipt of novel information.

As can be seen in Table 3, we find strong evidence confirming a tradeoff between the diversity of information received and the associated channel bandwidth. As employees communicated with densely connected contacts, the overall bandwidth of their communication channels to those contacts widened quite rapidly. After controlling for several relevant variables, we find that a one standard deviation increase in network constraint, i.e., reduced structural diversity, was associated on average with a 0.15 (Model 2, Table 3) standard deviation increase in channel bandwidth. On the other hand, a one standard deviation increase in channel bandwidth is associated with approximately a 0.09 (Model

	Model 1	Model 2
GenderDiff_i	0.020	-0.019
	(0.013)	(0.017)
HireDateDiff_i	0.014	-0.068
	(0.025)	(0.036)
NetworkSize_{it}	-0.967***	0.272***
	(0.027)	(0.046)
$\text{NetworkSize}^2_{it}$	0.277***	-0.024
	(0.012)	(0.018)
$\text{NetworkConstraint}_{it}$	-	0.146***
		(0.028)
$\text{ChannelBandwidth}_{it}$	0.087***	-
	(0.015)	
Constant	-0.276***	-0.187**
	(0.045)	(0.068)
Temporal Controls	Month	Month
R^2	0.38	0.07
Chis-quared statistic (df)	1394.9*** (16)	169.7*** (16)

Table 3: (Ego-level) The Network Diversity-Bandwidth Tradeoff (N=2300) using Random Effects regression. **(Model 1) Dependent Variable**= $\text{NetworkConstraint}_{it}$, **(Model 2) Dependent Variable**= $\text{ChannelBandwidth}_{it}$. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

1, Table 3) standard deviation increase in network constraint, again controlling for other factors. Hence, as networks become less diverse, the thickness of their communication channels increases.

Examining the effects of the DBT on information novelty, we find a substantial effect on the diversity of information received and the total amount of non-redundant information received. As networks become more structurally diverse, brokers experience an increase in the dispersion of the information and the non-redundant information they receive. Specifically, a one standard deviation increase in structural constraint is associated, on average, with a 0.61 (Model 1, Table 4) standard deviation decrease in information diversity, and a one standard deviation increase in channel bandwidth is associated with a 0.12 (Model 1, Table 4) standard deviation increase in information diversity. Network size yields a positive effect as well, with a 0.48 standard deviation increase in information diversity.

Considering the effects of network structure on the total volume of non-redundant information brokers receive, we find a negative relationship between network constraint and non-redundant information. As brokers' networks become more constrained, they receive less non-redundant information, confirming Burt's primary argument. In contrast, network size and channel bandwidth are both positively associated with non-redundant information. In particular, a one standard deviation increase in network constraint de-

	Model 1	Model 2
GenderDiff_i	-0.001	0.001
	(0.012)	(0.010)
HireDateDiff_i	-0.043*	-0.002
	(0.022)	(0.024)
NetworkSize_{it}	0.476***	0.930***
	(0.032)	(0.007)
$\text{NetworkSize}^2_{it}$	-0.198***	-0.028***
	(0.013)	(0.003)
$\text{NetworkConstraint}_{it}$	-0.614***	-0.275***
	(0.020)	(0.014)
$\text{ChannelBandwidth}_{it}$	0.126***	0.094***
	(0.015)	(0.012)
Constant	0.158***	-0.123*
	(0.042)	(0.050)
Temporal Controls	Month	Month
R^2	0.57	0.25
Chi-squared statistic (df)	3019.4*** (17)	742.3*** (17)

Table 4: (Ego-level) Predicting Information Diversity and Non-Redundant Information (N=2300) using Random Effects regression. **(Model 1) Dependent Variable**= $\text{InformationDiversity}_{it}$, **(Model 2) Dependent Variable**= $\text{NonRedundantInformation}_{it}$ (JointEntropy_{it}). *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

creases non-redundant information by approximately 0.28 standard deviations (Model 2, Table 4), and one standard deviation increase in channel bandwidth is associated with a 0.09 (Model 2, Table 4) standard deviation increase in non-redundant information.

These results confirm the Diversity-Bandwidth Tradeoff and validate and replicate the results of Aral and Van Alstyne (2011) in a completely different organizational setting. The robustness of the findings in this new setting, using a new panel dataset provides strong evidence that the Diversity-Bandwidth Tradeoff is a general phenomenon that holds at the heart of the vision advantage mechanism theorized to explain the Strength of Weak Ties and Brokerage Theory.

5.2 Unpacking the Vision Advantage of Brokers, Bridges, and Weak Ties.

To further understand the mechanics of vision advantages and the Diversity-Bandwidth Tradeoff, we next analyzed the communications network at the level of individual ties. This enabled us to uncover the underlying anatomy of the vision advantage. In particular, we were able to distinguish several different contributions of network constraint at the level of individual ties and analyze them separately. To unpack the relationship between

network constraint and access to novelty, we distinguished the two separate terms of Burt's original constraint variable into its dyadic components: direct investment and redundant investment, as described in Sections 3 and 4.

As can be seen from Table 5, we find a highly significant increase in information diversity received within ties. A one standard deviation increase in ego's direct investment in communicating with a particular alter (the proportion of communication volume they dedicate to that alter) is associated with a 0.19 (Model 1, Table 5) standard deviation increase in information diversity within that dyadic channel. This effect is corroborated by a roughly similar positive effect in redundant investment. Together, these results indicate that the diversity of information that an ego receives within a particular relationship increases with the amount of time and effort they invest, directly and via shared connections, in that peer. Further, we find that higher channel bandwidth also facilitates ($\beta = 0.354$) information diversity within ties. A greater volume of communication with a particular alter is associated with an increase in the diversity of information received.

	Model 1	Model 2	Model 3
GenderDiff_{ij}	0.011	0.036*	0.034*
	(0.012)	(0.014)	(0.015)
HireDateDiff_{ij}	-0.009	-0.011	-0.038***
	(0.006)	(0.007)	(0.008)
NetworkSize_{it}	0.140***	0.159***	0.128***
	(0.006)	(0.006)	(0.007)
$\text{NetworkSize}^2_{it}$	-0.016***	-0.022***	-0.076***
	(0.004)	(0.004)	(0.004)
$\text{DirectInvestment}_{itj}$	0.194***	0.026***	-0.063***
	(0.005)	(0.006)	(0.006)
$\text{RedundantInvestment}_{itj}$	0.088***	0.010	-0.07***
	(0.005)	(0.006)	(0.006)
$\text{ChannelBandwidth}_{itj}$	0.354***	0.186***	-0.087***
	(0.005)	(0.006)	(0.006)
Constant	-0.170***	-0.061***	0.043***
	(0.014)	(0.016)	(0.016)
Temporal Controls	Month	Month	Month
R^2	0.21	0.05	0.03
Chi-squared statistic(df)	12569.4***(18)	2643.2***(19)	1613.6***(18)

Table 5: (Dyad-level) Predicting Information Diversity within a dyad, Non-Redundant Information and Information Uniqueness (N=53079) using Random Effects regression. **(Model 1) Dependent Variable**= Information Diversity within a dyad ($\text{InformationDiversity}_{itj}$), **(Model 2) Dependent Variable**= Non-Redundant Information/Conditional Entropy ($\text{NonRedundantInformation}_{itj}$), **(Model 3) Dependent Variable**= Information Uniqueness ($\text{InformationUniqueness}_{itj}$). ***p < 0.001, **p < 0.01, *p < 0.05

Next, we consider the total amount of non-redundant information conveyed to the ego per tie, as measured by conditional entropy. Again, we consistently find positive associations for direct and redundant investment as well as channel bandwidth. The channel bandwidth seems the most notable variable, with a one standard deviation increase in bandwidth associated with a 0.19 standard deviation increase in non-redundant information. This is also consistent with our findings on information diversity within ties and the amount of non-redundant information ego receives across all his peers (measured by joint entropy).

Finally, when analyzing the information uniqueness between ties, with the receiving ego as the point of reference, we find the opposite effect of direct and redundant investment and bandwidth. Specifically, we observe a decrease in the information uniqueness a tie delivers with increased direct and redundant investment in that tie. This is further supported by the negative relationship between channel bandwidth and information uniqueness. In other words: weak bridging links provide information that is distant from or unique, compared to the information provided by other ties.

These results, considered together, paint a precise picture of the underlying mechanisms that enable vision advantages. Brokers receive more diverse information and more non-redundant information from strong, cohesive, and embedded ties. However, the information that a broker gets from structurally diverse, weak bridging ties is, on average, more unique or different (i.e., more remote in topic space) when compared to the information they receive from their other contacts. In networks comprised of weak bridging ties, the pairwise topical distance between the information provided by networks of strong and cohesive relations is, on average, relatively small. It indicates that structurally weak ties offer unique information, which is significantly different (distant) from the information provided by the brokers' core clique of communication partners. At the same time, the information supplied by structurally weak ties is also more specific or topically narrow (i.e., less diverse).

Reflecting further, the information diversity and the total non-redundant information that the brokers receive both decrease in cohesive or constrained networks (Table 4). Collectively, these results imply that the effect of information diversity within a channel and the effect of information uniqueness are countervailing. As information uniqueness increases, information diversity decreases; that is, the information provided by weak bridging ties is unique and topically narrow. Hence, the total amount of novel information an ego receives through all its contacts is driven more by connections providing unique information.

These results together tell a very compelling story about how vision advantages work —we depict the mechanisms of the vision advantage graphically in Figure 4. As structural diversity increases, the bandwidth of communication channels contracts (The Diversity-Bandwidth Tradeoff). For example, in diverse networks of weak, low bandwidth, and bridging ties, novelty measured across the ties is high (meaning each contact is providing information different from what other contacts are providing), but novelty provided within each channel is decreasing. On the other hand, in constrained networks of strong, high bandwidth embedded ties, novelty across ties decreases due to information overlap and

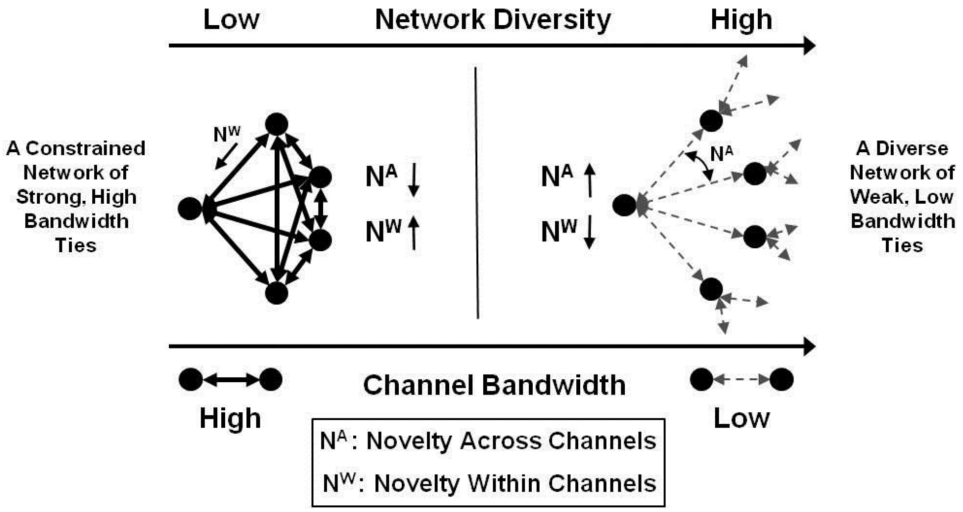

Figure 4: The Mechanics of the Vision Advantage.

redundancy across channels. At the same time, though, the novelty within each channel increases due to the rich, frequent, high bandwidth communication in these dyads.

The mechanisms of the vision advantage become even more apparent when we consider the effects of Longitudinal Entropy (Table 6). Our previous models explain how access to different kinds of novelty (diversity, total non-redundant information, uniqueness) changes as structural variables such as network constraint and channel bandwidth change. However, examining longitudinal measures of novelty allows us to explore how access to novel information changes as actors add new information to what they already know.

	Model 1	Model 2	Model 3	Model 4
ΔNetworkConstraint$_{it}$	-0.690***	-0.723***	-0.673***	-0.739***
	(0.029)	(0.029)	(0.027)	(0.026)
ΔChannelBandwidth$_{it}$	-	0.169***	-	0.313***
		(0.020)		(0.019)
Constant	0.004	0.005	0.139***	0.108***
	(0.008)	(0.008)	(0.026)	(0.025)
R^2	0.25	0.28	0.26	0.36
Chi-squared statistic(df)	577.4***(1)	675.0***(2)	599.0***(1)	956.1***(2)

Table 6: (Ego-level) N=1727 **(Models 1,2) Dependent Variable**= Longitudinal Entropy (Memoryless) LongitudinalEntropy$^{ml}_{it}$. **(Models 3,4) Dependent Variable**= Longitudinal Entropy (with memory) LongitudinalEntropy$^{mem}_{it}$. Only the people who received at least 1 email during all the 12 months of the panel are included. *** p <0.001, ** p < 0.01, * p < 0.05

As illustrated in Figure 5(a), longitudinal entropy systematically reduces over time in the memory model due to the effects of extended memory aggregation (as we aggregate

more information, the novelty of the information we receive is reduced in each subsequent period). As illustrated in Fig. 5(b), in the memoryless model, we find no such trend over time. If we quickly forget what we know, new information seems novel even though we may have seen it in the past. We further explore how the relationship between longitudinal entropy and features of network structure, such as channel bandwidth and network constraint, drive access to novelty over time for each user in the email network from the perspective of learning models with strong or weak memory processes in turn. The relationship between network structure and longitudinal entropy proves highly significant. It is most pronounced in the relationship between longitudinal entropy and network constraint, as illustrated in Fig 5(c). The figure displays the relationship for the memoryless model, but we can also make the same observation in the case of the memory model though it is less precisely estimated.

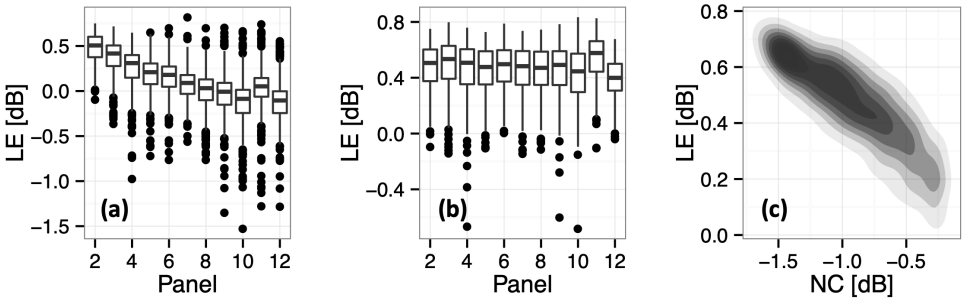

Figure 5: Distribution of longitudinal entropy per panel (month): (a) memory; (b) memoryless. (c) Density plot depicting relationship between longitudinal entropy (LE) and network constraint (NC) for the memoryless model.

In summary, network diversity (or, inversely, network constraint) is the dominant factor in the relationship between network structure and longitudinal entropy by roughly an order of magnitude compared to the channel bandwidth. This result suggests that weak bridging ties, which provide unique information through low bandwidth, structurally diverse channels, contribute the most to the aggregation of novel information over time compared to high-bandwidth, cohesive ties. This highlights the specific importance of weak bridging ties to vision advantages and access to novelty. The unique information provided by weak bridging ties —information that is topically distant from what other connections are providing —is more likely to be different than what we learned in the past or already knew, adding to the longitudinal aggregation of novelty (and avoiding temporal information redundancy) over time.

6 Conclusion

In this paper, we theorized and subsequently analyzed how network structure enables access to novel information. To do so, we conceptualized and developed three different dimensions of informational novelty—diversity, non-redundancy, and uniqueness. We

operationalized these novelty measures using data from a medium-sized digital media firm and used them to analyze the structure and content of the firm's dynamic email network. Finally, we performed empirical analyses to validate the vision advantage argument at the heart of the strength of weak ties and brokerage theories and to understand further the dynamic mechanisms that make vision advantages work. Three results emerged from our analysis.

First, we confirmed the diversity bandwidth tradeoff (DBT) at the heart of the vision advantage. As a broker's network becomes more diverse, the bandwidth of their communication channels contracts, creating countervailing effects on novel information access. These results replicated prior work on the DBT with remarkable fidelity. Second, our analysis uncovered the mechanics driving the DBT and highlighted differences in vision advantages offered by strong, cohesive ties and weak bridging ties. Strong and cohesive ties deliver greater information diversity and more total novelty. In contrast, weak bridging ties contribute the most uniqueness —information that is most different from what other contacts are delivering. Finally, longitudinal entropy, which measures the accumulation of non-redundant information over time, is predominantly driven by network diversity, with bandwidth having a relatively smaller impact. Compared with our former conclusions, this indicates that structurally weak ties, which provide unique information with limited bandwidth and diversity, will contribute most to the aggregation of novel information over time instead of high-bandwidth, cohesive ties.

The theory we propose and our empirical analysis represent the first steps toward a dynamic ego-and-dyadic level model of the vision advantages that have been hypothesized to explain The Strength of Weak Ties and Brokerage Theory for nearly fifty years. The work also highlights the power of combining network structure data with network content data to understand how the structure of social relationships is associated with the information content that flows through them (Sundararajan et al., 2013). Finally, all of these endeavors provide further evidence of the power of micro-level data to uncover social processes driving competitive advantages for networked actors.

Our empirical analyses are not without limitations. Due to the limitations of our setup, both the network structure and the novelty measures were derived using the same email communication data, which could potentially introduce some bias into our parameter estimates. It is an excellent avenue for future work to disentangle the network construction from novelty measurement. Second, our informational novelty measures do not quantify novelty along a particular topical dimension. Potential future work could build new generative models of text data to help us get around this difficulty.

References

Sinan Aral and Marshall Van Alstyne. The diversity-bandwidth trade-off1. *American Journal of Sociology*, 117(1):90–171, 2011.

Sinan Aral, Erik Brynjolfsson, and Marshall W Van Alstyne. Productivity effects of information diffusion in networks. *Available at SSRN 987499*, 2007.

Sinan Aral, Erik Brynjolfsson, and Marshall Van Alstyne. Information, technology, and information worker productivity. *Information Systems Research*, 23(3-part-2):849–867, 2012.

Noah Askin and Michael Mauskapf. What makes popular culture popular? product features and optimal differentiation in music. *American Sociological Review*, 82(5): 910–944, 2017.

Wayne E Baker. Market networks and corporate behavior. *American journal of sociology*, pages 589–625, 1990.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

Nathaniel Bulkley and Marshall W Van Alstyne. Why information should influence productivity. 2004.

Ronald S Burt. Structural holes and good ideas. *American journal of sociology*, 110(2): 349–399, 2004.

Ronald S Burt. *Brokerage and closure: An introduction to social capital*. OUP Oxford, 2005.

Ronald S Burt. Information and structural holes: comment on reagans and zuckerman. *Industrial and Corporate Change*, 17(5):953–969, 2008.

Ronald S Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009.

Wesley M Cohen and Daniel A Levinthal. Absorptive capacity: a new perspective on learning and innovation. *Administrative science quarterly*, pages 128–152, 1990.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Paramveer S Dhillon, Dean Foster, and Lyle Ungar. Multi-view learning of word embeddings via cca. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 199–207, 2011.

Paramveer S Dhillon, Jordan Rodu, Dean P Foster, and Lyle H Ungar. Two step cca: a new spectral method for estimating vector models of words. In *Proceedings of the*

29th International Coference on International Conference on Machine Learning, pages 67–74, 2012.

Paramveer S Dhillon, Dean P Foster, and Lyle H Ungar. Eigenwords: spectral word embeddings. *J. Mach. Learn. Res.*, 16:3035–3078, 2015.

Amir Goldberg, Michael T Hannan, and Balázs Kovács. What does it mean to span cultural boundaries? variety and atypicality in cultural consumption. *American Sociological Review*, 81(2):215–241, 2016.

Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

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

Morten T Hansen. Knowledge networks: Explaining effective knowledge sharing in multiunit companies. *Organization science*, 13(3):232–248, 2002.

Andrew Hargadon and Robert I Sutton. Technology brokering and innovation in a product development firm. *Administrative science quarterly*, pages 716–749, 1997.

Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *science*, 311(5757):88–90, 2006.

David Lazer and Allan Friedman. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4):667–694, 2007.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

C Manning and Hinrich Schütze. *Foundations of statical natural language processing*. MIT Press. Cambridge, MA, 1999.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review E*, 68(3):036122, 2003.

David Obstfeld. Social networks, the tertius iungens orientation, and involvement in innovation. *Administrative science quarterly*, 50(1):100–130, 2005.

John F Padgett and Christopher K Ansell. Robust action and the rise of the medici, 1400-1434. *American journal of sociology*, pages 1259–1319, 1993.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Joel M Podolny. Networks as the pipes and prisms of the market1. *American journal of sociology*, 107(1):33–60, 2001.

Ray Reagans and Ezra W Zuckerman. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization science*, 12(4):502–517, 2001.

Mark C Reynolds, Marshall Van Alstyne, and Sinan Aral. Functions that preserve privacy but permit analysis of text paper 246. 2009.

Simon Rodan and D Charles Galunic. More than network structure: how knowledge heterogeneity influences managerial performance and innovativeness. *Strategic Management Journal*, 25:541–556, 2004.

Jari Saramäki and Kimmo Kaski. Scale-free networks generated by random walkers. *Physica A: Statistical Mechanics and its Applications*, 341:80–86, 2004.

Arun Sundararajan, Foster Provost, Gal Oestreicher-Singer, and Sinan Aral. Research commentary-information in digital, economic, and social networks. *Information Systems Research*, 24(4):883–905, 2013.

Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

Brian Uzzi. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative science quarterly*, pages 35–67, 1997.

Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem1. *American journal of sociology*, 111(2):447–504, 2005.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

Marshall Van Alstyne and Erik Brynjolfsson. Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management Science*, 51(6): 851–868, 2005.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

Duncan J Watts and Steven H Strogatz. Collective dynamics of Ôsmall-worldÕnetworks. *nature*, 393(6684):440–442, 1998.

Fang Wu, Bernardo A Huberman, Lada A Adamic, and Joshua R Tyler. Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337(1-2): 327–335, 2004.

	Model 1	Model 2
GenderDiff_i	-0.002	0.001
	(0.019)	(0.012)
HireDateDiff_i	-0.061*	-0.004
	(0.029)	(0.031)
NetworkSize_{it}	0.512***	0.909***
	(0.026)	(0.012)
$\text{NetworkSize}^2_{it}$	-0.182***	-0.032***
	(0.019)	(0.009)
$\text{NetworkConstraint}_{it}$	-0.633***	-0.313***
	(0.026)	(0.020)
$\text{ChannelBandwidth}_{it}$	0.146***	0.083***
	(0.019)	(0.018)
Constant	0.143***	-0.101*
	(0.034)	(0.041)
Temporal Controls	Month	Month
R^2	0.53	0.31
Chi-squared statistic (df)	3277.7*** (17)	1094.8*** (17)

Table 7: **[Replicating results of Table 4 in the paper with topics estimated via doc2vec]** (Ego-level) Predicting Information Diversity and Non-Redundant Information (N=2300) using Random Effects regression. **(Model 1) Dependent Variable**= $\text{InformationDiversity}_{it}$, **(Model 2) Dependent Variable**= $\text{NonRedundantInformation}_{it}$ (JointEntropy_{it}). *** p < 0.001, ** p < 0.01, * p < 0.05.

7 Appendix

We test the robustness of our results in several ways.

7.1 Estimating Topics using Document Embeddings

We re-estimated the topics using an alternate approach—Document Embeddings (doc2vec) (Le and Mikolov, 2014). doc2vec embeds each document (email in our case) into a real-valued vector which can be used akin to the topics estimated by Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

For our analyses, we used the doc2vec implementation from Gensim[9]. In order to make an apples-to-apples comparison, we chose the dimensionality of the doc2vec embeddingss also as 50 (similar to that of LDA that we used in the paper). Below (Tables 7, 8, 9), we show results that replicate Tables 4,5,6 from the paper using doc2vec embeddings. All other details are the same as in the paper.

[9] https://radimrehurek.com/gensim/models/doc2vec.html

	Model 1	Model 2	Model 3
GenderDiff_{ij}	0.010	0.029*	0.038*
	(0.012)	(0.014)	(0.015)
HireDateDiff_{ij}	-0.009	-0.012	-0.037
	(0.006)	(0.007)	(0.008)
NetworkSize_{it}	0.140***	0.159***	0.133***
	(0.005)	(0.007)	(0.006)
$\text{NetworkSize}_{it}^2$	-0.015***	-0.024***	-0.075***
	(0.004)	(0.005)	(0.004)
$\text{DirectInvestment}_{itj}$	0.190***	0.024***	-0.062***
	(0.006)	(0.006)	(0.006)
$\text{RedundantInvestment}_{itj}$	0.090***	0.011	-0.075***
	(0.005)	(0.006)	(0.006)
$\text{ChannelBandwidth}_{itj}$	0.370***	0.188***	-0.086***
	(0.005)	(0.006)	(0.006)
Constant	0.832***	0.949***	2.04***
	(0.014)	(0.016)	(0.016)
Temporal Controls	Month	Month	Month
R^2	0.13	0.03	0.148
Chi-squared statistic(df)	11907.3***(18)	2505.8.4***(18)	1500.1***(18)

Table 8: [**Replicating results of Table 5 in the paper with topics estimated via doc2vec**] (Dyad-level) Predicting Information Diversity within a dyad, Non-Redundant Information and Information Uniqueness (N=53079) using Random Effects regression. **(Model 1) Dependent Variable**= Information Diversity within a dyad ($\text{InformationDiversity}_{itj}$), **(Model 2) Dependent Variable**= Non-Redundant Information/Conditional Entropy ($\text{NonRedundantInformation}_{itj}$), **(Model 3) Dependent Variable**= Information Uniqueness ($\text{InformationUniqueness}_{itj}$). ***p < 0.001, **p < 0.01, *p < 0.05

	Model 1	Model 2	Model 3	Model 4
ΔNetworkConstraint$_{it}$	-0.694***	-0.726***	-0.682***	-0.744***
	(0.033)	(0.034)	(0.032)	(0.031)
ΔChannelBandwidth$_{it}$	-	0.160***	-	0.292***
		(0.023)		(0.023)
Constant	0.999***	1.00***	1.14***	1.11***
	(0.009)	(0.010)	(0.027)	(0.026)
R^2	0.20	0.22	0.21	0.27
Chi-squared statistic (df)	423.9***(1)	482.2***(2)	449.6***(1)	649.4***(2)

Table 9: **[Replicating results of Table 6 in the paper with topics estimated via doc2vec]** (Ego-level) N=1727 **(Models 1,2) Dependent Variable**= Longitudinal Entropy (Memoryless) LongitudinalEntropy$^{ml}_{it}$. **(Models 3,4) Dependent Variable**= Longitudinal Entropy (with memory) LongitudinalEntropy$^{mem}_{it}$. Only the people who received at least 1 email during all the 12 months of the panel are included. *** p <0.001, ** p < 0.01, * p < 0.05

7.2 Estimating Topics using LDA with varying number of topics

We re-estimated our novelty measures using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to estimate the topics, but by choosing the number of topics to be 25 and 100. Below (Tables 10, 11, 12, 13, 14, 15), we show results that replicate Tables 4,5,6 from the paper using varying number of LDA topics. All other details are the same as in the paper.

7.3 Robustness to different similarity metrics

We re-estimated our novelty measures Information Diversity (at both ego and dyadic-level) and Information Uniqueness using Hellinger Distance and KL-Divergence as metrics to compute distances between the topic distributions. In the paper we used cosine similarity to compute these distances. The correlations between the novelty metrics reported in the paper (using cosine similarity) and those using Hellinger Distance and KL-Divergence are shown below in Tables 16, 17, and 18.

7.4 Robustness to different time-window for calculating Longitudinal measures

We re-estimated Table 6 in the paper using a different temporal window (t= 1 week) for calculating the different measures. The results are shown in Table 19. All other details are the same as in the paper.

	Model 1	Model 2
GenderDiff_i	-0.004	0.003
	(0.014)	(0.012)
HireDateDiff_i	-0.051*	-0.004
	(0.030)	(0.019)
NetworkSize_{it}	0.461***	0.947***
	(0.028)	(0.011)
$\text{NetworkSize}^2_{it}$	-0.174***	-0.033***
	(0.011)	(0.011)
$\text{NetworkConstraint}_{it}$	-0.577***	-0.323***
	(0.024)	(0.025)
$\text{ChannelBandwidth}_{it}$	0.141***	0.120***
	(0.017)	(0.016)
Constant	0.192***	-0.099*
	(0.042)	(0.048)
Temporal Controls	Month	Month
R^2	0.59	0.23
Chi-squared statistic (df)	3061.9*** (17)	889.1*** (17)

Table 10: [**Replicating results of Table 4 in the paper with topics estimated via LDA with 25 topics**] (Ego-level) Predicting Information Diversity and Non-Redundant Information (N=2300) using Random Effects regression. **(Model 1) Dependent Variable**= $\text{InformationDiversity}_{it}$, **(Model 2) Dependent Variable**= $\text{NonRedundantInformation}_{it}$ (JointEntropy_{it}). *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	Model 1	Model 2	Model 3
GenderDiff_{ij}	0.009	0.036*	0.034*
	(0.012)	(0.014)	(0.015)
HireDateDiff_{ij}	-0.009	-0.012	-0.038
	(0.006)	(0.007)	(0.008)
NetworkSize_{it}	0.140***	0.160***	0.128***
	(0.006)	(0.006)	(0.006)
$\text{NetworkSize}_{it}^2$	-0.015***	-0.021***	-0.076***
	(0.003)	(0.004)	(0.004)
$\text{DirectInvestment}_{itj}$	0.193***	0.027***	-0.063***
	(0.006)	(0.006)	(0.006)
$\text{RedundantInvestment}_{itj}$	0.088***	0.010	-0.07***
	(0.005)	(0.006)	(0.005)
$\text{ChannelBandwidth}_{itj}$	0.369***	0.187***	-0.086***
	(0.005)	(0.006)	(0.006)
Constant	0.828***	0.939***	0.043***
	(0.014)	(0.016)	(0.016)
Temporal Controls	Month	Month	Month
R^2	0.14	0.03	0.03
Chi-squared statistic(df)	12230.3***(18)	2578.4***(18)	1613.6***(18)

Table 11: [**Replicating results of Table 5 in the paper with topics estimated via LDA with 25 topics**] (Dyad-level) Predicting Information Diversity within a dyad, Non-Redundant Information and Information Uniqueness (N=53079) using Random Effects regression. **(Model 1) Dependent Variable**= Information Diversity within a dyad ($\text{InformationDiversity}_{itj}$), **(Model 2) Dependent Variable**= Non-Redundant Information/Conditional Entropy ($\text{NonRedundantInformation}_{itj}$), **(Model 3) Dependent Variable**= Information Uniqueness ($\text{InformationUniqueness}_{itj}$). ***p $<$ 0.001, **p $<$ 0.01, *p $<$ 0.05

	Model 1	Model 2	Model 3	Model 4
ΔNetworkConstraint$_{it}$	-0.668***	-0.705***	-0.658***	-0.723***
	(0.031)	(0.031)	(0.030)	(0.029)
ΔChannelBandwidth$_{it}$	-	0.182***	-	0.305***
		(0.021)		(0.021)
Constant	1.01***	1.00***	1.13***	1.10***
	(0.009)	(0.09)	(0.027)	(0.025)
R^2	0.21	0.24	0.21	0.29
Chi-squared statistic (df)	445.5***(1)	533.2***(2)	462.8***(1)	710.3***(2)

Table 12: **[Replicating results of Table 6 in the paper with topics estimated via LDA with 25 topics]** (Ego-level) N=1727 **(Models 1,2) Dependent Variable**= Longitudinal Entropy (Memoryless) LongitudinalEntropy$^{ml}_{it}$. **(Models 3,4) Dependent Variable**= Longitudinal Entropy (with memory) LongitudinalEntropy$^{mem}_{it}$. Only the people who received at least 1 email during all the 12 months of the panel are included. *** p <0.001, ** p < 0.01, * p < 0.05

	Model 1	Model 2
GenderDiff$_i$	-0.002	0.002
	(0.016)	(0.011)
HireDateDiff$_i$	-0.053*	-0.003
	(0.025)	(0.021)
NetworkSize$_{it}$	0.491***	0.956***
	(0.039)	(0.014)
NetworkSize$^2_{it}$	-0.211***	-0.033***
	(0.021)	(0.010)
NetworkConstraint$_{it}$	-0.599***	-0.301***
	(0.034)	(0.019)
ChannelBandwidth$_{it}$	0.139***	0.107***
	(0.010)	(0.018)
Constant	0.122***	-0.134*
	(0.037)	(0.069)
Temporal Controls	Month	Month
R^2	0.62	0.32
Chi-squared statistic (df)	3118.7*** (17)	955.4*** (17)

Table 13: **[Replicating results of Table 4 in the paper with topics estimated via LDA with 100 topics]** (Ego-level) Predicting Information Diversity and Non-Redundant Information (N=2300) using Random Effects regression. **(Model 1) Dependent Variable**= InformationDiversity$_{it}$, **(Model 2) Dependent Variable**= NonRedundantInformation$_{it}$ (JointEntropy$_{it}$). *** p < 0.001, ** p < 0.01, * p < 0.05.

	Model 1	Model 2	Model 3
GenderDiff_{ij}	0.011	0.035*	0.035*
	(0.012)	(0.014)	(0.015)
HireDateDiff_{ij}	-0.008	-0.010	-0.039
	(0.006)	(0.007)	(0.008)
NetworkSize_{it}	0.140***	0.160***	0.129***
	(0.006)	(0.006)	(0.007)
$\text{NetworkSize}^2_{it}$	-0.015***	-0.022***	-0.076***
	(0.006)	(0.004)	(0.004)
$\text{DirectInvestment}_{itj}$	0.193***	0.026***	-0.063***
	(0.006)	(0.006)	(0.006)
$\text{RedundantInvestment}_{itj}$	0.089***	0.010	-0.07***
	(0.005)	(0.006)	(0.006)
$\text{ChannelBandwidth}_{itj}$	0.367***	0.188***	-0.087***
	(0.005)	(0.006)	(0.006)
Constant	0.829***	0.941***	1.043***
	(0.014)	(0.016)	(0.016)
Temporal Controls	Month	Month	Month
R^2	0.14	0.04	0.07
Chi-squared statistic(df)	12440.3***(18)	2618.3***(18)	1584.2***(18)

Table 14: [**Replicating results of Table 5 in the paper with topics estimated via LDA with 100 topics**] (Dyad-level) Predicting Information Diversity within a dyad, Non-Redundant Information and Information Uniqueness (N=53079) using Random Effects regression. **(Model 1) Dependent Variable**= Information Diversity within a dyad ($\text{InformationDiversity}_{itj}$), **(Model 2) Dependent Variable**= Non-Redundant Information/Conditional Entropy ($\text{NonRedundantInformation}_{itj}$), **(Model 3) Dependent Variable**= Information Uniqueness ($\text{InformationUniqueness}_{itj}$). ***p < 0.001, **p < 0.01, *p < 0.05

	Model 1	Model 2	Model 3	Model 4
ΔNetworkConstraint$_{it}$	-0.695***	-0.730***	-0.672***	-0.737***
	(0.030)	(0.030)	(0.029)	(0.027)
ΔChannelBandwidth$_{it}$	-	0.175***	-	0.304***
		(0.021)		(0.020)
Constant	1.00***	1.00***	1.14***	1.11***
	(0.008)	(0.008)	(0.026)	(0.024)
R^2	0.24	0.28	0.24	0.33
Chi-squared statistic (df)	531.8***(1)	624.8***(2)	545.8***(1)	839.4***(2)

Table 15: [**Replicating results of Table 6 in the paper with topics estimated via LDA with 100 topics**] (Ego-level) N=1727 **(Models 1,2) Dependent Variable**= Longitudinal Entropy (Memoryless) LongitudinalEntropy$^{ml}_{it}$. **(Models 3,4) Dependent Variable**= Longitudinal Entropy (with memory) LongitudinalEntropy$^{mem}_{it}$. Only the people who received at least 1 email during all the 12 months of the panel are included. *** p <0.001, ** p < 0.01, * p < 0.05

Variable	1	2	3
1. Information Diversity (Ego/cosine)	1	-	-
2. Information Diversity (Ego/Hellinger)	0.942	1	-
3. Information Diversity (Ego/KL-Div)	0.969	0.935	1

Table 16: Pairwise Correlations between different variables of the panel of 232 employees.

Variable	1	2	3
1. Information Diversity (Dyad/cosine)	1	-	-
2. Information Diversity (Dyad/Hellinger)	0.955	1	-
3. Information Diversity (Dyad/KL-Div)	0.971	0.923	1

Table 17: Pairwise Correlations between different variables of the panel of 232 employees.

Variable	1	2	3
1. Information Uniqueness (Dyad/cosine)	1	-	-
2. Information Uniqueness (Dyad/Hellinger)	0.932	1	-
3. Information Uniqueness (Dyad/KL-Div)	0.957	0.944	1

Table 18: Pairwise Correlations between different variables of the panel of 232 employees.

	Model 1	Model 2	Model 3	Model 4
ΔNetworkConstraint$_{it}$	-0.756***	-0.828***	-0.597***	-0.812***
	(0.041)	(0.045)	(0.033)	(0.041)
ΔChannelBandwidth$_{it}$	-	0.213***	-	0.376***
		(0.026)		(0.024)
Constant	0.010	0.009	0.159***	0.092***
	(0.007)	(0.005)	(0.023)	(0.019)
R^2	0.22	0.31	0.29	0.33
Chi-squared statistic (df)	682.1***(1)	605***(2)	601.4***(1)	1020.2***(2)

Table 19: [**Replicating results of Table 6 in the paper with temporal window (t)= 1 week**] (Ego-level) N=8007 **(Models 1,2) Dependent Variable**= Longitudinal Entropy (Memoryless) LongitudinalEntropy$^{ml}_{it}$. **(Models 3,4) Dependent Variable**= Longitudinal Entropy (with memory) LongitudinalEntropy$^{mem}_{it}$. Only the people who received at least 1 email during all the 12 months of the panel are included. *** p <0.001, ** p < 0.01, * p < 0.05