

Essays on Information, Technology and Information Worker Productivity

by

Sinan Aral

M.P.P. Harvard University, 2001.
M.Sc. London School of Economics, 1999.
B.A. Northwestern University, 1996.

Submitted to the Alfred P. Sloan School of Management in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

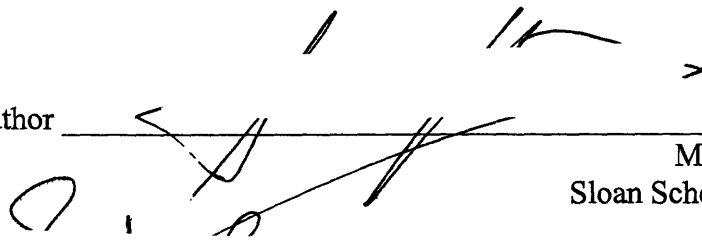
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February, 2007

© Sinan K. Aral. All rights reserved.

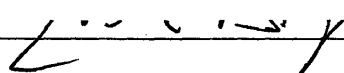
The author hereby grants MIT permission to reproduce and distribute publicly paper and
electronic copies of this thesis document in whole or in part.

Signature of Author



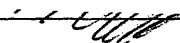
Management Science
Sloan School of Management
January 23, 2007

Certified by

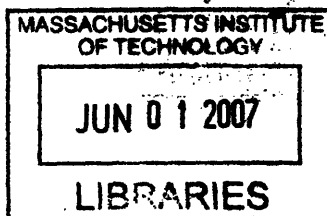


Erik Brynjolfsson
George and Sandi Schussel Professor of Management, MIT Sloan School of Management

Accepted by



Birger Wernerfelt
Professor of Management, MIT Sloan School of Management
Chair, PhD Program



ARCHIVES

Abstract

Essays on Information, Technology & Information Worker Productivity

Sinan Aral

Submitted to the Alfred P. Sloan School of Management in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy in Management Science

Abstract

I examine how information technology (IT) skills and use, communication network structures, and the distribution and flow of information in organizations impact individual information worker productivity. The work is divided into three essays based on the task level practices of information workers at a midsize executive recruiting firm:

Essay 1: “*Information, Technology and Information Worker Productivity: Task Level Evidence.*” I develop and econometrically test a multistage model of production and interaction activities at the firm, and analyze correlations among network structure, characteristics of information flow and real economic output. I find that (a) IT use is positively correlated with non-linear drivers of productivity; (b) the structure and size of workers’ communication networks are highly correlated with performance; (c) an inverted-U shaped relationship exists between multitasking and productivity such that, beyond an optimum, more multitasking is associated with declining project completion rates and revenue generation; and (d) asynchronous information seeking such as email and database use promotes multitasking while synchronous information seeking over the phone shows a negative correlation. These data demonstrate a strong correspondence among technology use, social networks, and productivity for project-based information workers.

Essay 2: “*Network Structure and Information Advantage: Structural Determinants of Access to Novel Information and their Performance Implications.*” I examine relationships between social network structure, information structure, and individual performance. I build and validate a Vector Space Model of information diversity, develop hypotheses linking two key aspects of network structure - size and diversity - to the distribution of novel information among actors, and test the theory using data on email communication patterns, message content and performance. Results indicate that access to diverse, novel information is related to network structure in non-linear ways, and that network diversity contributes to performance even when controlling for the positive performance effects of access to novel information, suggesting additional benefits to network diversity beyond those conferred through information advantage.

Essay 3: “*Organizational Information Dynamics: Drivers of Information Diffusion in Organizations.*” I examine drivers of the diffusion of different types of information through organizations by observing several thousand diffusion processes of two types of information – ‘event news’ and ‘discussion topics’ – from their original first use to their varied recipients over time. I then test the effects of network structure and functional and demographic characteristics of dyadic relationships on the likelihood of receiving each type of information and receiving it sooner. Discussion topics exhibit more shallow diffusion characterized by ‘back-and-forth’ conversation and are more likely to diffuse vertically up and down the organizational hierarchy, across relationships with a prior working history, and across stronger ties; while news, characterized by a spike in communication and rapid, pervasive diffusion through the organization, is more likely to diffuse laterally as well as vertically, and without regard to the strength or function of

Abstract

relationships. The findings highlight the importance of simultaneous considerations of structure and content in information diffusion studies.

Committee

Erik Brynjolfsson (Chair)

Director, Center for Digital Business, Sloan School of Management, MIT,
& George & Sandi Schussel Professor of Management, Sloan School of Management, MIT.

Peter Weill

Director, Center for Information Systems Research, Sloan School of Management, MIT,
& Senior Research Scientist, Sloan School of Management, MIT.

Marshall Van Alstyne

Associate Professor, Boston University School of Management,
& Visiting Professor, Center for Digital Business, Sloan School of Management, MIT.

Acknowledgements

Acknowledgements

I have been blessed with an incredible family of advisors, peers, role models, and friends, who over the years have provided insight, guidance, feedback, and unwavering support and friendship, without which I would have been lost. It is difficult to express how much these people mean to me, both professionally and personally. Without you, none of this would have been possible.

Thank you, first and foremost, to my wife Yonca. You are a blessing. Your love and support mean more to me than you will ever know, and your patience has at times been superhuman. You give me perspective and strength, and make me feel alive. Your love and friendship are my inspiration.

To the members of my committee Erik Brynjolfsson, Peter Weill and Marshall Van Alstyne – thank you! I am proud to call each of you a role model and true friend. Erik, I've never met anyone like you. You are the rarest combination of intellectual horsepower and modesty I have ever encountered. I thank you deeply for the opportunities you have given me, for your guidance which has shaped the way I see the world, and for believing in me and my ideas from day one. You have had a profound influence on me as an academic and as a person. I learn something new from you through every interaction and I continue to be amazed at how much you give to the academic community through your direct efforts and through your influence on others. Your intellectual curiosity is an inspiration to me every day. Peter, I am indebted to you in countless ways. You have been much more than an advisor to me, and I am grateful to have met you. Your unwavering commitment to 'getting it right,' and your leadership in teaching me the importance of communicating my ideas with clarity and precision are two of the most important lessons I have learned over the last six years. You are one of the most effective educators I have ever met and I look to you as a role model on so many professional and personal dimensions. Your intellectual creativity and your ability to orchestrate interactions inspire those who learn from you. I always look forward to our discussions and deeply appreciate the time you spent with me over the years guiding my development both as a person and as a researcher. Marshall, you are ahead of your time. Your intellectual guidance has been critical to the ideas I have developed in this thesis. I deeply appreciate your taking me on as a colleague and collaborator on this exciting project and for your hard work in collecting the initial dataset which, in conjunction with the data I collected, formed the basis of these analyses. Thank you for reminding me to 'focus on information' and the insight that emerges from such a perspective. Thank you for pushing me to always go the extra mile in validating what I learned – a lesson that will stay with me. Thank you also for being a close friend as well as an intellectual instigator. The two go well together. I look forward with great anticipation to future collaboration with all three of you.

Several other members of the MIT Sloan community also provided tremendous support and guidance. To Wanda Orlikowski, thank you for teaching me how to question my most fundamental assumptions and how to approach research with an open mind. To Tom Malone, Stuart Madnick, Benjamin Grosf and Chrysanthos Dellarocas, each of whom had a very direct, personal impact on my thinking, thank you for your support and guidance. Several MIT faculty members outside the IT group also helped me along the way. Roberto Fernandez taught me the value of exogenous change in crystallizing observation, Ezra Zuckerman communicated the importance of a nuanced perspective on network theory, Emilio Castilla helped me situate my thoughts in the context of research outside my immediate purview, and Jonathon Cummings introduced me to the landscape of communication networks research.

Friends and peers in various doctoral programs at MIT both contributed to my work and provided support. Carlos Osorio, my friend, thank you for everything. Our discussions inspired me and at times saved me. To the doctoral students of the Sloan IT group, Tanu Ghosh, David Fitoussi, Xiaoquan Michael Zhang, Yu Jeffrey Hu, Aykut Firat, Adam Saunders, Lynn Wu, and Sumit Bhansali, thank you for your inspiration and support. To the members of the 'dissertation club,' Lourdes Sosa, Hazhir Rahmandad, Kevin Boudreau, and Jeroen Struben you were my strength in the hour of chaos. I look forward to many stimulating interactions for years to come. Adam Seth Litwin, thanks for going beyond the expected. Your insights and methodological advice were essential, especially for the third essay. To Karim Lakhani, thank you for leading by example and interjecting playfulness into my creation process.

Acknowledgements

Over the years, I spent time in two research centers affiliated with the Sloan IT Group. These centers helped create the vibrant intellectual environment I experienced at MIT, and I am indebted to Jeanne Ross, Chuck Gibson, George Westerman and Nils Fonstad of the Center for Information Systems Research (CISR) for their lively and provocative discussions, their advice on my work and their friendship in the daily pursuit of knowledge. Chris Foglia made my last six years feel like home – your influence makes CISR a success. To David Fitzgerald and Julie Hammond-Coiro, you both made every day in the office a little bit easier and a lot more fun. To Carlene Doucette and Yubettys Baez, thanks for always looking out for me. Thank you to David Verrill, executive director of the Center for Digital Business (CDB), for taking me in organizationally and for having confidence in me to represent the CDB in important ways. Thank you to Steve Buckley for your technical support and friendship, to Stephanie Woerner for your friendship and your expert orchestration of the incredibly complex S.E.E.I.T. project. I am also thankful to Rob Laubacher, John Quimby, Peter Gloor and George Herman of the Center for Coordination Science for their support and feedback, and of course, to Sharon Cayley without whom I would not have been able to navigate the complexities of the Sloan PhD program. Your help was critical to my making it through this. Thank you to the National Science Foundation (grants IIS-9876233 & IIS-0085725), Cisco, Intel, and France Telecom for valuable funding; and to Emre Aciksoz, Tim Choe, Abraham Evans-El, Jia Fazio, Saba Gul, Davy Kim, Jennifer Kwon, Petch Manoharn and Jun Zhang for your tireless research assistance.

To my closest friends, Miles, Stefan and Paul, thank you for reminding me to be balanced, for helping me keep one foot in the ‘real world,’ and for looking out for my mental health when I wasn’t. Your spiritual guidance helped me see the finish line.

Finally to my family, thank you! I love you all so much. To my mother and father, you are my heroes. You taught me what it truly means to guide someone in life. I aspire to be half the person each of you is. To my extended family, Lili, Gugi, Altay, Delphine and Ella, thanks for opening your homes and your hearts and for your warmth and love. To my grandmothers, Ezher and Bedia, I love you both. You are a constant source of wisdom and inspiration. Rest in peace to Sekur and Farouk – you both still influence me every day.

Thank you all!

Sinan Aral,
Cambridge, MA.

Table of Contents

Table of Contents

Section	Page
I. Introduction	8
II. Essay 1: “Information, Technology and Information Worker Productivity: Task Level Evidence”	
1. Introduction	16
2. Theory and Literature	18
2.1. Information, Technology and Productivity	18
2.2. Information and Productivity	19
2.3. Social Structure, Information Flows and Information Advantage	19
3. Background and Data	20
3.1. Research Setting: The Role of Information and Technology	20
3.2. Data	22
4. Models and Hypotheses	26
4.1. A Production Model of Revenue and Project Output for Executive Recruiting	26
4.2. Project Level Multitasking	29
4.3. A Model of Project Duration	30
4.4. Alternate Hypotheses and Control Variables	30
5. Statistical Specifications	32
6. Results	33
6.1. Drivers of Production	33
6.2. Relationships Between IT, Information Flows and Multitasking	36
6.3. Relationships Between IT, Information Flows, Multitasking and Project Duration	40
7. Discussion and Conclusion	42
III. Essay 2: “Network Structure and Information Advantage: Structural Determinants of Access to Novel Information and their Performance Implications”	48
1. Introduction	49
2. Theory	52
2.1. Network Structure and Information Advantage: A Critical Inference	52
2.2. Network Determinants of Information Advantage	55
2.3. Non-Network Determinants of Information Advantage	58
2.4. The Setting – Executive Recruiting	59
3. Methods	60
3.1. Data	61
3.2. Modeling and Measuring Information Diversity	64
3.2.1. Modeling and Measuring Topics in Email: A Vector Space Model of Communication Content	64
3.2.2. Construction of Topic Vectors and Keyword Selection	65
3.2.3. Measuring Email Content Diversity	68
3.2.4. Validating Diversity Measures	69

Table of Contents

3.3. Statistical Specifications	71
4. Results	73
4.1. Network Structure and Access to Diverse, Non-Redundant Information	73
4.2. Tradeoffs between Network Size and Network Diversity	75
4.3. Network Structure, Information Diversity and Performance	77
5. Conclusion	79
IV. Essay 3: “Organizational Information Dynamics: Drivers of Information Diffusion in Organizations”	85
1. Introduction	86
2. Theory & Literature	87
2.1. The Central Role of Information in Diffusion Studies	88
2.2. Information Dynamics in Organizations	91
2.2.1. Social Drivers of Organizational Information Diffusion	94
2.2.2. Dimensions of Information Content that Affect Diffusion	97
3. Methods	100
3.1. Data	100
3.2. Identifying Heterogeneous Information Types	104
3.3. Data Structure	110
3.4. Statistical Specifications	110
4. Results	113
4.1. Estimation of the Diffusion of Information	114
4.2. Estimation of the Diffusion of Discussion Topics & Event News	116
5. Discussion & Conclusion	119
V. Conclusion	125

Introduction

Introduction

This thesis examines how information flows and the use of information technology (IT) impact the productivity of information workers. The goal of the research is to better understand the production process of information workers, to identify the intermediate drivers of information worker productivity, and to understand specifically how the flow of information in organizations, IT skills and use, and characteristics of the information to which individuals have access impact their individual and team-level productivity and effectiveness.

The thesis is divided into three essays. The first, entitled “*Information, Technology and Information Worker Productivity: Task Level Evidence*” attempts to characterize the production process of information workers, and to estimate the impact of information and technology on both intermediate and final output measures. The second essay, entitled “*Network Structure & Information Advantage: Structural Determinants of Access to Novel Information and their Performance Implications*” examines relationships between social network structure, information structure, and individual performance, specifically investigating which network structures influence access to diverse and novel information, and whether these relationships explain performance in information intensive work. The third essay, entitled “*Organizational Information Dynamics: Drivers of Information Diffusion in Organizations*” investigates whether social structure and informal relationships in organizations affect the flow of strategic information in email, and asks: Does network position affect who sees strategic information and who sees it sooner? Taken together, these three essays represent the beginning of a program of research dedicated to understanding the role of information and information technology in the productivity and performance of information intensive organizations at the micro level.

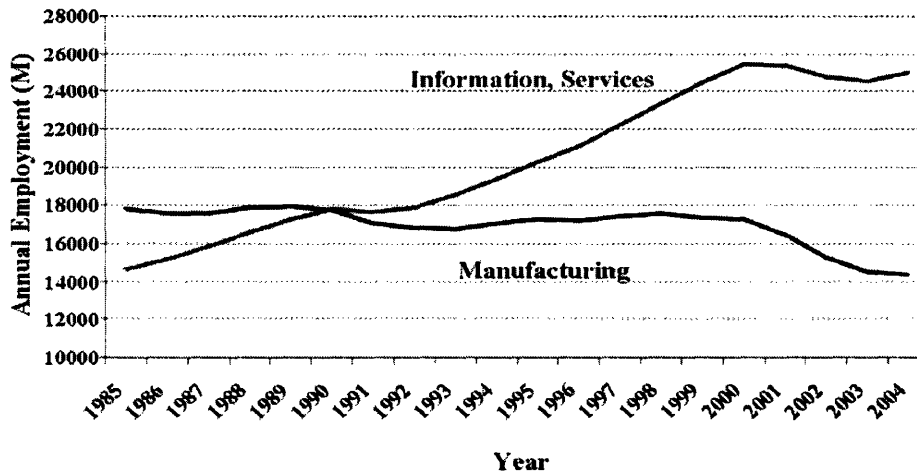
The importance of these topics to both management theory and practice is becoming undeniable as production in developed economies shifts dramatically toward information intensive work. The United States Bureau of Labor Statistics estimates that the number of information and services workers in the

Introduction

United States has nearly doubled since 1985, growing from approximately 14 million to nearly 26 million by 2005. In contrast, manufacturing employment has dropped from 18 million to 14 million (see graph).

U.S. Manufacturing and Information Services Employment: 1985-2005

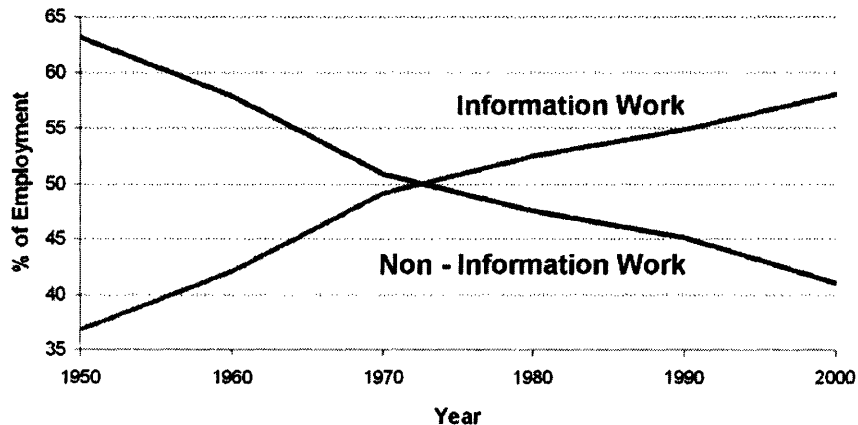
Source: Bureau of Labor Statistics (<http://www.bls.gov/lag/laghome.htm>)



In percentage terms, Wolfe (2005) estimates that in 1950, 36% of the U.S. labor force were information workers, while in 2000 they comprised nearly 60% of all U.S. employees.

Information and Non-Information Work as a % of Employment in the U.S. 1960 - 2000

Source: Wolff (2005)



By some estimates, information workers today make up 70% of the U.S. labor force and account for as much as 60% of the total value added in the U.S. economy (Apte & Nath 2004). Similar statistics are cited for advanced European economies, Japan and advanced regional pockets of emerging economies.

Introduction

That information work is now a cornerstone of production in developed economies is reason enough to justify its examination. But, more importantly, we know little about how to model and measure information worker productivity. Most neo-classical productivity frameworks are geared toward manufacturing work and the production of tangible goods rather than work that uses information to create information based products or services as outputs. For example, classical frameworks estimate the output returns to IT and non-IT capital stock in multifactor productivity estimation, rather than measuring the use of technology or the flow of information itself. As a consequence we know little about the production process or production function of information workers, and lack theories and measures that help us understand what drives productivity in information intensive settings. As Thomas Davenport (2005: 45) recently wrote: “Alas, there is no Fredrick Taylor equivalent for knowledge work. As a result we lack measures, methods and rules of thumb for improvement... Exactly how to improve knowledge-work productivity... is one of the most important economic issues of our time.”

The U.S. Bureau of Labor Statistics defines the “information sector” as “establishments engaged in the following processes: (a) producing and distributing information and cultural products, (b) providing the means to transmit or distribute these products as well as data or communications, and (c) processing data.” To develop a detailed understanding of the specific research questions posed in this thesis, and to provide a set of theory and empirical results that generalize to a well defined population, I examine a slightly narrower subset of this sector. Specifically, I focus on workers whose primary work inputs are information based, whose primary work practices involve assembling, analyzing, and making decisions based on information gathered from various sources, and whose primary work outputs are information based products or services delivered using information created, analyzed or used during production. More specifically, I study executive recruiters, whose work can be described as an information intensive ‘matching process’ that involves matching the right job candidates to clients’ requirements. The empirical evidence examined in this setting, I argue, generalizes to a focused but rather large subset of the information sector including accountants, lawyers, consultants, marketers, insurers, traders, underwriters, computer programmers, actuaries, creative content providers, information service providers, product

Introduction

support staff, as well as a host of other jobs that use, make decisions based on and produce information based goods and services.

My work draws on several rich fields for both theoretical and methodological scaffolding. In the process of answering questions about the role of information and information technology in information worker productivity, I draw on and attempt to contribute to fields as diverse as economics, sociology, and information systems. Streams of literature from these three fields have already accomplished a tremendous amount, addressing questions indirectly related to my area of inquiry, and therefore provide a solid basis of intellectual raw material from which I draw.

Specifically, I draw from a branch of sociology which has in recent years been labeled ‘economic sociology,’ and more specifically from several streams of literature that examine social networks and their implications for the movement of information in organizations and individuals’ performance. I also draw heavily from the information systems literature, which has in the last decade led the development of theory and evidence relating information technology to the productivity and performance of firms, industries and nations. Finally, I draw from the economics literature to form the basis of my thinking about production processes, the estimation of productivity and econometric techniques. In addition to these three main fields, I also rely on work in information retrieval, semantic text mining, and qualitative interviewing. None of these fields serves as a dedicated source of the ideas that make up this thesis. Rather, ideas from all these fields are used in concert to devise new ways of thinking about information, technology and information worker productivity.

The thesis makes use of detailed task level data from a medium sized executive recruiting firm, combined with qualitative interviews and several external sources of data that supplement data collected at the firm. The basic methodological approach in each essay is to begin with a theoretical puzzle, to examine interview data from the firm to understand how these puzzles are instantiated at the research site and to develop hypotheses, and then to specify estimating equations to test hypotheses that help unpack and shed light on the original puzzle. The strategy is akin to the “insider econometrics” described by Bartel, Ichniowski, & Shaw (2004). Rather than providing detailed descriptions of each essay, the

Introduction

methods used or the results obtained in the introduction, I provide detailed abstracts and introductions for each essay separately and leave to the conclusion a discussion of the results and their collective implications for theory and practice.

I find a substantial program of correspondence between information, technology and information worker productivity at the task level. Information technology use and skills are correlated with non-linear drivers of productivity, such as multitasking behavior – the degree to which information workers take on multiple simultaneous projects. Multitasking behavior has an inverted-U shaped relationship with productivity such that, beyond an optimum, more multitasking is associated with declining project completion rates and revenue generation. The structure and size of communication networks – the social means through which workers access strategic information – are highly correlated with performance. Workers with diverse networks, who are also in the thick of the flow of information in the firm, are more likely to multitask more effectively, increasing their productivity.

I also address a long standing inference in the social network literature that has never been empirically tested. A growing body of evidence links the structural properties of individuals' and groups' networked relationships to various dimensions of economic performance. However, the mechanisms driving this linkage, thought to be related to the value of the information flowing between connected actors, are typically inferred, and rarely empirically demonstrated. One of the most prominent mechanisms theorized to drive the relationship between social structure and performance is the existence of 'information benefits' to network structure – that actors in favorable structural positions enjoy social and economic advantages based on their access to specific types of information. By examining the intermediate mechanisms driving the relationship between network structure and performance, I find that diverse networks enhance performance by providing access to diverse information, but that network diversity contributes to performance even when controlling for the positive performance effects of access to novel information, suggesting additional benefits to network diversity beyond those conferred through information advantage. I also find nuanced non-linear relationships between network size and network diversity which help explain the limits of access to non-redundant contacts in bounded organizational

Introduction

networks. Specifically, I find that the total amount of novel information and the diversity of information flowing to actors are increasing in their network size and network diversity. However, the marginal increase in information diversity is decreasing in actors' network size, a result explained in part because as actors establish relationships with a finite set of possible contacts, the probability that a marginal relationship will be structurally non-redundant (and therefore provide novel information) decreases as possible contacts in the network are exhausted.

Finally, I find that the flow of information in the firm is guided by several social and structural factors that determine who is more likely to see strategic information and who is more likely to see it sooner – contributing to the relative success (or failure) of individual recruiters. I find that the diffusion of news, characterized by a spike in communication and rapid, pervasive diffusion through the organization, is influenced by demographic and network factors but not by functional relationships (e.g. prior co-work, authority) or the strength of ties. In contrast, diffusion of discussion topics, which exhibit more shallow diffusion characterized by 'back-and-forth' conversation, is heavily influenced by functional relationships and the strength of ties, as well as demographic and network factors. Discussion topics are more likely to diffuse vertically up and down the organizational hierarchy, across relationships with a prior working history, and across stronger ties, while news is more likely to diffuse laterally as well as vertically, and without regard to the strength or function of relationships, demonstrating the importance of simultaneous considerations of structure and content in information diffusion studies.

The methods developed in this thesis are replicable and portend new frontiers for both IT productivity and social network research. First, previous estimates of the relationship between IT and productivity demonstrated correspondence between firms' total IT capital stock and their productivity. However, estimation of relationships between information, technology and productivity at the task level enable examinations of *how* IT improves productivity, moving beyond the question of *whether* IT improves productivity. In addition, the use of email data to characterize and estimate communication networks is an important emerging area of social network research that can help overcome several known sources of estimation bias. However, such methods are not without their own idiosyncratic difficulties. I

Introduction

therefore discuss both methodological and empirical implications of this type of research. Finally, systematic examination of communication content opens new areas of inquiry into the ‘information benefits’ derived from social network structure and how information diffuses through social groups.

The methods developed and the empirical findings revealed in this thesis through task-level productivity estimation and message-level information and communication content estimation, are essential steps forward in our understanding of the roles of information and technology in the production process of information workers. Understanding information worker productivity (and how information and technology contribute to information worker productivity) is becoming more and more critical as information work begins to represent the lion’s share of production activities in developed economies.

References

- Apte, U., Nath, H. 2004. “Size, structure and growth of the U.S. economy.” Center for Management in the Information Economy, Business and Information Technologies Project (BIT) Working Paper.
- Bartel, A., C. Ichniowski, & K. Shaw. 2004. “Using insider econometrics to study productivity.” *American Economic Review* 2(1) 217-223.
- Davenport, T.H. 2005. Thinking for a living: How to get better performances and results from knowledge workers. Harvard Business School Press. Cambridge, MA.
- Wolfe, E.N. 2005. “The growth of information workers in the U.S. economy.” *Communications of the ACM* 48(10): 37-42.

Essay 1: “Information, Technology and Information Worker Productivity: Task Level Evidence”

Abstract

In an effort to reveal the fine-grained relationships between IT use, patterns of information flows, and individual information-worker productivity, we study task level practices at a midsize executive recruiting firm. We analyze both project-level and individual-level performance using: (1) detailed accounting data on revenues, compensation, project completion rates, and team membership for over 1300 projects spanning 5 years, (2) direct observation of over 125,000 e-mail messages over a period of 10 months by individual workers, and (3) data on a matched set of the same workers' self-reported IT skills, IT use and information sharing. These detailed data permit us to econometrically evaluate a multistage model of production and interaction activities at the firm, and to analyze the relationships among key technologies, work practices, and output. We find that (a) IT use is positively correlated with non-linear drivers of productivity; (b) the structure and size of workers' communication networks are highly correlated with performance; (c) an inverted-U shaped relationship exists between multitasking and productivity such that, beyond an optimum, more multitasking is associated with declining project completion rates and revenue generation; and (d) asynchronous information seeking such as email and database use promotes multitasking while synchronous information seeking over the phone shows a negative correlation. Overall, these data show statistically significant relationships among technology use, social networks, completed projects, and revenues for project-based information workers. Results are consistent with simple models of queuing and multitasking and these methods can be replicated in other settings, suggesting new frontiers for IT value and social network research.

Key words: Productivity, Information Worker, Information Technology, Task-Level Evidence, Social Networks, Multitasking, Production Function.

History: Awarded Best Paper at the 27th Annual International Conference on Information Systems, Milwaukee, WI, 2006.

Essay 1: Information, Technology & Information Worker Productivity

“In the physical sciences, when errors of measurement and other noise are found to be of the same order of magnitude as the phenomena under study, the response is not to try to squeeze more information out of the data by statistical means; it is instead to find techniques for observing the phenomena at a higher level of resolution. The corresponding strategy for [social science] is obvious: to secure new kinds of data at the micro level.”
-- *Herbert Simon*

1. Introduction

Information workers now account for as much as 70% of the U.S. labor force and contribute over 60% of the total valued added in the U.S. economy (Apte & Nath 2004). Ironically, as more and more workers focus on processing information, researchers have less and less information about how these workers create value, and managers have greater difficulty measuring, managing and optimizing work. Unlike bushels of wheat or tons of steel, the real output of most information workers is difficult to measure. Counting meetings attended or memos filed is not closely linked to the value these activities create. But, as the information content of work increases, the role of information becomes increasingly central to our understanding of the performance of individuals, groups and organizations. This paper explores the relationship between information, technology and information worker productivity, using detailed empirical evidence to examine how IT use and information seeking habits affect individual level output. Our findings not only uncover relationships among IT use and skill, communication behaviors, social networks, and productivity, they shed light on the underlying mechanisms that drive performance.

By studying a single industry in depth, Ichniowski, Shaw and Prennushi (1997) were able to specify a precise blue collar production function for steel finishing lines, and measure the effects of particular work practices and technologies on productivity. The corresponding strategy for comprehending information work is clear: to secure task-level data for a specific group of information workers. Our study focuses on executive recruiters, or “head hunters,” whose primary work involves filling clients’ job openings. Output is precisely observable in this setting because accounting data provide complete and detailed records of project-level and individual-level revenues, the number of projects completed, project duration, the number of simultaneous projects, and project and individual-

Essay 1: Information, Technology & Information Worker Productivity

level characteristics. With the company's and employees' cooperation, we also monitored email usage and conducted detailed surveys and interviews focusing on activities, skills, behaviors, and perceptions relevant to information work.

Our IT variables focus on the *use* of technology, not merely its presence, and include direct, message-level observation of communications volume, the size and shape of email contact networks, professed ability to use database technology, and relative time spent on various information seeking tasks. When combined with interviews and visits, these data enabled us to specify and estimate several equations relating technology, skill, social network structure, worker characteristics, task completion and revenue generation. Narrowly focusing on one industry allowed us to precisely define the white collar production process, and our concentrated data collection from one firm eliminates many sources of heterogeneity that confound productivity estimation at more aggregate levels of analysis.

Our results demonstrate that information flows and IT use do in fact predict significantly higher levels of economic productivity. Employees that use databases more also conduct more work simultaneously and finish projects faster. Heavier database users generate more revenue for the firm per unit time. But our analyses at the task level, designed to unpack the processes driving performance, also reveal some counterintuitive results. We find that individuals occupying central brokerage positions in the firm's communication network, who arguably have more structurally efficient access to novel information, are not necessarily more efficient *per project*. Instead, their higher levels of productivity are driven by higher capacities to multitask across simultaneous projects. Richer communications structure predicts greater multitasking, and multitasking drives productivity, demonstrating that technology use not only speeds work, it enables new ways of working that can make workers more productive. Our results reveal a substantial program of correspondence among information, technology and output, and motivate new questions regarding the tradeoffs between multitasking and the speed of work, and how information affects intermediate production processes in white collar work.

2. Theory and Literature

2.1. Information, Technology and Productivity

Historically, technological revolutions have triggered sustained increases in productivity (David 1990). In the information age, new technologies, new ways of working, and an increasing availability of information could significantly affect productivity growth, and specifically, the productivity of workers in information-intensive industries. From 1995 to 2005, annual productivity growth in the U.S. averaged more than 3%, more than doubling the rate in the preceding two decades. A growing body of literature links these productivity gains to IT-intensive industries and firms. Studies of the relationship between IT and economic productivity have examined empirical evidence at the country (e.g. Dewan & Kraemer 2000), industry (e.g. Jorgenson & Stiroh 2000), and firm (e.g. Brynjolfsson & Hitt 1996) levels, demonstrating a convincing positive relationship across distinct measures (Brynjolfsson & Hitt 2000, Bharadwaj et. al. 1999). A handful of task-level studies of IT and productivity have been conducted in recent years (e.g. Ichniowski, Shaw & Pernushi 1997, Barua, Kreibel & Mukhopadhyay 1994, Mukhopadhyay 1997, McAfee 2002). However, most of these studies focus on the manufacturing sector and measure outputs pertaining to the production or distribution of physical goods, leaving a number of important questions unanswered. For example, the mechanisms by which IT affects productivity are not well understood and the output and production function for information workers such as managers, consultants, researchers, marketers, lawyers and accountants remain poorly modeled and measured.

Information technologies may be particularly important for the productivity of information workers not only because IT enables information workers to search for, retrieve, analyze and store information, but because technologies such as email enable new forms of work organization and communication that are increasingly asynchronous, geographically dispersed and sustained over longer periods of the day (e.g. Hinds & Kiesler 2002). As information work represents a growing proportion of the GDP, and is increasingly leveraged with IT, understanding IT and productivity in the context of information work is especially important. Accordingly, we seek to explore a new frontier for IT

Essay 1: Information, Technology & Information Worker Productivity

productivity research by using detailed task-level data to open the black box of the firm and understand how information and technology affect information work at the level of individual workers.

2.2. Information and Productivity

Two broad theoretical arguments contend that information should enable increased productivity (Buckley & Van Alstyne 2004). First, reductions in uncertainty can improve resource allocations and decision making, and reduce delay costs (Cyert & March 1963, Galbraith 1973). In our context, more precise or accurate information about the candidate pool can reduce time wasted interviewing candidates unsuitable for a given executive search. Uncertainty exists for recruiters when information is inaccurate, out of date, hard to find, or imprecise, and decisions based on faulty or incomplete information make filling positions more difficult. Precise information also tempers risk aversion, enabling actors to make appropriate decisions faster (Arrow 1962, Stiglitz 2000). Reductions in uncertainty help recruiters place the right candidates in front of the right clients at the right time, increasing the likelihood of concluding searches faster and, therefore, increasing contract execution per unit time. Second, sharing procedural information or know-how can improve the efficiency with which employees handle recurrent search problems. Knowledge sharing on difficult recurring situations improves effectiveness (Szulanski 1996), although at times complex knowledge may be tied to particular contexts (Von Hippel 1998) or difficult to transfer (Hansen 1999). In interviews, executive recruiters report learning to deal with difficult situations through communication with peers.

2.3. Social Structure, Information Flows and Information Advantage

If information influences productivity, its distribution and diffusion patterns are likely to affect the *relative* productivity of individuals and groups. Over several decades, social network research has examined how information can alter competitive dynamics, access to resources, awareness of opportunity, negotiating leverage, teamwork and ultimately performance. For example, individuals whose networks contain many structural holes may derive information and control benefits from the lack of connectivity

Essay 1: Information, Technology & Information Worker Productivity

among people in their network (Burt 1992), with their access to more non-redundant information making them more likely to receive early promotion (Burt 1992), enjoy greater career mobility (Podolny & Baron 1997), and adapt more quickly to change (Gargiulo & Benassi 2000).

Others argue that cohesion is more important for group performance than structural diversity because information in cohesive groups is more complete, fostering stronger norms of trust, reciprocity and familiarity, and improving the precision with which actors understand their environments (Coleman 1988). Reagans & Zuckerman (2001) show that cohesion within and structural holes across groups improve the innovation output of R&D teams. Podolny & Baron (1997) find that while cohesive ties are beneficial in ‘buy-in’ networks and for those contacts that have control over the fate of employees, structural holes are important in advice and information networks. Hansen (2002) finds that business units with shorter path lengths to other units that possess related knowledge finish projects faster, and that advice giving and advice receiving ties have differential impacts on project duration.

We seek to complement and extend this body of work by addressing an understudied yet fundamental question at the heart of the relationship between social network structure and economic performance: Are individuals and groups in favorable structural positions actually more *productive*? By addressing a performance dimension whose evaluation is removed from social influence, we avoid the endogeneity of socially derived peer performance evaluation apparent in a great deal of social network research.

3. Background and Data

3.1. Research Setting: The Role of Information and Technology

We studied a medium-sized executive recruiting firm over five years, with fourteen regional offices throughout the U.S. The employees occupy three basic positions – partner, consultant and researcher, and conduct their ‘searches’ in teams. Our interviews indicate that the process for securing and executing a contract is relatively standard: A partner secures a contract with a client and assembles a

Essay 1: Information, Technology & Information Worker Productivity

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

The core of executive recruiters' work involves retrieving and understanding clients' requirements and matching candidates to those requirements.¹ This matching process is information-intensive and requires activities geared toward assembling, analyzing, and making decisions based on information gathered from various sources including team members, other firm employees, contacts outside the firm, and data on potential candidates in the internal proprietary database, external proprietary databases, and public sources of information.

Recruiters earn revenue for the firm by filling vacancies, rather than billing hourly. Therefore, the speed with which vacancies are filled is an important intermediate measure of workers' productivity. Contract completion implies that the search team has met the client's minimum thresholds of candidate fit and quality, and given controls for differences across characteristics of contracts (e.g. job type, location), project duration (in addition to the real dollar output value of each contract) can be interpreted as a quality controlled measure of team and worker productivity.

Interviews with the CIO and other employees indicate that the firm uses IT in essentially two ways: 1) as a communication vehicle (e.g. phone, voicemail, and email) and 2) as a central repository of information and knowledge about ongoing projects, potential candidates and internal task coordination.

¹ "Client" refers to a firm seeking to hire one or more executives; "candidate" refers to a potential hire; and "recruiter" refers to someone expert in locating, vetting, and placing candidates.

Essay 1: Information, Technology & Information Worker Productivity

Both of these functions facilitate the information exchanges teams require to systematically assemble, analyze, codify and share knowledge about candidates and clients.

The firm pays to use external databases and has its own proprietary Executive Search System (ESS), built from an off-the-shelf relational database. The ESS not only provides a repository of information on current and past projects, the firm's own employees (e.g. contact information, areas of expertise, work history and current assignments), clients, and potential candidates (e.g. resumes, prior due diligence, and notes or "work ups" on their previous jobs); it also helps employees coordinate and manage dependencies across projects. For example, when searching for potential candidates, employees must honor contractual obligations that prevent poaching employees of past clients for one year. The ESS maintains an up-to-date record of candidates that are 'frozen' due to prior client obligations and employees use this information to coordinate contractual obligations across projects while selecting potential candidates.

3.2. Data

Data for this study include three separate data sets from the firm and one from outside the firm. The first is exact internal accounting records of: (i) revenues generated by individual recruiters, (ii) contract start and stop dates, (iii) projects handled simultaneously by each recruiter, (iv) labor costs and compensation, (v) project team composition, (vi) job levels of recruiters, and (vii) job levels of placed candidates. Accounting data cover the period 2001-2005. These provide excellent output measures that can also be normalized for quality.

The second set of data covers 10 months of complete email history captured from the corporate mail server during two equal periods from October 1, 2002 to March 1, 2003, and from October 1, 2003 to March 1, 2004. Email data has the potential to overcome bias in survey respondent recall of their social networks (see the 'BKS Studies': e.g. Bernard et. al 1981) by objectively recording who is communicating with whom and when. However, it is not without its own limitations as a source of data. We therefore took great care in collecting and analyzing our social network data. We wrote and

Essay 1: Information, Technology & Information Worker Productivity

developed capture software specific to this project and took multiple steps to maximize data integrity and levels of participation. New code was tested at Microsoft Research Labs for server load, accuracy and completeness of message capture, and security exposure. To account for differences in user deletion patterns, we set administrative controls to prevent data expunging for 24 hours (Van Alstyne & Zhang, 2003). The project went through nine months of human subjects review prior to launch and content was masked using cryptographic techniques to preserve individual privacy. Spam messages were excluded by eliminating external contacts who did not receive at least one message from someone inside the firm. Participants received \$100 in exchange for permitting use of their data, resulting in 87% coverage of recruiters eligible to participate and more than 125,000 email messages captured.²

The third data set contains survey responses on information-seeking behaviors, perceptions, experience, education, human factors, and time allocation. Survey questions were generated from a review of relevant literature and interviews with recruiters. Experts in survey methods at the Inter-University Consortium for Political and Social Science Research vetted the survey instrument, which was then pre-tested for comprehension and ease-of-use. Individual participants received \$25 for completed surveys and participation exceeded 85%.

The fourth data set involves independent controls for placement cities to normalize for project difficulty and will be described below. Together, these data provide a desktop-level view of information flows and IT use that we matched to precise measures of individual performance. Aggregating individual revenues also provides a complete picture of firm-level revenues.

Following our qualitative assessment of the role of IT in the firm's production process, we concentrated our measurement of IT around (a) the intensity and skill with which employees used the ESS system, and (b) the frequency of use of different modes of communication in maintaining contacts and seeking information. In measuring ESS skill, we asked respondents to evaluate (i) their personal effectiveness using the ESS system and (ii) their ability to find, add, and modify the records it contains.

² *F*-tests comparing performance levels of those who opted out with those who remained did not show statistically significant differences. *F* (Sig): Revenue02 2.295 (.136), Compensation02 .837 (.365), Multitasking02 .386 (.538).

Essay 1: Information, Technology & Information Worker Productivity

As these two factors were highly correlated (Spearman = .88***, $\alpha = .94$), we combined them into a single measure. To measure ESS use intensity, we asked respondents to estimate the proportion of time they spent gathering information from the ESS and external databases in order to perform their work. Finally, we asked respondents to estimate the number of people they communicated with in a typical day face-to-face, over the phone, and over email.³

To measure information flows, we constructed variables for both the *levels* and *structure* of email traffic. Since teams at our research site are small – between one and five people – we focus on the global network structure of teams, rather than on their internal structure. Measures of the level of email traffic count the total number of emails sent and received, individuals' network size, and their in-degree and out-degree, which measure individuals' frequency weighted number of contacts. Measures of communication structure include the '*betweenness centrality*' of an individual's email network $B(n_i)$ (Freeman 1979),⁴ which measures the probability that the individual will fall on the shortest path between any two other individuals linked by email communication and the '*constraint*' of the network C_i (Burt 1992: 55),⁵ which measures the degree to which an individual's contacts are connected to each other (a proxy for the redundancy of contacts):

$$B(n_i) = \sum_{j < k} g_{jk}(n_i) / g_{jk};$$
$$C_i = \sum_j \left(p_{ij} + \sum_q p_{iq} p_{qj} \right)^2, \quad q \neq i, j.$$

We examine degree and betweenness centrality measures as proxies for the likelihood of being privy to a useful piece of strategic information, and structural holes to capture the efficiency with which teams and individuals have access to non-redundant information.

³ As we also have an objective measure of this value, we assessed the accuracy of survey responses. Respondents reported a mean number of email contacts equal to 28.1, while the email data revealed a mean of 34.8 (Individual mean email contacts = 28.1, team mean = 20.1). We could not reject the hypothesis that the difference between these means was zero at the 95% level.

⁴ Where g_{jk} is the number of geodesic paths linking j and k and $g_{jk}(n_i)$ is the number of geodesic paths linking j and k involving i .

⁵ Where $p_{ij} + \sum_q p_{iq} p_{qj}$ measures the proportion of i 's network contacts that directly or indirectly involve j and C_i sums this across all of i 's contacts.

Essay 1: Information, Technology & Information Worker Productivity

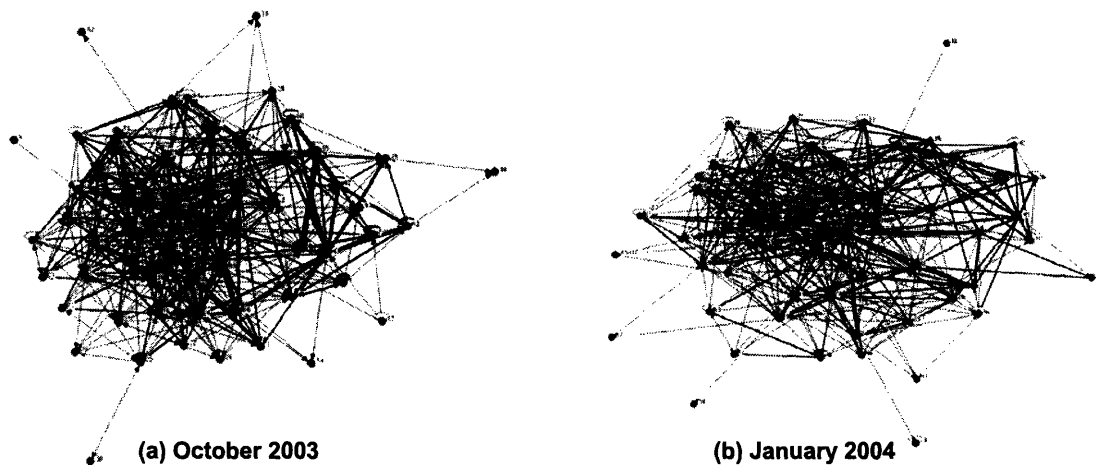


Figure 1. We use email messages to map the social network at this firm. Each node represents an individual in our data set, while the thicknesses of the links represent the amount of email traffic between individuals over two four week periods.

We distinguish between incoming and outgoing email to proxy for differences between information seeking and information provision and, in order to control for the overall level of communication, control for the total amount of email in our analyses. Two four week patterns of email traffic are shown in Figure 1 while Table 1 provides descriptive statistics for all variables, and Appendix A provides their descriptions and data sources.

Essay 1: Information, Technology & Information Worker Productivity

Table 1: Descriptive Statistics					
Variable	Obs.	Mean	SD	Min	Max
<i>Project Team Variables</i>					
Team Size	1382	1.98	.60	1	5
Age	1372	45.07	7.77	27	63
Yrs Education	1372	17.74	1.02	15	20
Industry Experience	1372	14.47	7.94	1	39
Multitasking	1382	8.86	2.84	1.60	18.31
Project Duration (Days)	1382	206.90	123.69	3	981
Project Revenue Value (\$)	1301	56962.5	25780.7	11666	237636
Team Interdependence	1382	1.36	.749	.05	4.65
Task Routiness	1382	1.18	.88	.05	4
F2F Contacts	1382	4.20	8.68	0	75
Phone Contacts	1382	15.76	10.54	1	70
Email Contacts	1382	20.14	18.46	1	100
ESS (Database) Skill	1382	3.10	1.92	.12	9.30
ESS (Database) Use (%)	1382	15.79	14.45	0	80
Total Emails	1382	1365.67	760.19	.6	3939
Total Emails Sent	1382	667.89	393.96	.3	1985
Total Emails Received	1382	697.79	378.34	.3	1954
Degree Centrality	1382	1295.23	720.24	.6	3584
In Degree	1382	632.66	373.01	.3	1804.2
Out Degree	1382	662.56	360.04	.3	1854
Network Size	1382	37.80	11.90	.6	79.36
Betweenness	1382	37.55	26.12	0	185.69
Constraint (1-Structural Holes)	1382	.18	.07	.02	.49
<i>City Characteristics</i>					
Cost of Living	1187	358.65	144.49	233.60	2059.60
Crime per Capita	1187	6262.40	2648.76	0	14603.80
Sunny Days per Annum	1187	212.15	33.93	23	300
Commute Time (Minutes)	1187	20.22	5.38	9	43
<i>Individual Variables - Daily</i>					
Daily Project Output	104982	.017	.017	0	.84
Daily Revenue Output	100815	694.82	690.24	0	3353.35
Multitasking	104983	6.55	5.51	0	28
Share Weighted Multitasking	104983	3.36	2.91	0	14.25
Average Project Duration	107658	212.01	158.55	0	1218.75

4. Models and Hypotheses

4.1. A Production Model of Revenue and Project Output for Executive Recruiting

A decade ago, moving from aggregate data to more fine grained data at the firm level helped resolve the ‘IT productivity paradox.’ Explorations at the firm level, however, are still constrained by the granularity of the data and thus can only explain *whether* IT increases productivity, not *how* IT increases productivity. Our data allow us to construct a detailed model of the production process of executive

Essay 1: Information, Technology & Information Worker Productivity

recruiters, and to test the impact of IT and information flows on intermediate process metrics and final output measures. We conduct both individual and project-level analyses that examine the specific mechanisms through which IT and information affect the production process of information workers.

As a first step in model development, we took a more traditional approach and examined the relationship between IT and revenues directly. We also evaluated a popular conception of how IT may improve productivity: by increasing the pace of work. There has been much discussion of how IT speeds work activities into the “fast lane” and drives business at “Internet speed.” All else being equal, faster completion of projects should lead to more revenues. Indeed, in our exploratory analysis, we did find a positive and statistically significant correlation between IT and revenue. However, to our surprise, we also found that our IT and information flow variables were actually correlated with *longer* project duration on average (see Table 2).

Variables	Model 1	Model 2	Model 3
<i>Dependent Variable:</i>	Revenue	Multitasking	Duration
Partner	354,668.03** (101188.43)	2.63 (2.06)	16.38 (36.72)
Consultant	420,625.63*** (86713.60)	2.39 (1.76)	20.13 (45.19)
Internal Email	11,657.50*** (2102.09)	.126** (.043)	1.91* (.987)
ESS Skill	326.32* (194.74)	.009** (.004)	.169** (.083)
<i>Controls</i>	Gender, Education, Industry Experience	Gender, Education, Industry Experience	Gender, Education, Industry Experience
Adj R ²	.53	.24	.18

***p<.001; **p<.05; *p<.10. OLS analysis on yearly variables in 2002.

This seeming paradox indicated that our simple model of production in recruiting firms was not accurate. While IT seemed to help individual workers bring more revenue to the firm, it was not simply speeding up their work. Further interviews revealed that employees often vary the number of projects they work on at a time such that workers’ revenues are a function not only of how fast they work, but also of how much they multitask.

Essay 1: Information, Technology & Information Worker Productivity

In our revised production model, employees work on projects whose number and duration determine total dollar “bookings” (contracts landed) and “billings” (contracts executed) revenue. If we consider white collar workers to be managing queued tasks, each with distinct start and stop points, we can measure the relationship between IT, information flows, and intermediate measures of output. In particular, data on project multitasking, and start and stop times over the sample period, index the rate at which projects are completed. These relationships are depicted in Figure 2.

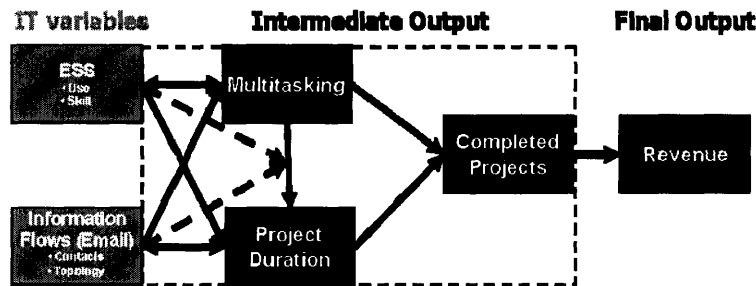


Figure 2. Our model of the production function represents a set of queued job tasks. The influence of IT and Information Flows can then be examined at the task level.

An aggregate model of production activity can be specified as in equation (1). This specification resembles that of Ichniowski, Shaw & Prennushi (1997), and increment to R^2 , PE and Box-Cox tests indicate this additive form is preferred to a multiplicative Cobb-Douglas specification.

$$(1) \quad Q_i = \alpha + \beta H_i + \gamma X_i + \delta Y_i + \varepsilon_i$$

The determinants of output (Q_i) in eq. (1) include dummy variables (H_i) for the job level of individual workers; human capital (X_i) reflected in recruiters’ age, gender, educational attainment and years of experience; IT and information flow variables (Y_i); constant (α) and error terms.⁶ In different models, Q_i represents revenues, completed projects, or the number of simultaneous projects -- depending on the hypothesis. In contrast to earlier work, IT capital and non-IT capital are constant across all observations (i.e. recruiters) and are thus included in the constant term. Instead, the IT variables of interest pertain to IT skill and use of the technology, not merely its presence. In another contrast to traditional IT-

⁶ i indexes either projects or individuals depending on the analysis. Output per unit time is measured at the individual level. But, project outcomes are a result of joint, not individual effort. For project-level outcomes, we therefore analyze and measure team-level variables of interest.

Essay 1: Information, Technology & Information Worker Productivity

productivity research, we include intermediate performance measures (e.g. multitasking, project duration) to estimate steps of the production process separately.

4.2. Project Level Multitasking

We developed project-level and individual-level measures of multitasking based on the multitasking profiles of each individual employee over every day of the five year time span of the study. A multitasking profile characterizes the projects an employee is engaged in during any given day, including not only the number of simultaneous contracts assigned to an employee, but also their relative share of project effort, the job types of the projects (e.g. the job classes of the projects and the cities in which they are based), and the dollar value of each project for the firm. With these data we constructed an individual multitasking measure weighted for effort share, and a team level multitasking measure tracking the average number of other projects a project team is working on during a focal project again weighted by assigned effort shares. Figure 3 displays a multitasking profile for one employee during the period 9/05/2002 to 11/26/2002.

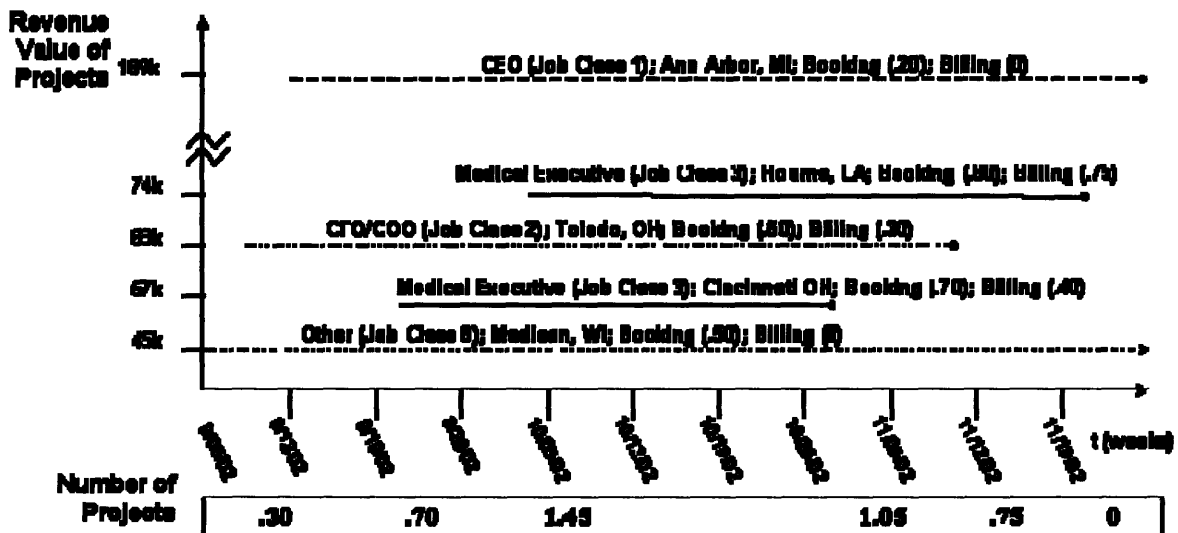


Figure 3. Multitasking Profile of Employee #102 (9/05/2002 – 11/26/2002). A Multitasking Profile displays all of an employee's ongoing projects during a particular period; each project's job class and city; and the level of attention given by the employee to booking (Booking %) and executing (Billing %) each project. The graphic below the profile displays the employee's effort share weighted number of projects over time.

4.3. A Model of Project Duration

To test whether IT, information flows and the level of multitasking are related to the speed with which teams execute projects, we developed a parsimonious model of project completion rate. As the dataset contains right censored data,⁷ ordinary least squares can produce biased and inconsistent results of rate analyses (Tuma & Hannan 1984). We therefore use a hazard rate model of the likelihood of a project completing on a given day, conditional on it not having been completed earlier. We employ a Cox proportional hazards model specification, formalized in equation (2), to estimate the relationships between IT use, information flows and the completion rate of projects:

$$(2) \quad R(t) = r(t)^b e^{\beta X},$$

where $R(t)$ represents the project completion rate, t is project time in the risk set, and $r(t)^b$ the baseline completion rate. The effects of independent variables are specified in the exponential power, where β is a vector of estimated coefficients and X is a vector of independent variables. The coefficients in this model have a straightforward interpretation: β represents the percent increase or decrease in the project completion rate associated with a one unit increase in the independent variable.⁸ Coefficients greater than 1 represent an increase in the project completion rate (equal to $\beta - 1$); coefficients less than 1 represent a decrease (equal to $1 - \beta$).

4.4. Alternate Hypotheses and Control Variables

Based on our interviews, we posit six broad factors that could influence our dependent variables besides the independent variables of interest:

Characteristics of Individual Recruiters. We included controls for traditional demographic and human capital variables (e.g. age, gender, level of education, industry experience and managerial level) to

⁷ This reflects projects that did not complete during the observation window.

⁸ Specification tests reveal no significant duration dependence in our explanatory variables, and the proportional hazards assumption is shown to be valid using both statistical and graphical tests.

Essay 1: Information, Technology & Information Worker Productivity

control for observable differences related to worker education, skill and experience. We also utilize fixed effects specifications to control for unobserved heterogeneity across individual recruiters.

Team Size. Adding more labor to a project may speed up work or slow it down depending on tradeoffs between the added complexity of a larger team and the output contribution of additional labor. We controlled for these effects by including a variable for the size of each team.

Job Type. Certain positions may be easier or harder to fill. Firms might, for instance, demand that a new CEO be named quickly. Senior executives also have more experience with recruiters and with job mobility. To control for the effect of *Job Type*, we include a dummy variable for the eight job classes the firm recognizes in its own records.⁹ We also control for *Task Characteristics*, measured by survey responses about the routineness and interdependence of tasks, for similar reasons.

City Characteristics. Crime rates, weather conditions, the cost of living and other city characteristics may increase or decrease the attractiveness of a position for candidates and may therefore influence contract completion due to placement difficulty. To control for these factors we collected data on the 768 cities in our sample from the web site Sperling's Best Places.¹⁰ Factor analysis revealed four underlying factors with significant results in our models: *cost of living*, *crime rates* (violent crime and property crime per capita), *weather conditions* (number of sunny days per annum) and *commute time*. We therefore included these controls in project-level analyses.¹¹

Revenue Value. The market price of a project contains information about the project's difficulty, value, priority and the market-assessed quality of work. As such, we include the revenue value of projects to control for differences in projects' difficulty, priority, and quality.

⁹ The firm categorizes jobs by the following categories: CEO, COO, CIO, Medical Executive, Human Resources Executive, Business Development Executive and 'Other.' We also ran specifications controlling for sub-categories of 'Other' jobs clustered by their project descriptions, which returned similar results. We therefore retain the firm's original classification.

¹⁰ <http://www.bestplaces.net/>

¹¹ We collected city level data on *tax rates for sales, income and property, the aggregate cost of living, home ownership costs, rate of home appreciation, air quality, water quality, number of superfund sites near the city, physicians per capita, health care costs per capita, violent and property crime per capita, public education expenditures per capita, average student to teacher ratio, an index of ultraviolet radiation levels, risk indices for earthquakes, tornadoes and hurricanes, average number of sunny, cloudy, and rainy days per year, average number of days below freezing per year and average commute time to work.*

Essay 1: Information, Technology & Information Worker Productivity

Temporal Variation. In order to isolate relationships between work process variables, such as multitasking, and output variables, we paid particular attention to the impact of both seasonal and transitory temporal shocks to the relationships. In our data, business exhibits seasonal variation. For instance, business picks up sharply in January and declines steadily through the next eight months. Given this variation, the exogenous shock of increased demand for executive recruiting services could drive increases in both the amount of work employees take on (multitasking), and the revenues they generate. In this case, we could find a spurious correlation between multitasking and revenues driven by an exogenous pulse in demand for the firms' services. There could also be non-seasonal transitory shocks to demand in a given year or a given month of a given year. For this reason, we control for both seasonal and transitory variation in our data with dummy variables for *year*, *month* and *year/month* separately.

5. Statistical Specifications

We tested three specifications of the relationship between revenues, completed projects, multitasking and project duration: Feasible Generalized Least Squares (FGLS) and Fixed Effects specifications at the daily level, and an OLS specification for the year 2002 independently. As daily regressions displayed significant levels of serial correlation based on Durbin-Watson tests and heteroskedasticity based on Breush-Pagan tests, we modeled these data in FGLS specifications using within-panel corrections for both heteroskedasticity and autocorrelation. The error term was modeled with autocorrelation diminishing uniformly over time: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$. We then examined OLS estimates of the relationships between our independent variables and multitasking at the project-level, with controls for job class, temporal variation and a variable indexing right censored project data. Finally, we employed a Maximum Likelihood specification to test the Cox proportional hazards models of project completion. We have reported standard errors according to the White correction (White 1980) for regressions that

Essay 1: Information, Technology & Information Worker Productivity

violated the assumption of no heteroskedasticity at the 5% level. As project outcomes may cluster on groups of project team members, we report robust standard errors clustered by project team.¹²

6. Results

6.1. Drivers of Production

We determined through interviews that, in our setting, the key driver of production is the number of projects completed per unit time. As recruiting teams complete projects, they generate revenue for the firm. Our model of the production process therefore hypothesizes that a key intermediate variable in the ‘black box’ is completed projects, shown in Figure 2: Completed Projects → Revenues.

We tested this hypothesis by examining the relationship between completed projects and revenue generation per person per day over the five year period. The results in Table 3, Model 1 demonstrate strong support for our basic model. The number of completed projects per day is a strong driver of individual information worker revenue generation. The coefficient indicates that the individual worker’s share of the revenue generated from a day’s work on an (eventually) completed project is worth, on average, \$2,149.19 dollars per day for the firm.

We then tested the second fundamental hypothesis of our model: that both revenues and completed projects are driven by the number of projects an individual works on per unit time, and by the length of time it takes to finish projects on average. We examined the relationship between multitasking, average duration, revenues and completed projects in Models 2-5 in Table 3. The results demonstrate that more simultaneous projects and faster completion times (shorter duration) are associated with greater project completion and revenue generation per person per day.¹³

¹² Clustered robust standard errors treat each project team as a super-observation for part of its contribution to the variance estimate (e.g. $\mathcal{E}_{ci} = \eta_c + U_{ci}$, where η_c is a group effect and U_{ci} the idiosyncratic error). They are robust to correlations within the observations of each group, but are never fully efficient. They represent conservative estimates of standard errors that are particularly conservative in our data because team members expend varying levels of effort across projects, such that teams with similar composition have relatively independent share weighted values of team participation.

¹³ As the variables for multitasking and duration are normalized with mean = 0 and s.d. = 1, the coefficients represent the standard deviation variation in the dependent variable associated with a one standard deviation change in the independent variable.

Essay 1: Information, Technology & Information Worker Productivity

Table 3: Panel Data Estimates of the Drivers of Project Completion and Revenue Generation

Variables	Model 1	Model 2	Model 3	Model 4	Model 5
<i>Dependent Variable:</i>	Revenues	nRevenues	nRevenues	nComp. Projects	nComp. Projects
<i>Specification</i>	<i>FGLS</i>	<i>FGLS</i>	<i>Fixed Effects</i>	<i>FGLS</i>	<i>Fixed Effects</i>
	<i>Daily</i>	<i>Daily</i>	<i>Daily</i>	<i>Daily</i>	<i>Daily</i>
nEducation	2.10 (1.62)	.003 (.003)		.001 (.001)	
Gender	-.73 (3.72)	.001 (.006)		-.007** (.003)	
Partner	654.17*** (11.42)	1.041*** (.029)		.418*** (.009)	
Consultant	521.14*** (10.23)	1.014*** (.028)		.350*** (.009)	
Completed Projects	2149.19*** (43.41)				
nMultitasking		.140*** (.010)	.987*** (.008)	.360*** (.007)	.722*** (.005)
nMultitasking Squared		-.089*** (.009)	-.272*** (.006)	-.147*** (.006)	-.146*** (.004)
nDuration		-.174*** (.004)	-.152*** (.003)	-.087*** (.003)	-.133*** (.002)
<i>Time Controls</i>	Month, Year	Month, Year	Month, Year	Month, Year	Month, Year
Log Likelihood	-370966.8	152093.6	-	133227.3	-
X ² (d.f) / F(d.f)	8976.9*** (20)	4045.32*** (22)	6836.79*** (18)	17285.69*** (22)	5035.76*** (18)
Observations	78201	78201	100816	81824	104983

***p<.001; **p<.05; *p<.10. "n"=Normalized Variable; Multitasking terms are effort share weighted.

We also find that the relationship between multitasking and output is non-linear. The coefficient on the multitasking squared term is negative and significant - implying a concave relationship, such that more multitasking is associated with greater revenue generation and project output to a point, after which there are diminishing marginal returns, then negative returns to increased multitasking. We considered four possible explanations for this inverted-U shaped relationship, and let the data speak to which is the most likely.

Explanation 1: A Fundamental Tradeoff between Workload and Efficiency. Perhaps the most intuitive explanation is that a fundamental trade-off exists between workload and efficiency, such that multitasking beyond a certain point reduces productivity. This explanation fits with empirical evidence on the cognitive costs of multitasking. Multitasking behavior has been associated with cognitive switching costs that reduce task completion rates and increase task error rates in experimental settings (e.g.

Essay 1: Information, Technology & Information Worker Productivity

Rubenstein et. al. 2001). When employees juggle too many simultaneous projects, work gets backed up and productivity suffers. The situation is analogous to congestion and throughput processes for queued activities. For example, the throughput of cars on a highway increases as more cars join traffic, but is reduced by congestion after a certain level of traffic is exceeded. Our interviews corroborate this story. As the CIO of the firm put it: “Everyone can only deal with so many balls in the air. When someone gets ‘too far in,’ they lose touch. They can’t tell one project from another.”

Explanation 2: Project Portfolios Differ by Employee Type. Correlated differences between individual workers and their project portfolios could also be driving the inverted-U shaped relationship between multitasking and output. For example, it could be that new, inexperienced workers take on fewer and less valuable projects, while the most experienced consultants take on the largest number. These two clusters of project portfolios would explain the first and last third of the inverted-U. Filling out the graph, partners may reserve the most important and valuable projects for themselves and work on fewer projects than consultants. Explanation 2 is consistent with several theoretical perspectives. Partners’ social and organizational power (e.g. Pfeffer 1981) could enable them to take on a relatively small number of high revenue value projects, creating a relationship between leisure (less multitasking) and revenues in the partner strata of our data. This explanation is also consistent with incentive theories of deferred compensation, where workers are underpaid during the early part of their careers (e.g. $[\text{Pay}=\text{f}(\text{revenues})] < \text{marginal revenue product}$) and paid more than their marginal revenue product later on (Lazear 1979).

Explanation 3: Unobserved Drivers of Multitasking and Output. There could also be unobservable drivers of both multitasking and output that create the inverted-U shaped relationship. For instance, the most productive workers could also spend time on other tasks we don’t observe (like networking) that drive them to work on fewer projects simultaneously while producing more output. If these highly productive workers worked on slightly more projects than inexperienced new workers, but fewer projects than experienced workers who did not spend time on these unobserved tasks, an inverted-U shaped relationship between multitasking and output could be observed.

Essay 1: Information, Technology & Information Worker Productivity

Explanation 4: Exogenous Temporal Variation. Clients may hire top management teams in groups, creating temporal clusters of contracts that are both few in number and high in revenue value. If this type of turnover happens seasonally – for example, near the beginning or end of the fiscal year – then temporal clusters of fewer high revenue value projects could create the inverted U-shaped relationship. Exogenous transitory shocks to client demand could also inspire ramping up of production, or large simultaneous layoffs in low revenue value positions. Given the right structure, it is possible that these temporal clusters could drive the inverted-U shaped relationship between multitasking and output.

Reconciling Explanations. While explanations 2-4 conform to theory and could explain the slope of this relationship at different levels of multitasking, our specifications suggest they are unlikely. In FGLS specifications, our controls for managerial level and industry experience go a long way toward holding constant variation driven by status, organizational power or career tenure. In addition, our estimates of the relationship between multitasking and output are robust to specifications controlling for unobserved heterogeneity across individuals, accounting for aspects of social and organizational power not captured by organizational level and tenure, for unobservable practices (e.g. networking) of highly productive workers, and for other characteristics of individual recruiters which could contribute to the shape of the relationship between multitasking and output. In addition, our controls for temporal variation (both seasonal variation and exogenous shocks to demand) discount explanations based on temporal clusters of projects of different types. As our quantitative and qualitative data discount explanations 2-4, we are drawn to interpret the results in Table 3 as evidence supporting explanation 1: that a fundamental tradeoff exists between workload and efficiency.¹⁴

6.2. Relationships between IT, Information Flows and Multitasking

To test whether IT use and skill, and properties of the flow of information in workers' email traffic are related to the intermediate output variables shown to drive production, we first tested the

¹⁴ Since we have not controlled for all possible sources of endogeneity or identified equilibrium values of multitasking and output, the optimal levels of multitasking implied by our parameter estimates may not be precise optima in equilibrium.

Essay 1: Information, Technology & Information Worker Productivity

relationship between our IT and information flow variables and project-level multitasking. Our analysis included controls for team characteristics and job class, but not for city characteristics, which are potentially salient for project duration but should not influence how many projects teams work on.^{15, 16}

Table 4: OLS Analysis of the Impact of IT on Multitasking at the Project-level

<i>Dependent Variable</i>	<i>Multitasking</i>		
	Model 1	Model 2	Model 3
	OLS-c	OLS-c	OLS-c
Team Size	.227** (.090)	.285** (.110)	.207* (.115)
Education	.102** (.052)	.077 (.055)	.101* (.051)
Industry Experience	-.002 (.007)	.001 (.006)	-.002 (.007)
nF2F Contacts	.030 (.036)		.038 (.036)
nPhone Contacts	-.224** (.090)		-.229** (.088)
nEmail Contacts	.320*** (.091)		.305*** (.093)
nESS Skill		.036 (.078)	-.029 (.081)
nESS Use		.114* (.064)	.061 (.061)
Constant	-1.98* (1.121)	-1.65 (1.146)	-1.90* (1.114)
Job Class Controls	YES	YES	YES
Time Controls	Year	Year	Year
Censor Dummy	YES	YES	YES
F Value	8.49***	7.90***	7.88***
(d.f)	(19)	(18)	(21)
R ²	.16	.12	.16
Obs.	1372	1372	1372

***p<.001; **p<.05; *p<.10. OLS-c = Robust Clustered Standard Errors (n = 505 Clusters),
“n” = Normalized Variable

The coefficients in Table 4, Models 1 and 3 demonstrate that teams whose members were heavy multitaskers communicated with more people over email, and significantly fewer people over the phone. Since the variables have been normalized, they can be interpreted as follows: a one standard deviation increase in the number of email contacts is associated on average with a .30 standard deviation increase in the number of simultaneous projects the team is working on during the focal project (see Model 3). We

¹⁵ We also ran the same analysis controlling for the revenue value of the project, with no qualitative change in the coefficients.

¹⁶ The models include a dummy variable for whether the project was right censored during the observation window.

Essay 1: Information, Technology & Information Worker Productivity

also see from the coefficient in Model 2 that teams who use the ESS system more to gather information work on more projects simultaneously.¹⁷ As synchronous technology (i.e. telephone) reduces multitasking while asynchronous technology (i.e. email, and to a lesser extent, ESS) supports multitasking, a manager seeking to juggle more projects might favor information access patterns that do not require coordinated scheduling.

Table 5: OLS Analysis of Relationship Between Information Flows & nMultitasking at the Project-level (“n” = Normalized Variable)

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
<i>Dependent Variable</i>	<i>Multitasking</i>					
	OLS-c	OLS-c	OLS-c	OLS-c	OLS-c	OLS-c
<i>Controls</i>						
Team Size	.223*** (.069)	.278*** (.074)	.221** (.071)	.226*** (.068)	.211** (.076)	.210** (.066)
Yrs Education	.090* (.046)	.085 (.051)	.079 (.048)	.091** (.044)	.084* (.048)	.087** (.042)
Industry Experience	.006 (.007)	-.002 (.006)	.006 (.007)	.007 (.006)	.003 (.006)	.003 (.005)
<i>Information Flow Level & Structure</i>						
nTotal Emails	.320*** (.067)					.266** (.104)
nNetwork Size		.304*** (.068)				
nIn Degree			.301*** (.068)			
nOut Degree				.361*** (.060)		
nBetweenness					.307*** (.060)	.126* (.076)
nConstraint (1-Structural Holes)					-.134* (.071)	-.181** (.090)
Constant	-2.128** (1.035)	-2.125* (1.093)	-1.870* (1.056)	-2.230** (.995)	-1.79* (1.082)	-.196* (1.002)
Job Class Controls?	YES	YES	YES	YES	YES	YES
Temporal Controls:	Year	Year	Year	Year	Year	Year
Censor Dummy?	YES	YES	YES	YES	YES	YES
F Value (d.f)	9.63*** (17)	8.73*** (17)	9.29*** (17)	10.60*** (17)	9.97*** (18)	10.03*** (19)
R ²	.20	.18	.19	.23	.20	.24
Observations	1372	1372	1372	1372	1372	1372
***p<.001; **p<.05; *p<.10. OLS-c = Robust Clustered Standard Errors (n = 505 Clusters)						

¹⁷ The coefficient on ESS Skill is positive and significant when entered alone, but not when controlling for ESS Use. ESS Use is significant at $p < .001$ in the full model (Model 3) when standard errors are robust but not clustered by project team.

Essay 1: Information, Technology & Information Worker Productivity

We also tested the analogous relationships between workers' email traffic and their amount of multitasking. The results for both the levels and structure of information flows in teams' email are reported in Table 5, Models 1-6.^{18 19} All four measures of communication levels demonstrate strongly that heavy multitaskers communicate more over email. These results strengthen and extend the result from the survey measure of email contacts reported in Table 4. When considering the structural properties of worker's email traffic, more multitasking is associated with greater betweenness centrality – a proxy for the probability of being privy to a given piece of information flowing through the communication network of the firm. Heavy multitaskers are in the 'thick' of the flow of information and are likely to be 'in between' a larger number of pairs of other employees in terms of their communication structure. In day-to-day terms, it pays to be a communications middleman. Peripheral individuals, with lower information flows, show fewer projects per unit time. Similarly, employees with 'redundant contacts' multitask less. The negative coefficient on the constraint variable shows that those entangled in closed networks (networks whose members are all closely tied together) work on fewer projects simultaneously. To untangle constrained social networks, organizations can diversify team assignments and job rotations. Both the level and the structure of information flows correspond strongly with multitasking behavior and structural parameters remain significant even when controlling for total email volume in Model 6.²⁰

These results demonstrate a strong correspondence between multitasking and the structure and level of email traffic. However, unobserved characteristics of project assignment may simultaneously drive multitasking and IT use. For example, it could be that multitasking is used more for simpler projects that are more readily accomplished via email. If so we may observe a correlation between multitasking and email use due to the nature of project assignment. To address these concerns, we examined the most

¹⁸ The models include a dummy variable for whether the project was right censored during the observation window. Estimates of information flow variables using non-clustered standard errors are all significant at $p < .001$.

¹⁹ Variables that are highly collinear are entered separately into the regressions.

²⁰ It seems intuitive that employees working on more projects at once need to be aware of more lines of communication and information, and thus appear in these structural positions. However, we cannot make causal claims about these results. Heavy multitaskers may seek more information and position themselves in the thick of information flows, or highly central employees may be chosen to conduct more tasks, or may chose to conduct more tasks on their own. Nevertheless, information flows are associated with the multitasking behavior of information workers in our data.

Essay 1: Information, Technology & Information Worker Productivity

likely sources of endogeneity in detail. We found that although simpler, lower revenue projects exhibited more multitasking (Revenue: $\beta = - 644$; $t = 2.27$), project revenue was associated with less total email ($\beta = - 4288.35$; $t = 2.61$) and with less database use ($\beta = - .112$; $t = 3.36$), discounting the possibility that simpler projects simultaneously drive more multitasking and more email and database use.²¹ We also find email and database use are associated with greater multitasking when controlling for project type and revenue value (see Table 4). Although project assignment may be non-random in our setting, it does not explain relationships between IT use, email and multitasking.

6.3. Relationships between IT, Information Flows, Multitasking and Project Duration

To test the relationships between multitasking, IT, information flows and project duration, we estimated the hazard rate model of project completion time. Our specification tests the relationship between explanatory variables and projects' instantaneous transition rate – a measure of the likelihood of project completion at time t , conditional on the project not having completed before t . Table 6 shows the analysis of the relationship between IT, multitasking and project completion rates, controlling for *job type*, *task characteristics*, and *city characteristics*.

Multitasking is strongly associated with slower completion rates. Teams with a one standard deviation increase in project-level multitasking (approximately 2.8 additional projects) complete about 15% fewer projects per month. These results corroborate our interpretation of the drivers of the inverted-U shaped relationship between multitasking and output. Teams that multitask more take longer to finish projects – a result consistent with a loss of efficiency at high levels of multitasking. Holding the level of multitasking constant, teams using the ESS to gather more information (S.D.= 15%) complete projects on average 11% faster, and teams that use the phone more also execute projects faster. These results are a

²¹ While projects in the medical field do exhibit more multitasking, total email use and the number of email contacts is fairly constant across project categories. This discounts the hypothesis that project assignment simultaneously drives IT use and multitasking, and is supported by analyses of IT use and multitasking that control for project type and revenue value.

Essay 1: Information, Technology & Information Worker Productivity

departure from our simple model of the impact of IT on project duration which did not control for the level of multitasking, indicating the analytical value of our more comprehensive model.²²

Table 6: Hazard Rate Analysis of the Impact of IT and Multitasking on Project Completion Rate (n=NormVar)

<i>Dependent Variable</i>	<i>Project Duration</i>		
	Model 1	Model 2	Model 3
Variables	RSE-c	RSE-c	RSE-c
<i>Controls</i>			
Team Size	.854** (.067)	.820** (.067)	.842** (.071)
Industry Experience	.989** (.004)	.990** (.005)	.991** (.005)
nCost of Living	.924* (.044)	.933 (.044)	.926 (.045)
nCrime Per Capita	.410** (.129)	.404** (.125)	.389** (.122)
nSunny Days	1.083** (.043)	1.058 (.043)	1.064 (.043)
nCommute Time	.953 (.032)	.962 (.033)	.959 (.033)
nRoutiness	1.003 (.042)	1.074* (.043)	1.042 (.047)
Interdepend.	.980 (.038)	.957 (.034)	.979 (.041)
nMultitasking	.858*** (.030)	.843*** (.029)	.851*** (.030)
<i>IT Variables</i>			
nFTF Contacts	1.002 (.027)		1.015 (.029)
nPhone Contacts	1.109* (.061)		1.073 (.062)
nEmail Contacts	.974 (.043)		.958 (.046)
nESS Use		1.118*** (.035)	1.114** (.038)
nESS Skill		.947 (.051)	.949 (.060)
Job Class Controls?	YES	YES	YES
Log Likelihood	-7080.3	-7077.6	-7076.03
X ² (d.f)	185.07*** (19)	193.16*** (18)	196.79*** (21)
Obs.	1180	1180	1180

***p<.001; **p<.05; *p<.10. RSE-c = Robust Clustered SE (n = 505 Clusters)

²² Team size and industry experience are associated with longer project duration and slower completion rates. Teams with more members may take longer to execute projects due to the added complexity of coordination, or firm may resort to 'throwing more labor at' difficult jobs or jobs that are taking longer to complete than expected. Controlling for team size therefore may also account for differences in project difficulty not picked up by controls for job type, task, and city characteristics. Industry experience also corresponds to longer project duration perhaps because less experienced employees receive less demanding work. Cost of living, crime rates, and greater commute times all reduce the project completion rate on average, meaning these characteristics may be less attractive to potential candidates, while good weather seems to boost the completion rate. Routine tasks consistently finish faster, and greater interdependence among team members is associated with slower completion rates.

Essay 1: Information, Technology & Information Worker Productivity

Our analyses of the impact of multitasking and IT use on the speed of work demonstrate two key findings: First, multitasking slows work, explaining a possible mechanism driving the inverted-U shaped relationship between multitasking and output. Second, IT use shifts production out, increasing output at all levels of multitasking by enabling greater workloads without a corresponding loss of efficiency.²³

We also analyzed how information flows and multitasking correspond to project completion rates.²⁴ All control variables and the multitasking variable display significant results of almost identical magnitude as reported in Table 6. However, none of the six information structure or flow variables (total emails, network size, in-degree, out-degree, betweenness and constraint) returned a significant parameter estimate.²⁵ While the levels and structure of information flows predict the level of multitasking, they do not predict the speed with which projects are completed, controlling for multitasking. This contrasts with Hansen (2002), which did not control for multitasking.

The strong positive coefficient on ESS use, together with survey and interview data, provides useful managerial insight. Although ESS use speeds projects by 11%, comfort with and ability to use these tools decline with age (Spearman's $\rho = -.47$, $p < .001$, Spearman's $\rho = -.31$, $p < .02$). This suggests that targeted training in ESS use could speed project completions at the firm. Overall, as multitasking reduces the per project completion rate, productive information workers trade longer task duration per project for more tasks per unit time by working on multiple projects in parallel. Productive executives then offset multitasking delay costs by using information technology more heavily.

7. Discussion and Conclusion

²³ We found no convincing evidence of any interaction effect of multitasking and IT use on the project completion rate in separate analyses, indicating that IT use enables greater project completion per unit time at all levels of multitasking.

²⁴ Detailed results on "Hazard Rate Analysis of the Impact of Information Flows and Multitasking on Project Completion Rate" are omitted due to space constraints, but are available upon request from the authors.

²⁵ When we remove city controls, network variables predict faster project completions (total email $\beta = 1.051$; $p < .10$, network size $\beta = 1.068$; $p < .05$, in-degree $\beta = 1.055$; $p < .10$, out-degree $\beta = 1.061$; $p < .10$), suggesting interdependence between geographic distribution and social network attributes – a result we intend to explore in future research.

Essay 1: Information, Technology & Information Worker Productivity

To date, important advances in assessing IT value have used more sophisticated econometric methods or more comprehensive firm-level and plant-level data. In contrast, our research seeks to open two new frontiers: (1) detailed task-level evidence of information worker output, and (2) objective measures of information flows through social networks. This approach provides a higher resolution microscope with which to study organizational phenomena, revealing finer grained relationships than would be possible with any amount of firm, industry, or country-level data.

Three contributions result from this approach. First, we show that information work can, in fact, be measured. We identified a context with objective performance metrics, built tools to directly observe behaviors and information flows in email, and gathered independent data on project quality controls. Our analyses of these data produce precise estimates of the productivity of information workers. While information work has often defied measurement in the past, we found it remarkably quantifiable in this setting.

Second, we build and validate multitasking and hazard rate models of project completions at both individual and team levels. These models highlight intermediate production processes and directly explore the association between using technology, juggling more tasks, and the ability to complete tasks faster. In effect, we used better data to reveal the production function of information workers. We find that individual differences in IT use behaviors correspond with differences in performance. On average, workers using more asynchronous email and database tools handle substantially more projects simultaneously. In contrast, traditional synchronous communication modes such as phone calls correlate with less multitasking. Further, there were speed implications. People who multitasked heavily benefited from also using the ESS heavily to speed their work, enabling them to complete more projects per unit time. These results, together with the survey data, imply that targeted ESS training could improve speed and thus firm performance.

Finally, and perhaps most interestingly, when we apply social network analysis to our email data, we find that position and flow are critically important. Betweenness centrality shows a positive association with ability to multitask, as do in-degree, out-degree, and network size. Among information

Essay 1: Information, Technology & Information Worker Productivity

workers, it pays to be a communications middleman. Peripheral employees, outside the communication flow, work on fewer projects over time. The total volume of communication is also statistically significant as is the measure of constraint, demonstrating that constrained networks and redundant contacts correspond to less multitasking. An implication of these results for managers is that untangling social networks through strategic job rotation could lead to more efficient multitasking. Strikingly, we also find that richer information flows alone do not necessarily increase the speed with which individuals complete their projects. Central information brokers boost their productivity by multitasking more effectively rather than by working faster.

In sum, we find a substantial correspondence among information, technology, and output in this setting. It is not just having IT but how one uses it that predicts differences in performance. Tools and techniques developed during this research can be readily applied to other project-level information work involving email and databases including sales, consulting, law, medicine, software development, venture capital, banking, insurance, and architecture, among others. This portends a substantial improvement in our understanding of the relationship between information, technology, and value creation, and reveals important managerial implications related to organizational structure, team assignment, job rotation, IT use and training, and the management of organizational communication.

Acknowledgments

We are grateful to the National Science Foundation (Career Award IIS-9876233 and grant IIS-0085725), Intel Corporation, the Marvin Bower Fellowship, and the MIT Center for Digital Business for generous funding. We thank Abraham Evans-El, Jia Fazio, Saba Gul, Davy Kim, Jennifer Kwon and Jun Zhang for their remarkable and tireless research assistance, and Julie Hilden, and seminar participants at NYU, MIT, Georgia Tech, the Workshop on Information Systems and Economics for valuable comments.

Essay 1: Information, Technology & Information Worker Productivity

References

- Apte, U., Nath, H. 2004. "Size, structure and growth of the U.S. economy." Center for Management in the Information Economy, Business and Information Technologies Project (BIT) Working Paper.
- Arrow, K.J. 1962. "The Economic Implications of Learning by Doing." *Rev. Econ. Stud.* (29:3): 155-173.
- Barua, A., C. H. Kriebel, Mukhopadhyay, T. 1995. "Information technology and business value: An analytical and empirical investigation." *Information Systems Research* (6:1), March: 2-23.
- Bernard, H.R., Killworth, P., & Sailor, L. 1981. "Summary of research on informant accuracy in network data and the reverse small world problem." *Connections*, (4:2): 11-25.
- Bharadwaj, A. S., S. G. Bharadwaj, Konsynski, B.R. 1999. "Information technology effects on firm performance as measured by Tobin's q." *Management Science* (45:7): 1008-1024.
- Breusch, T., Pagan, A. 1979. "A simple test for heteroscedasticity and random coefficient variation." *Econometrica* (47:5), September, 1287-1294.
- Brynjolfsson, E., Hitt, L. 1996. "Paradox lost? Firm-level evidence on the returns to information systems spending." *Management Science* (42:4), April, 541-558.
- Brynjolfsson, E., Hitt, L. 2000. "Beyond computation: Information technology, organizational transformation and business performance." *Journal of Economic Perspectives* (14:4), Fall, 23-48.
- Bulkley, N., Van Alstyne, M. 2005. "Why Information Should Influence Productivity" in *Network Society: A Cross Cultural Perspective*, M. Castells (Ed.), Edward Elgar, Northampton, MA, 145-173.
- Burt, R. 1992. *Structural Holes: The Social Structure of Competition*. Harvard Press, Cambridge, MA.
- Coleman, J.S. 1988. "Social Capital in the Creation of Human Capital" *Amer. Jrnl. Soc.*, (94): S95-S120.
- Cyert, R.M., March, J.G. 1963. *A Behavioral Theory of the Firm*, Malden, MA, Blackwell Publishers.
- David, P.A. 1990. "The Dynamo and the Computer: A Historical Perspective on the Modern Productivity Paradox." *Amer. Econ. Rev. Papers and Proceedings* (80:2), May, pp. 355-361.
- Dewan, S., Kraemer, K. 2000. "Information technology and productivity: evidence from country-level data." *Management Science* (46:4), April, 548-562.
- Freeman, L. 1979. "Centrality in social networks: Conceptual clarification." *Soc. Networks* (1:3): 215-34.
- Galbraith, J.R. 1973. *Designing Complex Organizations*. Reading, MA, Addison-Wesley.
- Gargiulo, M., Benassi, M. 2000. "Trapped in your own net? Network cohesion, structural holes, and the adaptation of social capital." *Organization Science* (11:2), March-April, 183-196.
- Granovetter, M. 1973. "The strength of weak ties." *American Journal of Sociology* (78:6): 1360-1380.
- Hansen, M. 1999. "The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits." *Admin. Sci. Quart.* (44:1), March, 82-111.
- Hansen, M. 2002. "Knowledge networks: Explaining effective knowledge sharing in multiunit companies." *Organization Science* (13:3), May/June, 232-248.
- Hinds, P.J., Kiesler, S. 2002. *Distributed Work*. Cambridge, MA. MIT Press.
- Ichniowski, C., K. Shaw, Prennushi, G. 1997. "The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines." *Amer. Econ. Rev.* (87:3): 291-313.
- Jorgenson D., Stiroh, K. 2000. "US Economic Growth at the Industry Level." *Amer. Econ. Rev.* (90:2): 161-7.

Essay 1: Information, Technology & Information Worker Productivity

- Lazear, E. 1979. "Why is there mandatory retirement?" *J. Political Economy* (87:6), 1261-1264.
- McAfee, A. 2002. "The impact of enterprise technology adoption on operational performance: An empirical investigation." *Production and Operations Management Journal* (11:1), Spring, 33-53.
- Mukhopadhyay, T., R. Surendra, Srinivasan, K. 1997. "Information Technology Impact on Process Output and Quality." *Management Science* (43:12), December, 1645-1659.
- Pfeffer, J. 1981. *Power in Organizations*, Pitman, Boston.
- Podolny, J., Baron, J. 1997. "Resources and relationships: Social networks and mobility in the workplace." *American Sociological Review* (62:5), October, 673-693.
- Reagans, R., Zuckerman, E. 2001. "Networks, diversity, and productivity: The social capital of corporate R&D teams." *Organization Science* (12:4), July-August, 502-517.
- Rubinstein, J., D. Meyer, Evans, E. 2001. "Executive Control of Cognitive Processes in Task Switching." *Journal of Experimental Psychology: Human Perception and Performance* (27:4): 763-797.
- Stiglitz, J. 2000. "The Contributions of the Economics of Information to Twentieth Century Economics." *Quarterly Journal of Economics* (115:4), November, 1441-1478.
- Szulanski, G. 1996. "Exploring internal stickiness: Impediments to the transfer of best practice within the firm." *Strategic Management Journal* (17), Winter, 27-43.
- Tuma, N.B., Hannan, M.T. 1984. *Social Dynamics: Models and Methods*. Academic Press, New York.
- Van Alstyne, M. Zhang J. 2003. "EmailNet: A System for Automatically Mining Social Networks from Organizational Email Communication," NAACSOS.
- von Hippel, E. 1998. "Economics of Product Development by Users: The Impact of "Sticky" Local Information" *Management Science* (44:5), May, 629-644.
- White, H. 1980. "A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity." *Econometrica* (48:4), May, 817-838.

Appendix A: Description of Variables and Data Sources

Variable	Source	Description
<i>Project Team Variables</i>		
Team Size	Accounting	Number of team members assigned to project.
Age	Survey	Age of employees, Average age of team members.
Yrs Education	Survey	Years of education, Average years of education of team members.
Industry Experience	Survey	Years of industry experience, Average years of industry experience of team members.
Multitasking	Accounting	Share weighted number of simultaneous projects, weighted by the billing % assignment of each employee to each project.
Project Duration (Days)	Accounting	Days from project start to project end.
Project Revenue Value (\$)	Accounting	Revenue value of project.
Team Interdependence	Survey	1-7: "My job tasks are highly interdependent with other people's tasks. I must often coordinate with other team members."
Task Routiness	Survey	1-7: "My data requirements are highly routine. I could specify all I need on standard forms."
F2F Contacts	Survey	How many people do you communicate with on a typical day face to face?
Phone Contacts	Survey	How many people do you communicate with on a typical day by phone?
Email Contacts	Survey	How many people do you communicate with on a typical day by email?
ESS (Database) Skill	Survey	Combined: (1-7) "I am highly effective at using our in-house proprietary search tools. This means I know what information they contain and can easily find, add, and modify the records I need." & "I have control over the information I use; I can access and modify it at will."
ESS (Database) Use (%)	Survey	"What proportion of your time do you spend gathering information from the internal database and external proprietary databases?"
Degree Centrality	Email	Number of ties to others. Row or column sums of adjacency matrix.
In Degree	Email	Number of incoming ties to others. Column sums of adjacency matrix.
Out Degree	Email	Number of outgoing ties to others. Row sums of adjacency matrix.
Network Size	Email	Number of unique contacts linked to ego email, plus ego.
Betweenness	Email	The percentage of all geodesic paths from neighbor to neighbor that pass through ego.
Constraint	Email	Measures the extent to which ego's connections are to others who are connected to one another. (Burt 1992: 55)
<i>City Characteristics</i>		
Cost of Living	City Data	The average cost of living in the following categories weighted as follows: Housing (30%), food and groceries (15%), transportation (10%), utilities (6%), health care (7%), miscellaneous – clothing, services and entertainment (32%). State and local taxes not included.
Crime per Capita	City Data	Violent and property crime per capita.
Sunny Days per Annum	City Data	Average number of sunny days per year.
Commute Time	City Data	Average number of minutes to work one way.

Essay 2: “Network Structure & Information Advantage: Structural Determinants of Access to Novel Information and their Performance Implications”

Abstract

We examine relationships between social network structure, information structure, and individual performance. Specifically, we investigate which network structures influence access to diverse and novel information, and whether these relationships explain performance in information intensive work. We build and validate an analytical model of information diversity, develop hypotheses linking two key aspects of network structure - size and diversity - to the distribution of novel information among actors, and test our theory using empirical evidence from a ten month panel of email communication patterns, message content and performance among employees of a medium sized executive recruiting firm. Our results indicate that: (1) the total amount of novel information and the diversity of information flowing to actors are increasing in their network size and network diversity. However, (2) the marginal increase in information diversity is decreasing in actors' network size, a result explained in part because (3) network diversity is increasing in network size, but with diminishing marginal returns. (4) Network diversity contributes to performance even when controlling for the positive performance effects of access to novel information, suggesting additional benefits to network diversity beyond those conferred through information advantage. Surprisingly, (5) traditional demographic and human capital variables have little effect on access to diverse information, highlighting the importance of network structure for information advantage. The methods and tools developed are replicable and can be readily applied to other settings in which email is widely used and available, opening a new frontier for the analysis of networks and information content.

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

1. Introduction

A growing body of evidence links the structural properties of individuals' and groups' networked relationships to various dimensions of economic performance. However, the mechanisms driving this linkage, thought to be related to the value of the information flowing between connected actors, are typically inferred, and rarely empirically demonstrated. As a consequence, our understanding of how and why social structure explains economic outcomes remains underdeveloped, and competing explanations of the causal mechanisms underlying structural advantage proliferate. For instance, we know little about the relative importance of information and control benefits to social structure, and numerous puzzles remain concerning the situational importance of network cohesion and brokerage (Burt 1992, Coleman 1988), and the tradeoffs between the knowledge and power benefits derived from network structures (Reagans & Zuckerman 2006). At the heart of these puzzles lie foundational questions about the degree to which social structure creates intermediate information benefits, and how different network topologies enable these benefits. Comprehensive theories of the structure-performance relationship require a more thorough examination of the intermediate mechanisms through which social structure affects economic advantage. The strategy of this paper is to narrowly examine one of these mechanisms – the relationship between network structure and information benefits – in detail.

One of the most prominent mechanisms theorized to drive the relationship between social structure and performance is the existence of 'information benefits' to network structure. According to this line of argument, actors in favorable structural positions enjoy social and economic advantages based on their access to specific types of information. Burt (1992) convincingly shows that individuals with structurally diverse networks (networks low in (a) cohesion, and (b) structural equivalence) are more successful in terms of wages, promotion, job placement, and creativity (Burt 2004a). He argues that these performance differentials can be explained in part by actors' access to diverse pools of knowledge, and

Essay 2: Network Structure & Information Advantage

their ability to efficiently gather non-redundant information.²⁶ Aral, Brynjolfsson and Van Alstyne (2006) demonstrate that structural diversity is associated with higher levels of economic productivity for task-based information workers. These studies, and numerous others, infer that network diversity is associated with performance in part because diverse contacts provide access to novel information. Novel information is thought to be valuable due to its local scarcity. Actors with scarce, novel information in a given network neighborhood are better positioned to broker opportunities, use information as a commodity, or apply information to problems that are intractable given local knowledge.

While theories of the value of information and empirical evidence on the relationship between network structure and performance exist, little theory, and almost no empirical evidence addresses how network structure influences the nature of the information distributed across a network - the network's 'information structure.'²⁷ To build theory relating network structure to information structure we explore characteristics of the information that accrues to actors in different network positions. We investigate how topological properties of individuals' network positions impact the abundance and diversity of the information they receive and distribute, and ask: Which network structures influence access to diverse, novel information? Do actors with larger networks access more novel information? Are individuals with structurally diverse networks privy to more diverse information? Do information benefits explain performance?

To address these questions, we build hypotheses linking two key aspects of network structure – network size and network diversity – to the distribution of novel information among actors, and to actors' subsequent performance. We test the implications of our theory using empirical evidence from a ten month panel of email communication patterns and message content among information workers in a

²⁶ Coleman's (1988) argument, that focused information from cohesive networks provides more precise signals of actors' environments, also assumes that cohesive networks provide focused (while diverse networks provide diverse) information.

²⁷ The term 'information structure' is used in the economics literature to denote a mapping from the true state of an environment to the signals an actor receives about that environment as a function of the "information gathering methods" used. Information structure combined with decision rules for action are referred to by economists as "organizational form" (see Marschak and Radner 1972 for a thorough review). We define information structure more narrowly in § 2.2.

Essay 2: Network Structure & Information Advantage

medium sized executive recruiting firm. We build and validate an analytical model to measure the diversity of information flowing to and from employees in email communication, and test whether information diversity can be predicted by individuals' network size and network diversity. Our hypotheses address the relationship between network structure and information access, and the tradeoffs between network size and network diversity in generating information advantage. In particular, we argue that while they should both be associated with access to more diverse and novel information, network size and network diversity should also constrain each other in bounded organizational networks.

Our findings indicate that: (1) the total amount of novel information and the diversity of information flowing to actors are increasing in actors' network size and network diversity, while (2) the marginal increase in information diversity is decreasing in network size. We also find evidence of a fundamental tradeoff between network size and network diversity. Part of the explanation for the decreasing marginal contribution of network size to information diversity is that (3) network diversity is increasing in network size, but with diminishing marginal returns. As actors establish relationships with a finite set of possible contacts in an organization, the probability that a marginal relationship will be non-redundant, and provide access to novel information, decreases as possible alters in the network are exhausted. We also find that (4) network diversity contributes to performance even when controlling for the positive performance effects of access to novel information, suggesting additional benefits to network diversity beyond those conferred through information advantage. Surprisingly, (5) traditional demographic and human capital variables (e.g. age, gender, industry experience, education) have little effect on access to diverse information, highlighting the importance of network structure for information advantage.

Our results represent some of the first empirical evidence on the relationship between network structure and information content and our methods for analyzing network structure and information content in email data can be replicated in other settings, opening a new line of inquiry into the relationship between network structure and information advantage.

2. Theory

2.1. Network Structure & Information Advantage: A Critical Inference

The assumption that network structure influences the distribution of information and knowledge underpins a significant amount of theory linking social structure to economic performance. Networks are thought to drive performance in part by influencing characteristics of the information to which individuals have access. For example, Granovetter (1973) argues that topological properties of friendship networks, constrained by basic norms of social interaction, empower weak friendship ties to deliver information about socially distant opportunities more effectively than strong ties. He posits that contacts maintained through weak ties typically “move in circles different from our own and thus have access to information different from that which we receive... [and are therefore]... the channels through which ideas, influence, or information socially distant from ego may reach him” (Granovetter 1973: 1371). Building on this theory, Burt (1992) argues that networks rich in structural diversity confer “information benefits,” by providing access to diverse perspectives, ideas and information. As information in local network neighborhoods tends to be redundant, structurally diverse contacts provide channels through which novel information flows to individuals from distinct pools of social activity. Redundant information is less valuable because many actors are aware of it at the same time, reducing opportunities associated with its use. Structural redundancy is also inefficient from the perspective of information benefits because actors incur costs to maintain redundant contacts while receiving no new information from them (Burt 1992).

In contrast, exposure to diverse ideas, perspectives, and solutions is thought to enable information arbitrage, the creation new innovations, and access to economic opportunities. Empirical evidence on the importance of knowledge brokerage to innovation provides an important example. Network studies of innovation demonstrate that individuals in brokerage positions spanning diverse pools of knowledge are more successful innovators. Hargadon and Sutton (1997) describe how engineers working at the design firm IDEO use their structural positions between diverse engineering and scientific disciplines to broker the flow of information and knowledge from unconnected industrial sectors, creating locally novel design solutions. As Burt (2004b) puts it, “creativity is an import-export game,... not a creation game.”

Essay 2: Network Structure & Information Advantage

The economic value of information in a network derives from its uneven distribution across actors. Value resides in pockets of distinct and diverse pools of information and expertise found in local network neighborhoods. Actors with access to these diverse pools “benefit from disparities in the level and value of particular knowledge held by different groups...” (Hargadon & Sutton 1997: 717), and one of the key mechanisms through which network structures are theorized to improve performance is through access to novel, non-redundant information (Burt 1992).

While the argument that network structures influence performance through their effect on the distribution of information is intuitive and appealing, empirical evidence and detailed theory on this intermediate mechanism remain quite limited. The vast majority of empirical work on networks and information advantage remains ‘content agnostic’ (Hansen 1999: 83), and infers the relationship between network structure and information structure from evidence of a link between networks and performance (e.g. Gargiulo and Benassi 2000, Sparrowe et al. 2001, Cummings & Cross 2003). For example, Reagans & Zuckerman (2001) infer that productivity gains from the external networks of corporate R&D teams are due in part to “information benefits,” “a broader array of ideas and opportunities,” and access to “different skills, information and experience.” Burt (1992, 2004a) also makes this empirical leap, inferring that the observed co-variation of wages, promotion, job placement, and creativity with network diversity is due in part to access to diverse and novel information. Others equate network content with the social function of relationships. For example, Burt (2000: 45) refers to “network content” as “the substance of relationships, qualities defined by distinctions such as friendship versus business versus authority.” In one of the first studies to explore this type of network content, Podolny & Baron (1997) showed that while cohesive ties are beneficial in ‘buy-in’ networks and for those contacts that have control over the fate of employees, structural holes are important for collecting advice and information. As we will describe, we take a different view of network content – one focused on the subject matter of communication rather than the social function of relationships.

The limited research that does empirically examine networks and information content has either focused on identifying tie characteristics, network characteristics and information characteristics that

Essay 2: Network Structure & Information Advantage

facilitate effective knowledge transfers; or on the types of information (e.g. complex or simple; tacit or explicit) most effectively transferred through different types of ties. As a result, the fundamental assumption that structurally diverse network contacts provide access to diverse and novel information remains unexplored. For example, several studies examine how characteristics of dyadic relationships, like the strength of ties, impact the effectiveness of knowledge transfer, and how knowledge transfer processes in turn affect performance (Granovetter 1973, Uzzi 1996, 1997, Hansen 1999). These studies infer the impact of network structure on the effectiveness of knowledge sharing from the strength of individual dyadic relationships. Reagans & McEvily (2003) extend this work by simultaneously examining the effects of tie strength and network structure on the ease of transferring knowledge between individuals. These studies either examine the strength of dyadic ties or the impact of network structure on discrete dyadic information transfer events, instead of the information that accrues to actors from all their network contacts in concert. Other studies examine characteristics of the information transferred across different types of network ties. For example, Hansen (1999, 2002) and Uzzi (1996, 1997) explore the degree to which knowledge being transferred is tacit or codifiable, simple or complex, and related or unrelated to a focal actor's knowledge.

To complement this research, we ask a related, yet fundamentally different question: Do networks affect the acquisition of diverse and novel information and to what extent does this intermediate mechanism drive performance? In pursuing this question, we undertake two fundamental departures from the current literature. First, by exploring the relative similarity or difference of the information content individuals receive from different network contacts, rather than information content in discrete transfers, we explore an actor's information diversity in relation to the body of information available in the network. Second, we focus on subject matter. Rather than characterizing the simplicity or complexity of information, or the degree to which knowledge is codifiable or tacit, we explore the topical content being discussed. Both simple and complex information can be either focused or diverse in terms of subject matter. Degrees of complexity and codifiability do not describe whether information is topically similar or dissimilar, or novel relative to a larger body of knowledge.

Essay 2: Network Structure & Information Advantage

These two departures from previous research are important to detailed exploration of information advantage because the theoretical mechanism linking structure to performance through information rests on the relative diversity of the information to which actors have access. Examining topics in information content that accrue to actors from their collection of network contacts is critical to effectively observing the relative content diversity of information, and thus the aspects of information theorized to drive value. In the following section, we develop hypotheses linking network structure to performance through access to diverse and novel information.

2.2. Network Determinants of Information Advantage

The empirical distribution of information across a network is likely determined in part by the structural pattern of relationships through which information is shared. All else equal, actors with strong ties and frequent communication tend to share information and accumulate similar knowledge. This relationship between social structure and information structure should create pools of similar information content collected in densely connected local network neighborhoods, and the structural properties of actors' positions in the social structure should in turn help determine the diversity or focus of the pools of information to which they have access. Two network characteristics in particular are theorized to drive access to diverse, novel information: network size and network diversity. These characteristics are fundamental because they represent the two dimensions of structure most directly related to information acquisition. As Burt (1992: 16) argues "everything else constant, a large, diverse network is the best guarantee of having a contact present where useful information is aired..."

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

Essay 2: Network Structure & Information Advantage

breadth. The greater the size of an actor's network, the more likely she is to have access to more information and to multiple social circles increasing the diversity of her information. However, size may not matter if each additional contact is embedded in the same social circles. Network diversity may therefore be more important in providing access to diverse information.

Network Diversity. Network diversity determines the number of non-redundant pools of information to which an actor is connected and therefore the channels through which new, diverse information can flow. Network diversity describes the degree to which contacts are structurally 'non-redundant,' and there are both first order and second order dimensions of redundancy as shown in Figure 1. In the first order, direct contacts can be connected to each other. Individuals who are in contact are likely to share information and be aware of the same opportunities, ideas and expertise. Formally, networks in which contacts are highly connected are termed 'cohesive.' In the second order, contacts in a network can themselves be connected to the same people, connecting the focal actor indirectly to redundant sources of information. Contacts that are themselves connected to the same people are termed 'structurally equivalent.'

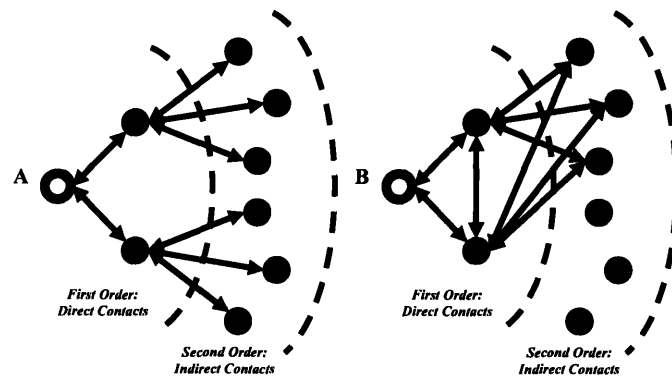


Figure 1. Structurally diverse networks are low in a) cohesion and b) structural equivalence. Actor A has two unconnected contacts which display no structural equivalence, while B has two redundant contacts that are connected and maximally structurally equivalent.

We measure redundancy in the first order of direct contacts by the lack of constraint in actors' networks, and in the second order by the average structural equivalence of actors' contacts. We define the

Essay 2: Network Structure & Information Advantage

constraint C_i (Burt 1992: 55)²⁸ of an actor's network as the degree to which an individual's contacts are connected to each other, such that $C_i = \sum_j \left(p_{ij} + \sum_q p_{iq} p_{qj} \right)^2$, $q \neq i, j$; and the structural diversity D_i of an actor's network as $1 - C_i$. We use the standard definition of the structural equivalence of two actors, measured as the Euclidean distance of their contact vectors.²⁹ As individuals communicate with more contacts, and as individuals' networks connect them to actors that are themselves unconnected and structurally non-equivalent, we expect the information they receive to be more diverse and we expect them to receive more total novel information:

H1: Network size and network diversity are positively associated with greater information diversity and with access to more non-redundant information.

While a greater number of contacts are likely to provide access to more diverse, non-redundant information, the probability that an additional contact will have novel information is likely decreasing in the size of an individual's network. Social networks tend to cluster into homophilous cliques (Blau 1986, for a review see McPherson, Smith-Loving, & Cook 2001). Since individuals usually make connections through contacts they already have, in bounded networks the likelihood that a marginal contact will be redundant should increase in the number of people I already know.³⁰ As actors establish relationships with a finite set of alters, the probability that a marginal relationship will be structurally non-redundant should decrease as possible alters in the network are exhausted. We therefore expect marginal increases in information diversity and network diversity are decreasing in network size:

H2a: The marginal increase in information diversity is decreasing in network size.

H2b: The marginal increase in structural network diversity is decreasing in network size.

²⁸ Where $p_{ij} + \sum_q p_{iq} p_{qj}$ measures the proportion of i 's network contacts that directly or indirectly involve j and C_i sums this across all of i 's contacts.

²⁹ The Euclidean distance measures the square root of the sum of squared distances between the contact vectors, approximating the degree to which contacts are connected to the same people. The structural equivalence score of a focal actor is calculated as the average structural equivalence of the actor's direct contacts.

³⁰ We focus on internal networks due to difficulties in collecting reliable data outside the firm and in estimating accurate network structures without access to whole network data (see Barnes 1979, Marsden 1990 and footnote #4). As Burt (1992: 172) demonstrates however "little evidence of hole effects [are] lost... when sociometric choices [are] restricted to relations within the firm."

2.3. Non-Network Determinants of Information Advantage

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

Demography & Human Capital. Demography can influence performance, learning capabilities and the variety of ideas to which individuals have access (e.g. Ancona & Caldwell 1992, Reagans & Zuckerman 2001). Older employees may have prior knowledge on a wider variety of topics or may be more aware of experts in the organization. Employment discrimination and interpersonal difference could also impact the relative performance and information seeking habits of men and women. We therefore control for the age and gender of employees. Greater industry experience, education or individuals' organizational position could also create variation in access to diverse and novel information and performance. As individuals gain experience, they may collect expertise across several domains, or specialize and focus their work and communication on a limited number of topics. We therefore control for education, industry experience and organizational position.³¹

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

Individual Characteristics & Temporal Shocks. Some employees may simply be more social or more ambitious, creating variation in information seeking habits and performance. To control for unobservable individual characteristics we test fixed effects specifications of each hypothesis. At the same time, temporal shocks could affect demand for the firm's services and information seeking activities

³¹ Employees are partners, consultants or researchers – we include dummy variables for each of these positions.

associated with more work.³² These exogenous shocks to demand could drive increases in project workload, information seeking, and revenue generation creating a spurious correlation between information flows and output. We therefore control for temporal variation with dummy variables for each month and year.

2.4. The Setting – Executive Recruiting

We studied a medium-sized executive recruiting firm with fourteen offices in the U.S. Interviews revealed that the core of executive recruiters' work involves matching job candidates to clients' requirements.³³ This matching process is information-intensive and requires activities geared toward assembling, analyzing, and making decisions based on information gathered from team members, other firm employees, and contacts outside the firm. Qualitative studies demonstrate that executive recruiters fill "brokerage positions" between clients and candidates and rely heavily on flows of information to complete their work effectively (Finlay & Coverdill 2000). In our context, more precise or accurate information about the candidate pool reduces time wasted interviewing unsuitable candidates and poor decisions based on faulty or incomplete information, making recruiters more effective. In addition, the sharing of procedural information can improve the efficiency with which employees handle recurrent search problems. Knowledge sharing on difficult recurring situations improves effectiveness (Szulanski 1996) and executive recruiters report learning to deal with difficult situations through communication with peers.

Recruiters generate revenue by filling vacancies rather than billing hourly. Therefore, the speed with which vacancies are filled is an important intermediate measure of workers' productivity. Contract completion implies that recruiters have met a client's minimum thresholds of candidate fit and quality.

³² In our data, business exhibits seasonal variation, with demand for the firm's services picking up sharply in January and declining over the next eight months. There may also be transitory shocks to demand in a given month or year.

³³ "Client" refers to a firm seeking to hire one or more executives; "candidate" refers to a potential hire; and "recruiter" refers to someone expert in locating, vetting, and placing candidates.

Essay 2: Network Structure & Information Advantage

Project duration can therefore be interpreted as a quality controlled measure of productivity. In assessing the performance of individual recruiters, we measure revenues generated per month, projects completed per month and average project duration per month. Effective recruiters rely on being “in the know” and delivering candidates that display specific professional and personal attributes. To accomplish this, recruiters must be aware of several different information channels to match different candidates with different client requirements. We therefore expect recruiters with diverse and non-redundant information to complete more projects, to complete projects faster and to generate more revenue for the firm per unit time.

H3: Access to non-redundant and diverse information is positively associated with more project completions, faster project completions and more revenue generated per unit time.

While we expect network structure to impact performance through its effects on access to diverse and novel information, there could be other intermediate mechanisms tying structure to performance. Network contacts could provide resources other than information (e.g. Podolny & Barron 1997), there could be power or control benefits to network structure independent of information flows (e.g. Burt 1992), and structural diversity could reduce dependence, place individuals in favorable trading relationships (e.g. Emerson 1962) or entitle them to benefits from informal reciprocity (e.g. Cook, Emerson & Gilmore 1983). If network diversity is positively associated with performance holding access to novel information constant, then there may be additional benefits to network structure beyond those conferred through information advantage. Given these alternative intermediate mechanisms, we hypothesize that network diversity is positively associated with performance even controlling for access to novel information.

H4: Network diversity is positively associated with more project completions, faster project completions and more revenue generated per unit time, controlling for access to novel information.

3. Methods

By analyzing email communication patterns and message content, we are able to match information channels to the subject matter of the content flowing through them. Our empirical approach

Essay 2: Network Structure & Information Advantage

also addresses another methodological puzzle that has historically troubled network research: In traditional network studies, a fundamental tradeoff exists between comprehensive observation of whole networks and the accuracy of respondents' recall. Most research elicits network data from respondents who have difficulty recalling their networks (e.g. Bernard et. al 1981), especially among individuals socially distant to themselves (Krackhardt & Kilduff 1999). The inaccuracy of respondent recall and the bias associated with recall at social distance creates inaccurate estimates of network variables (Kumbasar, Romney & Batchelder 1994), forcing most empirical studies to artificially limit the boundary of estimated networks to local areas around respondents (e.g. Reagans & McEvily 2003). Such empirical strategies create estimation challenges due to the sensitivity of network metrics to the completeness of data (Marsden 1990). If important areas of the network are not captured, estimates of network positions can be bias. Furthermore, as our content measures consider the similarity of topics across the entire network, poor coverage of the firm could bias our estimates of the relative novelty or diversity of topics discussed via email. We therefore take several steps to ensure a high level of participation (described below). As 87% of eligible recruiters agreed to participate, and given that our inability to observe the remaining 13% is limited to messages between two employees who both opted out of the study, we collect email network and content data with nearly full coverage of the firm.³⁴

3.1. Data

Our data come from four sources: (i) detailed accounting records of individual project assignments and performance, (ii) email data from the corporate server, (iii) survey data on demographic characteristics, human capital and information seeking behaviors, and (iv) data from the web site Wikipedia.org used to validate our analytical models of information diversity.

Internal accounting data describe: (i) revenues generated by individual recruiters, (ii) contract start and stop dates, (iii) projects handled simultaneously by each recruiter, (iv) project team composition,

³⁴ *F*-tests comparing performance levels of those who opted out with those who remained did not show statistically significant differences. *F* (Sig): Rev02 2.295 (.136), Comp02 .837 (.365), Multitasking .386 (.538).

Essay 2: Network Structure & Information Advantage

and (v) job levels of recruiters and placed candidates. These provide excellent performance measures that can be normalized for quality.

Email data cover 10 months of complete email history at the firm. The data were captured from the corporate mail server during two equal periods from October 1, 2002 to March 1, 2003 and from October 1, 2003 to March 1, 2004. Participants received \$100 in exchange for permitting use of their data, resulting in 87% coverage of recruiters eligible to participate and more than 125,000 email messages captured. Details of email data collection are described by Aral, Brynjolfsson & Van Alstyne (2006).³⁵

The third data set contains survey responses on demographic and human capital variables such as age, education, industry experience, and information-seeking behaviors. Survey questions were generated from a review of relevant literature and interviews with recruiters. Experts in survey methods at the Inter-University Consortium for Political and Social Science Research vetted the survey instrument, which was then pre-tested for comprehension and ease-of-use. Individual participants received \$25 for completed surveys and participation exceeded 85%.

The fourth dataset is a set of 291 entries collected from Wikipedia.org, which we describe in detail in the section pertaining to the validity of our information diversity metrics (see § 3.2.4).

Descriptive statistics and variable correlations are provided in Tables 1 and 2 respectively.

³⁵ We wrote and developed capture software specific to this project and took multiple steps to maximize data integrity. New code was tested at Microsoft Research Labs for server load, accuracy and completeness of message capture, and security exposure. To account for differences in user deletion patterns, we set administrative controls to prevent data expunging for 24 hours. The project went through nine months of human subjects review prior to launch and content was masked using cryptographic techniques to preserve individual privacy. Spam messages were excluded by eliminating external contacts who did not receive at least one message from someone inside the firm. Email data were collected using EmailNet, created by Jun Zhang and Marshall Van Alstyne.

Table 1: Descriptive Statistics

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

Table 2: Pair Wise Correlations Between Independent Variables

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

* p < .05

3.2. Modeling & Measuring Information Diversity

To test the relationship between network structure and information diversity, we analyze data on a relatively complete dimension of the social information to which employees in our firm have access – email communication. While email is not the only source of employees’ social communication and information gathering, it is one of the most pervasive communication media that preserves records of communication content, and is a good proxy for other social sources of information in organizations where email is widely used.³⁶ Our interviews indicate that in our firm, email is a primary communication media.

3.2.1. Modeling & Measuring Topics in Email: A Vector Space Model of Communication Content

We model and measure the diversity of information in individuals’ email inboxes and outboxes using a Vector Space Model of the topics present in email content (e.g. Salton et. al. 1975). Vector Space Models are widely used in information retrieval and search query optimization algorithms to identify documents that are similar to each other or pertain to topics identified by search terms. Vector Space Models represent textual content as vectors of topics in multidimensional space based on the relative prevalence of topic keywords. In our model, each email is represented as a multidimensional ‘topic vector’ whose elements are the frequencies of keywords in the email. The prevalence of certain keywords indicates that a topic that corresponds to those keywords is being discussed. For example, an email about pets might include two mentions of the word “dog,” two of “cat,” and three of “veterinarian;” while an email about econometrics might include three uses of the word “variance,” two of “specification,” and three of “heteroskedasticity.” The relative topic similarity of two emails can then be assessed by the degree to which their topic vectors converge (point toward the same topics in vector space) or diverge (point in orthogonal directions in vector space).³⁷

³⁶ In our data, the average number of contacts by phone ($\rho = .30$, $p < .01$) and instant messenger ($\rho = .15$, $p < .01$) are positively and statistically significantly correlated with email contacts.

³⁷ Each email may pertain to multiple topics based on keyword prevalence, and topic vectors representing emails can emphasize one topic more than another based on the relative frequencies of the keywords that pertain to different

Essay 2: Network Structure & Information Advantage

To measure content diversity, we characterize all emails as topic vectors and measure the variance or spread of topic vectors in individuals' inboxes and outboxes. Emails about similar topics contain similar language on average, and vectors used to represent them are therefore closer in multidimensional space, reducing their collective variance or spread. These methods are regularly used to automate topic discovery in documents and communications. In the next two sections we present our modeling strategies for construction of topic vectors, selection of keywords, and content diversity measurement.

3.2.2. Construction of Topic Vectors & Keyword Selection³⁸

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

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

where k_{ij} represents the weight of the j th keyword. A three dimensional vector space model of three documents is shown in Figure 2.

topics. In this way, our framework captures nuances of emails that may pertain to several topics of differing relative emphasis.

³⁸ We thank Petch Manoharn for his tireless coding efforts that extracted and manipulated the email data described in the next three sections.

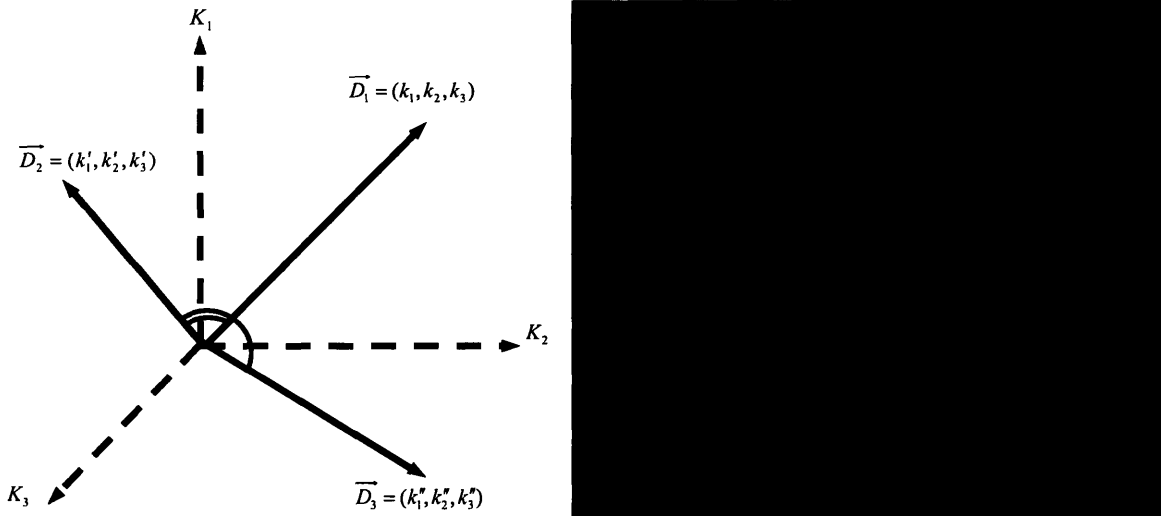


Figure 2. A three dimensional Vector Space Model of three documents is shown on the left. A Vector Space Model containing a test inbox with emails clustered along three dimensions is shown on the right.

Weights define the degree to which a particular keyword impacts the vector characterization of a document. Words that discriminate between documents or topics are weighted more heavily than words less useful in distinguishing topics. As terms that appear frequently in a document are typically thematic and relate to the document’s subject matter, we use the term frequency of keywords in email as weights to construct topic vectors and refine our keyword selection with criteria designed to select words that *distinguish and represent topics*.³⁹

In order to minimize their impact on the clustering process, we initialized our data by removing common “stop words,” such as “a,” “an,” “the,” “and,” and other common words with high frequency across all emails that are likely to create noise in content measures. We then implemented an iterative, k-means clustering algorithm to group emails into clusters that use the same words, similar words or words that frequently appeared together.⁴⁰ The result of iterative k-means clustering is a series of assignments of

³⁹ Another common weighting scheme is the ‘term-frequency/inverse-document frequency.’ However, we use a more sophisticated keyword selection refinement method specific to this dataset described in detail in the remainder this section.

⁴⁰ K-means clustering generates clusters by locally optimizing the mean squared distance of all documents in a corpus. The algorithm first creates an initial set of clusters based on language similarities, computes the ‘centroid’ of each cluster, and then reassigns documents to clusters whose centroid is the closest to that document in topic space.

emails to clusters based on their language similarity. Rather than imposing exogenous keywords on the topic space, we extract topic keywords likely to characterize topics by using a series of algorithms guided by three basic principles.

First, in order to identify distinct topics in our corpus, keywords should *distinguish* topics from one another. To achieve this goal, we chose keywords that maximize the variance of their mean frequencies across clusters produced by the k-means clustering procedure. This refinement favors words with widely differing mean frequencies across clusters, suggesting an ability to distinguish between topics. In our data, we find the coefficient of variation of the mean frequencies across topics to be a good indicator of this dispersion.⁴¹

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

Second, keywords should *represent* the topics they are intended to identify. In other words, keywords identifying a given topic should frequently appear in emails about that topic. To achieve this goal we chose keywords that minimize the mean frequency variance within clusters, favoring words that are consistently used across emails discussing a particular topic.⁴²

$$ITF_i = \frac{\sqrt{\sum_c \sum_i (f_i - \bar{M}_c)^2}}{\bar{M}_c}$$

Third, keywords should not occur too infrequently. If there are very few instances of keywords in the email corpus, keywords will not represent or distinguish topics and will create relatively sparse topic vectors that are difficult to compare. We therefore select high frequency words (not eliminated by the

The algorithm stops iterating when no reassignment is performed or when the objective function falls below a pre-specified threshold.

⁴¹ The coefficient of variation is particularly useful due to its scale invariance, enabling comparisons of datasets, like ours, with heterogeneous mean values (Ancona & Caldwell 1992). To ease computation we use the square of the coefficient of variation, which produces a monotonic transformation of the coefficient without effecting our keyword selection.

⁴² i indexes emails and c indexes k-means clusters. We squared the variation to ease computation as in footnote 42.

“stop word” list of common words) that maximize the inter-topic coefficient of variation and minimize intra-topic mean frequency variation. The results of this process are a set of keywords describing subject topics generated from usage characteristics of the email communication of employees at our research site.⁴³

3.2.3. Measuring Email Content Diversity

Using the keywords generated by our usage analysis, we populated topic vectors representing the subject matter of the emails in our data. We then measured the degree to which the emails in an individual employee’s inbox or outbox were focused or diverse by measuring the spread or variance of their topic vectors. We created five separate diversity measurement specifications based on techniques from the information retrieval, document similarity and information theory literatures. The approach of all five measures is to compare individuals’ emails to each other, and to characterize the degree to which emails are about a set of focused topics, or rather about a wider set of diverse topics. We used two common document similarity measures (Cosine similarity and Dice’s coefficient) and three measures enhanced by an information theoretic weighting of emails based on their “information content.”⁴⁴ Detailed descriptions of diversity measures and their correlations are provided in Appendix A. As all diversity measures are highly correlated (\sim corr = .98; see Appendix A), our specifications use the average cosine distance of employees’ incoming email topic vectors d_{ij}^I from the mean vector of their topic space M_i^I to represent incoming information diversity (ID_i^I):

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

⁴³ We conducted sensitivity analysis of our keyword selection process by choosing different thresholds at which to select words based on our criteria and found results were robust to all specifications and generated keyword sets more precise than those used in traditional term frequency/inverse document frequency weighted vector space models that do not refine keyword selection.

⁴⁴ Information Content is used to describe how informative a word or phrase is based on its level of abstraction. Formally, the information content of a concept c is quantified as its negative log likelihood $-\log p(c)$.

Essay 2: Network Structure & Information Advantage

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

3.2.4. Validating Diversity Measures

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

We created a range of high to low diversity clusters of Wikipedia entries by selecting entries from either the same sub-category in the topic hierarchy to create focused clusters, or from a diverse set of unrelated subtopics to create diverse clusters. For example, we created a minimum diversity cluster (Type-0) using a fixed number of documents from the same third level sub-category of the topic hierarchy, and a maximum diversity cluster (Type-9) using documents from unrelated third level sub-categories. We then constructed a series of document clusters (Type-0 to Type-9) ranging from low to

high topic diversity from 291 individual entries as shown in Figure 3.⁴⁵ The topic hierarchy from which documents were selected is shown in Appendix B.

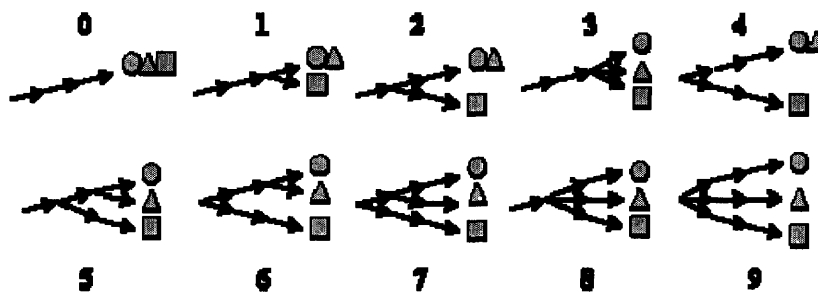


Figure 3. Document clusters selected from Wikipedia.org

If our measurement is robust, our diversity measures should identify Type-0 clusters as the least diverse and Type-9 clusters as the most diverse. We expect diversity will increase relatively monotonically from Type-0 to Type-9 clusters, although there could be debate for example about whether Type-4 clusters are more diverse than Type-3 clusters.⁴⁶ After creating this independent dataset, we used the Wikipedia entries to generate keywords and measure diversity using the methods described above.

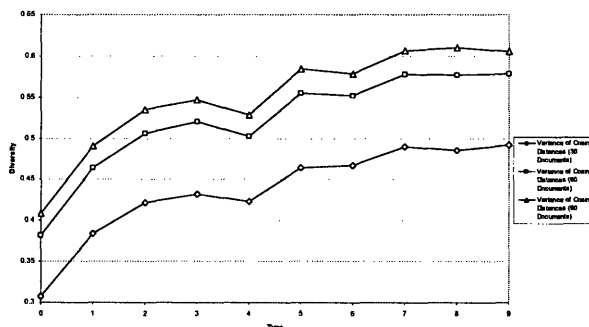


Figure 4. Results of Diversity Measurement validation using data from Wikipedia.org.

Our methods were very successful in characterizing diversity and ranking clusters from low to high diversity. Figure 4 displays cosine similarity metrics for Type-0 to Type-9 clusters using 30, 60, and 90 documents to populate clusters. All five diversity measures return increasing diversity scores for

⁴⁵ We created several sets of clusters for each type and averaged diversity scores for clusters of like type. We repeated the process using 3, 6 and 9 document samples per cluster type to control for the effects of the number of documents on diversity measures.

⁴⁶ Whether Type-3 or Type-4 clusters are more diverse depends on whether the similarity of two documents in the same third level sub category is greater or less than the difference of similarities between documents in the same second level sub category as compared to documents in categories from the first hierarchical layer onwards. This is, to some extent, an empirical question.

clusters selected from successively more diverse topics.⁴⁷ Overall, these results give us confidence in the ability of our diversity measurement to characterize the subject diversity of groups of text documents of varying sizes.

3.3. Statistical Specifications

We began by examining the structural determinants of access to diverse and novel information. We first estimated an equation relating network structure to the diversity of information flowing into actors' email inboxes using pooled OLS specifications controlling for individual characteristics and fixed effects models on monthly panels of individuals' networks and information diversity.⁴⁸ We focused on the size and structural diversity of individuals' networks and controlled for temporal variation by month. The estimating equation is specified as follows:

$$ID_{it}^I = \gamma_i + \beta_1 E_{it}^I + \beta_2 NS_{it} + \beta_3 NS_{it}^2 + \beta_4 ND_{it} + \beta_5 SE_{it} + \sum_j B_j HC_{ji} + \sum_m B_m Month + \varepsilon_{it} \quad [1],$$

where ID_{it}^I represents the diversity of the information in a given individual's inbox, E_{it}^I represents the total number of incoming messages received by i , NS_{it} represents the size of i 's network, NS_{it}^2 represents network size squared, ND_{it} represents structural diversity (measured by one minus constraint), SE_{it} represents structural equivalence, $\sum_j \beta_j HC_{ji}$ represents controls for human capital and demographic variables (Age, Gender, Education, Industry Experience, and Managerial Level), and $\sum_m \beta_m Month$ represents temporal controls for each month/year.

⁴⁷ The measures produce remarkably consistent diversity scores for each cluster type and the diversity scores increase relatively monotonically from Type-0 to Type-9 clusters. The diversity measures are not monotonically increasing for all successive sets, such as Type-4, and it is likely that the information contained in Type-4 clusters are less diverse than Type-3 clusters due simply to the fact that two Type-4 documents are taken from the same third level sub category.

⁴⁸ We focus in this paper on incoming information for two reasons. First, we expect network structure to influence incoming information more than outgoing information. Second, the theory we intend to test is about the information to which individuals have access as a result of their network structure, not the information individuals send. These dimensions are high correlated.

Essay 2: Network Structure & Information Advantage

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

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

To explore the mechanisms driving the non-linear relationship between network size and information diversity, we tested our hypothesis (2b) that while structural diversity is increasing in size, there are diminishing marginal diversity returns to size in bounded networks. If this is the case, we should see a non-linear positive relationship between network size and structural diversity, such that the marginal increase in structural diversity is decreasing in size. To test this hypothesis, we specified the following model:

$$ND_{it} = \gamma_i + \beta_1 NS_{it} + \beta_2 NS_{it}^2 + \sum_j B_j HC_{ji} + \sum_m B_m Month + \varepsilon_{it} \quad [3].$$

Finally, we tested the performance implications of network and information diversity. We tested the relationship between non-redundant information (NRI'_{it}) and performance (P_{it}), and included our measure of structural network diversity (ND_{it}) in the specification.

$$P_{it} = \gamma_i + \beta_1 NRI'_{it} + \beta_2 ND_{it} + \sum_j B_j HC_{ji} + \sum_m B_m Month + \varepsilon_{it} \quad [4].$$

If information benefits to network diversity exist, network diversity should be positively associated with access to diverse and non-redundant information, and non-redundant information should be positively associated with performance. If network diversity confers additional benefits beyond information advantage (such as power or favorable trading conditions) network diversity should contribute to performance beyond its contribution through information diversity.⁴⁹

⁴⁹ We were unable to reject the hypothesis of no heteroscedasticity and report standard errors according to the White correction (White 1980). White's approach is conservative. Estimated coefficients are unbiased but not efficient. In small samples, we may observe low t-statistics even when variables exert a real influence. As there may be idiosyncratic error at the level of individuals, for OLS analyses we report robust standard errors clustered by

4. Results

4.1. Network Structure & Access to Diverse, Non-Redundant Information

We first estimated the relationships between network size, network diversity and access to diverse information controlling for demographic factors, human capital, unobservable individual characteristics, temporal shocks and the total volume of communication.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
<i>Dependent Variable:</i>	Information Diversity	Information Diversity	Information Diversity	Information Diversity	Information Diversity	Information Diversity
<i>Specification</i>	<i>Fixed Effects</i>	<i>OLS-c</i>	<i>Fixed Effects</i>	<i>OLS-c</i>	<i>Fixed Effects</i>	<i>OLS-c</i>
Age		.006 (.009)		-.001 (.006)		-.001 (.006)
Gender		.003 (.135)		.142 (.092)		.135 (.097)
Education		-.061 (.006)		-.001 (.041)		-.002 (.042)
Industry		.010 (.010)		-.003 (.006)		-.001 (.007)
Experience		-.147 (.284)		.258 (.161)		.175 (.188)
Partner		-.006 (.246)		.159 (.150)		.122 (.168)
Consultant						
Total Email	-.001 (.001)	.000 (.001)	.001 (.001)	.001 (.001)	-.001 (.001)	.001 (.001)
Incoming						
Network	1.299*** (.133)	1.38*** (.301)			.474*** (.114)	.296* (.138)
Size						
Network	-.880*** (.112)	-1.048*** (.266)			-.272** (.089)	-.240* (.139)
Size-Squared						
Network			.239*** (.048)	.338*** (.073)	.128** (.052)	.268*** (.072)
Diversity						
Structural			-.018 (.032)	.063 (.094)	-.005 (.033)	.062 (.096)
Equivalence						
Constant	.059 (.094)	.863 (.895)	.056 (.072)	-.039 (.635)	.128* (.075)	.016 (.634)
Temporal						
Controls	Month	Month	Month	Month	Month	Month
F-Value (d.f.)	13.70*** (11)	3.76*** (17)	4.57*** (11)	4.86*** (17)	5.61*** (13)	5.03*** (19)
R ²	.24	.38	.10	.23	.14	.24
Obs.	563	448	540	434	540	434

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

Essay 2: Network Structure & Information Advantage

Our results demonstrate that the diversity of information flowing to an actor is increasing in the actor's network size and network diversity, while the marginal increase in information diversity is decreasing in network size, supporting hypotheses 1 and 2a. In all specifications, the size of recruiters' networks is positively correlated with information diversity; while the coefficient on network size squared is negative and significant. Network diversity is also positively and significantly associated with greater information diversity in incoming email. The first order diversity variable which measures the lack of constraint in the an actor's network, or the degree to which an actor's network contacts are unconnected, is highly significant in all specifications, while the average structural equivalence of actors' contacts does not influence access to diverse information controlling for network size and first order structural diversity. These results demonstrate that the structure of recruiters' networks influence access to diverse information and support the hypothesis that large diverse networks provide access to diverse, novel sets of information. They also show diminishing marginal returns to network size. As actors add network contacts, the contribution to information diversity lessens, implying that information benefits to network size are constrained in bounded networks.

We then tested relationships between network size, network diversity and the total amount of novel information that accrues to recruiters in incoming email. Our results, shown in Table 4, demonstrate that the amount of novel information flowing to an actor is increasing in the actor's network size and network diversity. Network diversity has a strong positive relationship with the total amount of novel information flowing into actors' inboxes, but is not significant when controlling for network size, implying that the larger of two similarly diverse networks receives the most non-redundant information. The impact of size on total novel information dominates that of structural diversity because of the strong relationship between size and total incoming email, a critical driver of the total amount of novel information (pair wise correlation: $\rho = .98, p < .01$). The implication of this result is to highlight the importance of flows of information over time. If actors are constantly receiving information from each of their network contacts, the flow of unique bits of information in two networks of similar structural diversity will be greater in the larger network. We would expect network diversity to drive greater access

Essay 2: Network Structure & Information Advantage

to non-redundant information controlling for the size of the network. As Burt (1992: 16-23) argues:

“given two networks of equal size, the one with more non-redundant contacts provides more benefits.”

However, our results imply a more complicated relationship between network size and network diversity

– an implication we explore in greater detail in the next section.

Table 4. Network Structure & Access to Non-Redundant Information

	Model 1	Model 2	Model 3	Model 4
<i>Dependent Variable:</i>	Non-Redundant Information	Non-Redundant Information	Non-Redundant Information	Non-Redundant Information
<i>Specification</i>	<i>Fixed Effects</i>	<i>OLS-c</i>	<i>Fixed Effects</i>	<i>OLS-c</i>
Age		.000 (.012)		-.005 (.010)
Gender		-.006 (.188)		-.127 (.155)
Education		-.068 (.062)		-.098 (.053)
Industry		-.029**		-.015
Experience		(.013)		(.010)
Partner		-.480 (.437)		-.244 (.395)
Consultant		-.839 (.318)		-.403 (.296)
Network Size			.711*** (.127)	1.195*** (.234)
Network Size-Squared			-.109 (.103)	-.518* (.263)
Network Diversity	.229*** (.061)	.530*** (.131)	-.070 (.060)	-.138 (.103)
Structural Equivalence	-.053 (.043)	-.000 (.006)	.022 (.037)	-.138 (.102)
Constant	-.281*** (.079)	1.655** (1.090)	-.247*** (.068)	1.784** (.890)
Temporal Controls	Month	Month	Month	Month
F-Value (d.f.)	10.54*** (10)	12.86*** (16)	25.05*** (12)	15.85*** (18)
R ²	.19	.35	.40	.55
Obs.	540	434	540	434

4.2. Tradeoffs between Network Size & Network Diversity

There is a strong, positive, but non-linear relationship between network size and network diversity in our data: structural diversity is increasing in network size, but with diminishing marginal returns (see Table 5). The coefficient on the size variable is positive and significant in relation to first order network diversity, and the size squared term is negative and significant. This result supports

Essay 2: Network Structure & Information Advantage

hypothesis 2b, and demonstrates why information benefits to larger networks may be constrained in bounded organizational networks. As recruiters contact more colleagues, the contribution of a marginal contact to the structural diversity of a focal actor's network is increasing, but with diminishing marginal returns, exploring more deeply the structural antecedents to information advantage.

	Model 1	Model 2	Model 3	Model 4
<i>Dependent Variable:</i>	Network Diversity	Network Diversity	Structural Equivalence	Structural Equivalence
<i>Specification</i>	<i>Fixed Effects</i>	<i>OLS-c</i>	<i>Fixed Effects</i>	<i>OLS-c</i>
Age		-.005 (.006)		.016** (.005)
Gender		-.156* (.091)		.024 (.102)
Education		-.030 (.034)		.011 (.045)
Industry		.025** (.009)		-.012 (.007)
Experience		-.004 (.186)		-1.012*** (.202)
Partner		.192 (.140)		-.940*** (.167)
Consultant		1.585*** (.113)	1.626*** (.209)	-.077 (.145)
Network Size		-1.038*** (.098)	-1.069*** (.190)	-.109 (.122)
Network Size-Squared		.083 (.064)	.651 (.630)	-.907*** (.074)
Constant				-.946 (.784)
Temporal Controls	Month	Month	Month	Month
F-Value (d.f.)	33.39*** (10)	15.58*** (16)	62.39*** (10)	59.97*** (16)
R ²	.41	.64	.58	.58
Obs.	563	448	540	434

The implications of a fundamental trade off between size and structural diversity complement Burt's (1992: 167) concepts of "effective size" and "efficiency."⁵⁰ Figure 5 displays a graph relating network size and network diversity, clearly showing the positive, non-linear relationship.

⁵⁰ In fact, Burt (1992: 169) finds stronger evidence of hole effects with the constraint measures we employ than with effective size, demonstrating "exclusive access is a critical quality of relations that spans structural holes."

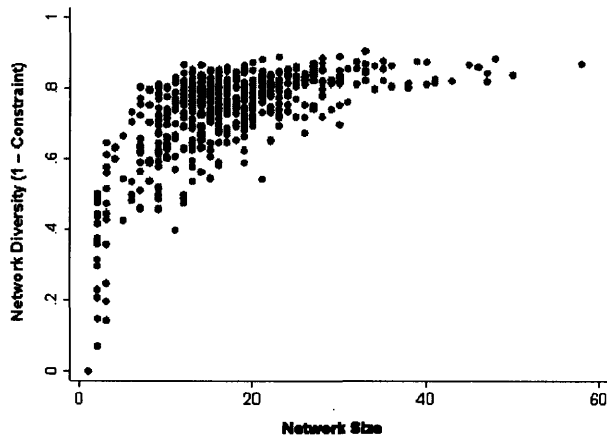


Figure 5. Graph of the relationship between network size and network diversity.

4.3. Network Structure, Information Diversity & Performance

Finally, we test the implications of network structure and access to diverse, non-redundant information for the performance of executive recruiters across revenues generated per month, projects completed per month, and the average duration of projects.⁵¹ Table 6 displays the results of fixed effects, between and pooled OLS estimates of these relationships. The results show strong evidence of a positive relationship between access to non-redundant information and performance. In fixed effects models, which control for variation explained by unobserved, time invariant characteristics of individuals, a one unit increase in the amount of non-redundant information flowing to individuals is associated on average with \$3,806.12 more revenue generated, an extra one tenth of one project completed, and 14 days shorter average project duration per person per month. Between estimates are all in the same direction and of similar magnitude, although only the relationship with revenue is significant. Pooled OLS estimates also show across the board that access to non-redundant information is associated with greater revenue generation, more completed projects per unit time and faster project completions.

⁵¹ As there are some employees who do not take on projects or who are not involved in any projects in a given month, we only estimate equations for individuals with non-zero revenues in a given month.

Table 6. Network Structure, Non-Redundant Information and Individual Performance

<i>Dependent Variable:</i>	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
	Revenue <i>Fixed Effects Within</i>	Revenue <i>Between Estimator</i>	Revenue <i>OLS-c</i>	Completed Projects <i>Fixed Effects Within</i>	Completed Projects <i>Between Estimator</i>	Completed Projects <i>OLS-c</i>	Fixed Effects <i>Within</i>	Project Duration <i>Between Estimator</i>	Project Duration <i>OLS-c</i>
Age			-241.75 (294.08)			-.006 (.005)			.344 (2.147)
Gender			-6217.33 (3816.54)			-.096 (.056)			-12.155 (26.346)
Education			-774.60 (1103.03)			-.003 (.022)			17.769 (10.686)
Industry			-91.58 (278.91)			-.002 (.006)			4.251 (2.529)
Experience			12979.80 (8533.10)			.156 (.159)			-83.392 (79.325)
Partner			9277.93 (6763.74)			.250** (.121)			-104.555 (57.056)
Consultant			7709.13** (3143.62)		.084 (.059)	.172** (.050)			-26.461* (14.931)
Non-Redundant Information	3806.12** (1211.06)	4726.45* (2783.69)		.097*** (.024)			-14.211** (5.44)	-35.233 (25.516)	
Network	165.14 (931.52)	5558.04* (3268.62)	3202.45* (1779.18)	.212 (.018)	.070 (.069)	.057* (.032)	-12.735** (4.18)	33.238 (29.961)	-14.764 (11.499)
Diversity	35238.48*** (1442.79)	28921.45* (16214.11)	56129.10** (20886.57)	.660*** (.028)	.402 (.344)	.873** (.431)	288.926*** (6.482)	243.027 (148.623)	-36.571 (190.419)
Temporal Controls	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year	Month / Year
F-Value	2.16** (10)	5.21*** (8)	3.97*** (16)	3.15*** (10)	3.46** (8)	4.72*** (16)	3.32*** (10)	1.72 (8)	4.06*** (16)
R ²	.06	.49	.24	.08	.39	.27	.08	.24	.28
Obs.	420	420	320	420	420	320	420	420	320

Essay 2: Network Structure & Information Advantage

These results support hypothesis 3 and provide evidence for ‘information advantages’ to network structure. Tables 3, 4 and 6 together demonstrate that diverse networks provide access to diverse, non-redundant information, which in turn drives performance in information intensive work.

We also uncover evidence of alternative mechanisms linking network structure to performance. Table 6 shows network diversity is positively associated with performance even when holding access to novel information constant, providing preliminary evidence of additional benefits to network structure beyond those conferred through information advantage. Controlling for access to novel information, network diversity is associated with greater revenue generation in fixed effects and pooled OLS specifications, more completed projects in pooled OLS specifications, and with faster project completion in fixed effects specifications. These results leave open the possibility that some benefits to structural network diversity come not from access to novel, non-redundant information, but rather from other mechanisms, like access to job support or organizational leverage, not tied to the novelty of the *information* they receive.

5. Conclusion

We present some of the first empirical evidence on the relationship between network structure and the content of information flowing to and from actors in a network. We develop theory detailing how network structures enable information benefits with measurable performance implications, and build and validate an analytical model to measure the diversity of information in email communication. Our results demonstrate a relationship between network structure and information structure, and lend broad support to the argument that network structures drive performance through their impact on individuals’ access to diverse information.

Specifically, we find that: (1) the total amount of novel information and the diversity of information flowing to actors are increasing in actors’ network size and network diversity, while (2) the marginal increase in information diversity is decreasing in network size. We also find evidence of a fundamental tradeoff between network size and network diversity. Part of the explanation for the

Essay 2: Network Structure & Information Advantage

decreasing marginal contribution of network size to information diversity is that (3) network diversity is increasing in network size, but with diminishing marginal returns. As actors establish relationships with a finite set of possible contacts, the probability that a marginal relationship will be non-redundant, and provide access to novel information, decreases as possible alters in the network are exhausted. (4) Network diversity contributes to performance even controlling for the positive performance effects of access to novel information, suggesting additional benefits to network diversity beyond those conferred through information advantage. Surprisingly, (5) traditional demographic and human capital variables (e.g. age, gender, industry experience, education) have little effect on access to diverse information, highlighting the importance of network structure for information advantage. The methods and tools developed are replicable and can be readily applied to other settings in which email is widely used and available, opening a new frontier for the analysis of networks and information content.

Acknowledgments

We are grateful to the National Science Foundation (Career Award IIS-9876233 and grant IIS-0085725), Cisco Systems and the MIT Center for Digital Business for generous funding. We thank Tim Choe, Petch Manoharn and Jun Zhang for their tireless research assistance, and seminar participants at the Workshop on Information Systems Economics for valuable comments.⁵²

References

- Ancona, D.G. & Caldwell, D.F. 1992. Demography & Design: Predictors of new Product Team Performance. *Organization Science*, 3(3): 321-341.
- Aral, S., Brynjolfsson, E., & Van Alstyne, M. 2006. "Information, Technology and Information Worker Productivity: Task Level Evidence." *Proceedings of the 27th Annual International Conference on Information Systems*, Milwaukee, Wisconsin.
- Bernard, H.R., Killworth, P., & Sailor, L. 1981. "Summary of research on informant accuracy in network data and the reverse small world problem." *Connections*, (4:2): 11-25.

⁵² The original metrics and correlations in Appendix A and B were created from the email data by Petch Manoharn as part of his research assistantship and Master's thesis work under my supervision as well as that of Erik Brynjolfsson and Marshall Van Alstyne.

Essay 2: Network Structure & Information Advantage

- Blau, P. 1964. Exchange and Power in Social Life. J. Wiley Press, New York, NY.
- Burt, R. 1992. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA.
- Burt, R. 2000. "The network structure of social capital" In B. Staw, & Sutton, R. (Ed.), *Research in organizational behavior* (Vol. 22). New York, NY, JAI Press.
- Burt, R. 2004a. "Structural Holes & Good Ideas" *American Journal of Sociology*, (110): 349-99.
- Burt, R. 2004b. "Where to get a good idea: Steal it outside your group." As quoted by Michael Erard in *The New York Times*, May.
- Coleman, J.S. 1988. "Social Capital in the Creation of Human Capital" *American Journal of Sociology*, (94): S95-S120.
- Cook, K.S., Emerson, R.M., Gilmore, M.R., & Yamagishi, T. 1983. "The distribution of power in exchange networks." *American Journal of Sociology*, 89: 275-305.
- Cummings, J., & Cross, R. 2003. "Structural properties of work groups and their consequences for performance." *Social Networks*, 25(3):197-210.
- Emerson, R. 1962. "Power-Dependence Relations." *American Sociological Review*, 27: 31-41.
- Finlay, W. & Coverdill, J.E. 2000. "Risk, Opportunism & Structural Holes: How headhunters manage clients and earn fees." *Work & Occupations*, (27): 377-405.
- Gargiulo, M., M. Benassi, 2000. "Trapped in your own net? Network cohesion, structural holes, and the adaptation of social capital." *Organization Science* (11:2): 183-196.
- Granovetter, M. 1973. "The strength of weak ties." *American Journal of Sociology* (78):1360-80.
- Hansen, M. 1999. "The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits." *Administrative Science Quarterly* (44:1):82-111.
- Hansen, M. 2002. "Knowledge networks: Explaining effective knowledge sharing in multiunit companies." *Organization Science* (13:3): 232-248.
- Hargadon, A. & R, Sutton. 1997. "Technology brokering and innovation in a product development firm." *Administrative Science Quarterly*, (42): 716-49.
- Krackhardt, D. & Kilduff, M. 1999. "Whether close or far: Social distance effects on perceived balance in friendship networks." *Journal of personality and social psychology* (76) 770-82.
- Kumbasar, E., Romney, A.K., and Batchelder, W.H. 1994. Systematic biases in social perception. *American Journal of Sociology*, (100): 477-505.
- Marschak, J. & Radner, R. 1972. Economic Theory of Teams, Yale University Press, CT.
- Marsden, P. 1990. "Network Data & Measurement." *Annual Review of Sociology* (16): 435-463.
- McPherson, M., L. Smith-Lovin & J. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415-444.
- Podolny, J., Baron, J. 1997. "Resources and relationships: Social networks and mobility in the workplace." *American Sociological Review* (62:5): 673-693.
- Reagans, R. & McEvily, B. 2003. "Network Structure & Knowledge Transfer: The Effects of Cohesion & Range." *Administrative Science Quarterly*, (48): 240-67.
- Reagans, R. & Zuckerman, E. 2001. "Networks, diversity, and productivity: The social capital of corporate R&D teams." *Organization Science* (12:4): 502-517.

Essay 2: Network Structure & Information Advantage

- Reagans, R. & Zuckerman, E. 2006. "Why Knowledge Does Not Equal Power: The Network Redundancy Tradeoff" *Working Paper Sloan School of Management* 2006, pp. 1-67.
- Salton, G., Wong, A., & Yang, C. S. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM*, 18(11): 613-620.
- Sparrowe, R., Liden, R., Wayne, S., & Kraimer, M. 2001. "Social networks and the performance of individuals and groups." *Academy of Management Journal*, 44(2): 316-325.
- Szulanski, G. 1996. "Exploring internal stickiness: Impediments to the transfer of best practice within the firm." *Strategic Management Journal* (17): 27-43.
- Uzzi, B. 1996. "The sources and consequences of embeddedness for the economic performance of organizations: The network effect" *American Sociological Review*, (61):674-98.
- Uzzi, B. 1997. "Social structure and competition in interfirm networks: The paradox of embeddedness." *Administrative Science Quarterly*, 42: 35-67.
- Van Alstyne, M. & Zhang, J. 2003. "EmailNet: A System for Automatically Mining Social Networks from Organizational Email Communication," NAACSOS.
- White, H. 1980. "A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity." *Econometrica* (48:4): 817-838.

Appendix A. Descriptions & Correlations of Information Diversity Metrics
Metric
Description and Purpose

VarCos

Variance based on cosine distance (cosine similarity):

$$ID'_i = \frac{\sum_{j=1}^N (Cos(d'_{ij}, M'_i))^2}{N}, \text{ where } Cos(d_{ij}, M) = \frac{d_i \cdot M_i}{|d_i| |M_i|} = \frac{\sum_j w_{ij} \times w_{Mj}}{\sqrt{\sum w_{ij}^2} \sqrt{\sum w_{Mj}^2}}$$

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

VarDice

$$\text{Variance based on Dice's Distance and Dice's Coefficient: } VarDice'_i = \frac{\sum_{j=1}^N (DistDice(d'_{ij}))^2}{N},$$

where

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

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

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

AvgCommon

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

$$AvgCommon'_i = \frac{\sum_{j=1}^N (CommonDist(d'_{1j}, d'_{2j}))}{N}, \text{ where } (d'_{1j}, d'_{2j}) \text{ represents a given pair of}$$

 documents (1 and 2) in an inbox and j indexes all pairs of documents in an inbox, and where:

$$CommonDist(d'_{1j}, d'_{2j}) = 1 - CommonSim(d'_{1j}, d'_{2j})$$

$$CommonSim(d'_{1j}, d'_{2j}) = \frac{\sum Iterations_in_same_cluster}{\sum Iterations}$$

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

AvgCommonIC

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

Essay 2: Network Structure & Information Advantage

$$CommonICSim(D_1, D_2) = \frac{1}{\log\left(\frac{1}{\|all_documents\|}\right)} \cdot \frac{\sum_{D_1, D_2 \text{ in same bucket}} \log\left(\frac{\|documents_in_the_bucket\|}{\|all_documents\|}\right)}{total_number_of_bucket_levels}$$

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

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

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

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

$$DocBucketDist(D_1, D_2) = \frac{1}{\|cluster_iterations\|} \cdot \sum_{i \in cluster_iterations} (BucketDist(B_{iteration=i, D_1}, B_{iteration=i, D_2})),$$

and:

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

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

Correlations Between the Five Measures of Information Diversity					
Measure	1	2	3	4	5
1. VarCosSim	1.0000				
2. VarDiceSim	0.9999	1.0000			
3. AvgCommon	0.9855	0.9845	1.0000		
4. AvgCommonIC	0.9943	0.9937	0.9973	1.0000	
5. AvgBucDiff	0.9790	0.9778	0.9993	0.9939	1.0000

Appendix B. Wikipedia.org Categories

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

Essay 3: “Organizational Information Dynamics: Drivers of Information Diffusion in Organizations”

Abstract

We examine what drives the diffusion of different types of information through organizations. We ask: What predicts the likelihood of an individual becoming aware of a strategic piece of information, or becoming aware of it sooner? Do different types of information exhibit different diffusion patterns, and do different characteristics of social structure, relationships and individuals in turn affect access to different kinds of information? We hypothesize that the dual effects of content and structure jointly predict the diffusion path of a given piece of information. While one type of information may be more likely to diffuse upward through the organizational hierarchy or strictly across functional relationships, a different type of information may diffuse laterally or without regard to function or hierarchy. To test our hypotheses, we characterize the social network of a medium sized executive recruiting firm using accounting data on project co-work relationships and ten months of email traffic observed over two five month periods. We identify two distinct types of information diffusing over this network – ‘event news’ and ‘discussion topics’ – by their usage characteristics, and observe several thousand diffusion processes of each type of information from their original first use to their varied recipients over time. We then test the effects of network structure and functional and demographic characteristics of dyadic relationships on the likelihood of receiving each type of information and receiving it more quickly. Our results demonstrate that the diffusion of news, characterized by a spike in communication and rapid, pervasive diffusion through the organization, is influenced by demographic and network factors but not by functional relationships (e.g. prior co-work, authority) or the strength of ties. In contrast, diffusion of discussion topics, which exhibit more shallow diffusion characterized by ‘back-and-forth’ conversation, is heavily influenced by functional relationships and the strength of ties, as well as demographic and network factors. Discussion topics are more likely to diffuse vertically up and down the organizational hierarchy, across relationships with a prior working history, and across stronger ties, while news is more likely to diffuse laterally as well as vertically, and without regard to the strength or function of relationships. Our findings highlight the importance of simultaneous considerations of structure and content in information diffusion studies.

Keywords: Social Networks, Information Diffusion, Information Dynamics.

1. Introduction

The process of information diffusion through social groups lies at the heart of numerous phenomena in a variety of disciplines from finance, to marketing, to innovation. Theories on subjects as wide ranging as the diffusion of innovations (e.g. Rogers 1995), dynamic trading behavior (e.g. Hirshleifer et. al. 1994), and the mechanics of word of mouth marketing (e.g. Dellarocas 2003), rely on information diffusion as a central theoretical building block, making important assumptions about how information spreads between individuals. Timely access to strategic information, innovative ideas, or current news can also highlight hidden opportunities, provide negotiating leverage (Burt 1992), promote innovation (Hargadon & Sutton 1997, Burt 2004), and ultimately drive economic performance (Reagans & Zuckerman 2001, Hansen 2002, Aral, Brynjolfsson & Van Alstynne 2006). But, while theories based on information diffusion proliferate, empirical evidence on how information spreads through social groups remains scarce.

Diffusion studies typically observe adoption or purchase decisions rather than the diffusion of information itself. The studies that do focus on information diffusion are typically theoretical, or derive results from computer simulations of information passing among a handful of actors. Existing theory focuses mainly on which global social structures maximize diffusion, and although we know that transfers of certain types of information are easier than others (Von Hippel 1998), diffusion studies typically treat information as a uniform concept, making variation in diffusion patterns across different information types and social structures difficult to theorize. The apparent value of timely access to strategic information, which rests on the likelihood of seeing the information and seeing it sooner than others, gives rise to a natural set of questions about the dynamic movement of information through populations: How does information diffuse through a given social group? What makes someone more likely to be exposed to an idea as it spreads? Do different types of information diffuse differently?

To complement studies of the economic value of information, and provide evidence on how and when information diffuses in organizations, we study the movement of different types of information through one organization over a period of two years. We argue that the social nature of information

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

diffusion necessitates simultaneous examination of both the type of information and the type of social relationship or structure through which it diffuses. In organizations, the dual effect of content and structure jointly predict the diffusion path of a given piece of information. While one type of information may be more likely to diffuse upward through the organizational hierarchy or strictly across functional relationships, a different type of information may diffuse laterally or without regard to function or hierarchy.

To test our theory, we characterize the social network of a medium sized executive recruiting firm using ten months of email data observed over two five month periods and accounting data detailing project co-work relationships. We identify two distinct types of information diffusing over this network – ‘event news’ and ‘discussion topics’ – by their usage characteristics, and observe several thousand diffusion processes of each type of information from their original first use to their varied recipients over time. We then test the effects of network structure and functional and demographic characteristics of dyadic relationships and individuals on the likelihood of receiving each type of information and receiving it more quickly.

Our results demonstrate that the diffusion of news, characterized by a spike in communication and rapid, pervasive diffusion through the organization, is influenced by demographic and network factors but not by functional relationships (e.g. prior co-work, authority) or the strength of ties. In contrast, diffusion of discussion topics, which exhibit more shallow diffusion characterized by ‘back-and-forth’ conversation, is heavily influenced by functional relationships and the strength of ties, as well as demographic and network factors. Discussion topics are more likely to diffuse vertically up and down the organizational hierarchy, across relationships with a prior working history, and across stronger ties, while news is more likely to diffuse laterally as well as vertically, and without regard to the strength or function of relationships. Our findings highlight the importance of simultaneous considerations of structure and content in information diffusion studies.

2. Theory & Literature

2.1. The Central Role of Information in Diffusion Studies

Most diffusion studies take information diffusion as a central starting point from which to investigate the propagation of innovations or the potential success of targeted marketing campaigns. As a result, our understanding of underlying information dynamics that drive influence remains underdeveloped. Ironically, small variations in assumptions about how information spreads can drastically alter insights provided by diffusion models. For example, if news of an innovation spreads differently than knowledge about how to use it, or if status, authority or the functions of relationships moderate the flow of information from one individual or group to another, markedly different diffusion outcomes can be derived from the same basic propagation models. In this section we highlight the role of information in current diffusion studies and describe critical areas of knowledge that remain underdeveloped.

Theories of the diffusion of innovations (e.g. Rogers 1995), which since early studies of the diffusion of hybrid corn (Grilliches 1957, Ryan & Gross 1943) have been applied to phenomena as wide ranging as the spread of democracy (e.g. Wejnert 2002) and the adoption of technological innovations (e.g. Coleman, Katz & Menzel 1966), rely on information diffusion as a central mechanism driving adoption decisions. Potential adopters are exposed to new innovations and are convinced to adopt through “processes by which participants create and share information with one another in order to reach mutual understanding” (Rogers 1995: 17). As Rogers (1995: 17-18) describes, “the essence of the diffusion process is the information exchange through which an individual communicates a new idea to one or several others.”

Information diffusion drives word of mouth contagion and viral marketing. Much of this research is concerned with maximizing the spread of influence through a social network by identifying influential nodes likely to “trigger” pervasive information cascades (e.g. Domingos & Richardson 2001, Kempe, Kleinberg, Tardos 2003), or enumerating characteristics of information cascades, such as the empirical distributions of their depth and structure (e.g. Leskovec, Singh, Kleinberg 2006). For example, Leskovec, Singh & Kleinberg (2006: 1) find that cascades in online recommendation networks “tend to be shallow,

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

but occasionally large bursts of propagation appear” such that “the distribution of cascade sizes is approximately heavy-tailed.” They also find different cascade properties across different products (DVDs, books and music), and identify various structural cascade patterns that recur in their data. Agent based models also apply notions of information cascades to revenue forecasts for various product categories such as movie box office receipts (De Vany & Lee 2001).

Two fundamental models have emerged to explain the diffusion of influence in contagion and innovation diffusion: threshold models and cascade models. Threshold models posit that individuals adopt innovations after reaching and surpassing their own private “threshold” of influence (e.g. Granovetter 1978, Schelling 1978). As others close to them adopt an innovation, each subsequent adoption⁵³ brings an actor closer to the threshold of influence required for their personal adoption. Individuals may be “close” in terms of their physical proximity, their direct interactions, or the equivalence of their social status or roles (Burt 1987). Cascade models on the other hand posit that each time a proximate individual adopts, the focal actor adopts with some probability that is a function of their relationship (e.g. Kempe, Kleinberg, Tardos 2003).

In both threshold models and cascade models, information sharing encourages adoption by spreading awareness (Meyer & Rowan 1977, Dewar & Dutton 1986), facilitating mimetic pressures (e.g. Coleman et. al. 1966), or through direct peer-to-peer influence (e.g. Rogers 1995). Each of these mechanisms involves significant information exchanges between adopters and non-adopters about the existence of a given innovation or product (awareness), its contextual uses and advantages (influence), or positive signals from the adoption of others (imitation, mimetic pressure). However, these communication processes, which are themselves complex, varied and subject to a host of underlying social norms and structural constraints, are rarely theorized. For instance, a great deal less information is necessary to communicate the existence of an innovation than its relative costs and benefits. In certain circumstances information may flow freely, spreading through social interaction without regard to status or competition,

⁵³ Often weighted by some factor that describes why one proximate node is more influential than another.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

whereas in competitive or hierarchical environments the desire to maintain information advantages or authority may facilitate some information transfers while limiting others. Both threshold and cascade models assume an information transmission between adopters and non-adopters, but rarely specify the nature of the information or the conditions under which exchanges take place. Rather, the diffusion process is typically tested under various assumptions about the distribution of thresholds or dyadic adoption probabilities in the population. In fact, as Kempe, Kleinberg, Tardos (2003: 2) explain “the fact that [thresholds] are randomly selected is intended to model our lack of knowledge of their values.”⁵⁴

Information diffusion also underlies several well known theories of dynamical trading behavior in financial markets. Hirshleifer et. al. (1994) demonstrate that temporal asymmetries in the diffusion of information to various traders create abnormal profits for those that are informed (and informed early) and explain seemingly irrational trading equilibria, such as “herding” or outcomes based on “follow the leader” strategies, that seem to contradict rational private valuations. Since the consequences of financial trading decisions are actualized in relation to the collective decisions of the market, the receipt of market information and the exact timing of when investors become aware of a given piece of information “may be even more important than the accuracy of the information” in explaining profits and collective trading behaviors (Hirshleifer et. al. 1994: 1688). In this way, cascades of information that pulse through the population of financial traders can separate winners from losers, determine equilibria, and create collective tendencies that contradict privately held information. Yet, in these models temporal asymmetries in information acquisition are taken as given, and how and why these systematic asymmetries arise remains unknown.

Finally, there is a body of literature on knowledge transfers and performance, which explores relationships between network structure, knowledge sharing and the performance of teams (e.g. Reagans & Zuckerman 2001, Cummings 2004). These studies show that knowledge sharing and the network

⁵⁴ A related body of literature, stemming from Stanley Milgram’s famous study of the “small-world phenomenon,” studies network structures that support efficient search in local networks to explain the ability of actors to find short paths to targets through large networks with only local information (e.g. Watts 1999, Kleinberg 1999, 2001; Adamic et. al. 2001).

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

structure that guides information flows can impact the productivity and performance of work groups. However, most of this work remains “agnostic with respect to content” (Hansen 1999: 83) and only considers whether knowledge is flowing rather than the type of knowledge being transferred. A related literature examines the conditions under which knowledge and information flow efficiently between business units and individuals (e.g. Hansen 1999, 2002, Reagans and McEvily 2003), although this work focuses on dyadic transfers of information between business units or individuals rather than on the diffusion of information from an originator to all potential recipients in the organization.

While these related literatures highlight the importance of information dynamics and in particular information diffusion in organizations, there is scant research on the subject. Although processes and circumstances that govern the movement of information through populations form a core micro-level foundation on which diffusion models are based, little research examines enablers and constraints of the movement of information itself. In order to understand the underlying dynamics driving the diffusion of innovations, dynamical trading behavior, information cascades that empower viral marketing, and global properties of how information transfers in organizations affect individual and group performance, it is important to develop theory about organizational information dynamics – how information moves through social groups within and between organizations.

2.2. Information Dynamics in Organizations

In this section we develop theory about a limited subset of information dynamics: the combined effects of information content and social structure on the diffusion of information in organizations. More specifically, we focus on the diffusion of event news and discussion topics as a function of both dyadic and individual network, demographic, authority, hierarchy and co-work characteristics.

Although some information diffusion studies exist, they typically rely on computer simulations of a handful of agents (e.g. Yamaguchi 1994, Buskens & Yamaguchi 1999, Newman et. al. 2002, Reagans & Zuckerman 2006), treat information as a uniform, homogeneous concept (e.g. Buskens & Yamaguchi 1999, Wu et. al. 2004, Newman et. al. 2002, Reagans & Zuckerman 2006), and focus on global properties

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

that maximize the diffusion of a given piece of information (e.g. Newman et. al. 2002). While these studies provide an important starting point for nascent explorations of information dynamics, significant theoretical underpinnings remain unexplored.

Most current conceptualizations of information dynamics assume information is homogeneous, without separately theorizing different types of information. For example, Bushkens & Yamaguchi (1999) improve on Yamaguchi (1994) by introducing the assumption that information exchange is non-rival (that information is retained by the sender in an exchange) into their agent based model, but maintain the tradition of assuming information homogeneity. Reagans & Zuckerman (2006) model the passage of “bits” through different network structures and assume the “uniqueness” of each bit, but theorize that each bit passes between individuals with uniform probability and without regard to how the uniqueness of a bit is related to its transmission probability. Wu et. al. (2004) test decay processes in information diffusion using email data from employees at HP Labs. However, by studying the diffusion of attachments and links, they too assume different types of information diffuse uniformly.

The assumption of information homogeneity is problematic in light of prior evidence on differences in information transfer effectiveness across different types of information. Some information is simply “stickier” (Von Hippel 1998) and more difficult to transfer (Hansen 1999) due to its specificity (Rosenberg 1982, Nelson 1990), complexity (Uzzi 1996, 1997, Hansen 1999), the amount of related knowledge of the receiver (Cohen & Levinthal 1990, Hansen 2002), and the degree to which the information is declarative or procedural (Cohen & Bacdayan 1994). These factors make it unlikely that all types of information exhibit uniform transfer rates or diffusion patterns across different relationships or social structures. As Wu et. al. (2004: 328) point out: “There are ... differences between information flows and the spread of viruses. While viruses tend to be indiscriminate, infecting any susceptible individual, information is selective and passed by its host only to individuals the host thinks would be interested in it.” We take this departure from epidemic models of disease a step further. Many other factors can influence the diffusion of a given type of information beyond the senders’ perception of the receivers’ interest. We hypothesize that the strength and function of social relationships, geographic

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

proximity, organizational boundaries, and hierarchy, authority and status differences across social groups effect the movement of information, and have different effects across different types of information.

Finally, current work focuses more on global social network properties that maximize the diffusion of a given piece of information through a population (e.g. Watts & Stogatz 1998), than on drivers of individual access to information cascades. There is also a lack of robust empirical evidence on different types of information diffusing through real social groups. Although Wu et. al. (2003) study the diffusion of attachments and links in email, which are content agnostic, and Newman et. al. (2002) model the spread of computer viruses over an observed sample of email address books, neither of these works nor any other work we found examines empirical evidence on the diffusion patterns of different types of information and the predictors of access to those diffusion processes among individuals in an organization.

We therefore propose three extensions to the current body of work on information dynamics. First, we propose that in addition to global structural properties of social groups, there exist hierarchal, demographic and functional enablers and constraints of information diffusion. For example, we hypothesize that information may diffuse more readily vertically (or laterally) through an organizational hierarchy due to authority or status differences, or more quickly through functional relationships than strong ties per se. Second, we hypothesize that different types of information content diffuse differently. We argue that characteristics of information are pertinent to how it diffuses. Third, we argue that content and structure jointly predict the diffusion path of a given piece of information, and that different social and structural factors will govern the diffusion of different types of information. We hypothesize that simple, declarative news will diffuse relatively indiscriminately without regard to the strength of ties, the function of relationships, or authority; in contrast, the diffusion of discussion topics, characterized by shallow cascades and back and forth conversations, will be influenced by the strength of ties, the function of relationships and by authority and hierarchy. We develop the theory behind each of these extensions in the next two sections, and then test our theory by studying several thousand diffusion processes observed over two years in our email data.

2.2.1. Social Drivers of Organizational Information Diffusion

Several factors could influence the flow of information in organizations. Theories on how status hierarchies, demography, social networks and formal organizational structures facilitate and hinder relationship building guide our thinking on how information may diffuse. We also utilize semi-structured interview data from employees of our research site to identify potential drivers of information diffusion in our setting. In the end, we hypothesize four categories of factors that may impact information dynamics in organizations: demography, organizational hierarchy, tie and network characteristics and functional task characteristics. Each of these categories includes individual and dyadic dimensions of interest.

Demography. Individuals' demographic characteristics and dissimilarity are likely to affect social choices about information seeking and information transmission. Similar individuals tend to flock together in social relationships – a phenomenon known as homophily (McPherson, Smith-Loving, & Cook 2001), creating parity in perspectives, information and resources across demographically similar individuals in organizations (Burt 1992, Reagans & Zuckerman 2001). Demographic diversity has also been shown to introduce social divisions and create tension in organizational work groups (Pfeffer 1983), reducing the likelihood that individuals of dissimilar demographic backgrounds will go to each other for advice or pass information to one another. We therefore measure the demographic characteristics of individuals and the demographic dissimilarity of pairs of individuals at our research site, focusing on age, gender, and education, three of the most important variables in organizational demography.⁵⁵

Organizational Hierarchy. There is good reason to suspect that information flows are affected by formal organizational structures. Formal structures define reporting relationships and work dependencies that necessitate communication and coordination (Mintzberg 1979). Managers and employees frequently communicate to manage administrative tasks even when they are not working on the same projects, and the importance of notification for accountability, and recognition for upward mobility encourages

⁵⁵ Unfortunately, we do not have access to race or organizational tenure variables, although we do have measures of industry tenure which we consider a functional dimension of social distance rather than a demographic one.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

dialogue and information exchange along hierarchical lines. Embedded within formal organizational hierarchies are gradients of status and authority that may also guide information flows. Task level knowledge and familiarity with customers, specialized technology, competitors and new market opportunities often resides with mid-level managers rather than upper management, leading some firms to decentralize decision rights to take advantage of local knowledge (Dessein 2002). If this is the case, we might expect information to flow most quickly to employees in mid-level management positions. In our organization, teams are composed hierarchically, generally consisting of one partner, one consultant, and one researcher. As project teams are organized hierarchically, task related information is likely to flow vertically rather than laterally across individuals of the same organizational rank.

Tie & Network Characteristics. Informal networks are also likely to impact information diffusion in organizations. A vast literature treats the relationship between social network structure and organizational performance (e.g. Burt 1992, Gargiulo and Benassi 2000, Ahuja 2000, Sparrowe et al. 2001, Cummings & Cross 2003). Although most of this work does not measure information flows explicitly, evidence of a relationship between performance and network structure is typically assumed to be due in part to the information flowing between connected actors (Burt 1992, Reagans & Zuckerman 2001). Some network research does explicitly treat information transmission. For example, several studies have shown that the strength of ties increases the transfer of information and knowledge in dyadic relationships and in particular improves the effectiveness of transfers of tacit knowledge (Uzzi 1996, 1997). As individuals interact more frequently, they are likely to pass information to one another. We therefore measure the *strength of communication ties* by the total volume of email passing between each pair of individuals in our network. Other studies demonstrate that '*betweenness centrality*' $B(n_i)$ (Freeman 1979),⁵⁶ which measures the probability that the individual will fall on the shortest path between any two other individuals linked by email communication, predicts the total amount of knowledge acquired from other parts of the network (Hansen 1999), and that actors with high network

⁵⁶ Where g_{jk} is the number of geodesic paths linking j and k and $g_{jk}(n_i)$ is the number of geodesic paths linking j and k involving i .

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

constraint C_i (Burt 1992: 55),⁵⁷ which measures the degree to which an individual's contacts are connected to each other (a proxy for the redundancy of contacts), are less privy to new information (Burt 1992). We therefore measure individuals' betweenness centrality and their constraint as follows:

$$B(n_i) = \sum_{j < k} g_{jk}(n_i) / g_{jk};$$
$$C_i = \sum_j \left(p_{ij} + \sum_q p_{iq} p_{qj} \right)^2, \quad q \neq i, j.$$

The total amount of email communication that individuals engage in is also likely to drive their access to information and how quickly they see it (Cummings & Cross 2003). Finally, a great deal of evidence links physical proximity to communication between actors (e.g. Allen 1977), however, in the case of email communication, it could be that greater geographic distance is associated with more email communication between actors who find it costly to meet and communicate face to face. We therefore measure physical proximity by whether or not two people work together in one of the firm's fourteen offices.

Task Characteristics. Working relationships are conduit of communication and information flow. As individuals work together, they develop stronger bonds of trust and experience that encourage them to exchange information. Working relationships also necessitate exchange of task related information and create relatively stable and enduring ties that individuals rely on for advice on future projects. However, relationships can decay over time when they are not active (Burt 2002), and repeated relationships are more likely to create long term conduits through which information diffuses. We therefore measure the strength of project co-work relationships by the number of projects employees have worked on together. We also know from the literature on absorptive capacity (Cohen & Levinthal 1990) that related knowledge helps individuals consume new information, and individuals in related fields and of related expertise are more likely to swim in the same pools of information. We therefore also measure whether or

⁵⁷ Where $p_{ij} + \sum_q p_{iq} p_{qj}$ measures the proportion of i 's network contacts that directly or indirectly involve j and C_i sums this across all of i 's contacts.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

not employees work in the same expertise area in the firm. For instance, some recruiters focus on health care positions, or technology positions, as well as in certain regions. We measure specialization using a dummy variable whose value is one for employees working in the same expertise area. Finally, as with demographic distance, we expect information to diffuse more easily between employees in the same industry cohort. Pfeffer (1983) has noted the importance of organizational cohorts in maintaining lines of communication and organizational relationships. The same logic can be applied to industry cohorts. Employees with the same industry tenure have likely been through similar work related milestones and may already be familiar with each other through industry relationships. In addition, more experienced workers may rely on other experienced workers for information, while status and authority prevent the less experienced from sharing as much information across industry tenure gradients.

2.2.2. Dimensions of Information Content that Affect Diffusion

Apart from social context, characteristics of information itself are likely to affect diffusion patterns and the probability that a given individual sees a piece of information. Previous research has shown that certain types of information are “stickier” and have higher transfer costs (Von Hippel 1998), and several dimensions of information may influence when it is shared and how it diffuses. In what follows we describe two contrasting types of information we call ‘event news’ and ‘discussion topics,’ which serve as vignettes for comparison across information types. These vignettes are intended as archetypes, not mutually exclusive categories. Information is contextual – one may have lengthy discussions about event news or hear news about an important discussion. The point of illustrating these two archetypes is to evoke the underlying characteristics of information that are most likely correlated with what we eventually measure and test: the diffusion patterns and usage characteristics of particular words in email. Our contention is that information of the types described here are likely to diffuse in a certain way and that words that exhibit these diffusion patterns proxy for information content with the characteristics we describe. However, the relationship between the types of information described and the diffusion patterns observed is not critical to the argument. In the end, our goal is to demonstrate that

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

different characteristics of people, relationships and social structure affect access to different types of information with different aggregate diffusion patterns.

Event News. We define ‘event news’ as simple, declarative, factual information that is likely triggered by an external event and is of general interest to many people in the organization. In the context of our research site, employees may learn of forthcoming layoffs at a source company, a forthcoming change in company policy or a significant change in top management through a rapid pervasive information cascade that travels quickly and pervasively throughout the organization. Such information is likely to be simple, declarative and factual, informing recipients of an event that has or will soon take place. Such information is of general interest to all employees in the firm and is likely to be widely shared amongst many people and across organizational and hierarchical boundaries.

Discussion Topics. We define ‘discussion topics’ as more specific, complex, and procedural, characterized by back and forth discussion of interest to limited and specialized groups of people. At this firm, work groups discuss particular projects, and most frequently have back and forth discussion about particular candidates or clients. So, for instance, a particular candidate’s name may be discussed back and forth as their merits for a particular job are being considered. Teams specializing in filling nursing job vacancies in the south eastern United States may circulate names amongst other recruiters who specialize in the same type of job in the same region. When we presented this typology to an IT employee of a different firm, they immediately related to this category of information by giving the example of “ITIL” (Information Technology Infrastructure Library), a framework of best practice approaches for delivering IT services that has since received ISO accreditation. This informant indicated that the IT staff in her firm would frequently discuss “ITIL,” and that we would likely find reference to it in back and forth email discussions in her group, but that it was unlikely that anyone else in the firm would discuss it.

Theories of information transfer support our distinctions between event news and discussion topics. Hansen (1999) demonstrates that complex knowledge is more difficult and costly to transfer, and shows that in dyadic relationships, strong ties are necessary to effectively transfer complex knowledge. There is also a theoretical distinction made between declarative and procedural information (Cohen &

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

Bacdayan 1994: 557), with the former consisting of “facts, propositions and events,” and the later of information about how to accomplish tasks, activities or routines. We argue that event news is more likely to be simple and declarative, and thus more easily transferred widely amongst different types of people. Rosenberg (1976), Nelson (1990) and Von Hippel (1998) also make the distinction between “specific” and “generic” information and knowledge, arguing that, in contrast to the specific, “generic knowledge not only tends to be germane to a wide variety of uses and users. Such knowledge is the stock in trade of professionals in a field ... so that when new generic knowledge is created anywhere, it is relatively costless to communicate to other professionals” (Nelson 1990: 11-12, as quoted in Von Hippel 1998: 431).⁵⁸ Finally, transfers of information and knowledge are more effective among individuals with related knowledge (Cohen & Levinthal 1990, Hansen 2002). Those with similar expertise or specialization are more likely to share information more often and more effectively due not only to their shared common interest in certain information, but also their ability to more effectively communicate ideas based on their “common ground” (Cramton 1991). We therefore hypothesize that diffusion of event news, which is likely to be simple, declarative, factual and relevant to a wide variety of users, will be driven by the demographic and network factors theorized to constrain interactions due to homophily and network constraints.

H1: Access to event news is driven by demographic similarity, and structural characteristics of network position such as betweenness centrality, constraint and path length.

On the other hand, we argue that information passed back and forth amongst small groups is likely to be task specific and reflect information relevant to those socially and organizationally proximate to the originator. At our research site, since work groups are organized vertically along the organizational hierarchy, with teams composed of one member from each organizational level, we expect task related

⁵⁸ As pointed out by Orlikowski (2002), there is an important distinction to be made between knowledge and information; and in her case between knowledge and knowing. Without exploring the vast theoretical details of this distinction, we assume that the characteristics that make knowledge complex (and therefore costly to transfer) in turn influence characteristics of the information employees in this firm send and receive. Specifically, we argue that they are likely to *send* generic information to “a wide variety of users” (Nelson 1990: 11-12), without making the deeper assumption that what makes knowledge complex also makes information that communicates that knowledge complex.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

information to be passed vertically up and down the organizational hierarchy, rather than laterally between members of the same organizational level. We therefore hypothesize that diffusion of discussion topics, which are likely to be complex, specific, procedural and discussed in small groups, will be driven not only by demographic and network factors, but also by project co-work relationships and organizational hierarchy.

H2: Access to discussion topics is driven by demographic similarity, and structural characteristics of network position such as betweenness centrality, constraint and path length, as well as by task characteristics and organizational hierarchy.

3. Methods

3.1. Data

Data for this study come from three sources: (i) accounting data detailing project co-work relationships, organizational positions and physical locations, (ii) email data captured from the firm's corporate email server, and (iii) survey data that capture demographic characteristics, education, and industry tenure.

Email data cover 10 months of complete email history at the firm. The data were captured from the corporate mail server during two equal periods from October 1, 2002 to March 1, 2003 and from October 1, 2003 to March 1, 2004. We wrote and developed capture software specific to this project and took multiple steps to maximize data integrity and levels of participation. New code was tested at Microsoft Research Labs for server load, accuracy and completeness of message capture, and security exposure. To account for differences in user deletion patterns, we set administrative controls to prevent data expunging for 24 hours. The project went through nine months of human subjects review prior to launch and content was masked using cryptographic techniques to preserve individual privacy. Spam messages were excluded by eliminating external contacts who did not receive at least one message from someone inside the firm.⁵⁹ Participants received \$100 in exchange for permitting use of their data,

⁵⁹ In this study we focus on email sent to and from members of the firm due the difficulty of estimating accurate social network structures without access to whole network data (see Marsden 1990).

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

resulting in 87% coverage of recruiters eligible to participate and more than 125,000 email messages captured. Details of data collection are described by Aral, Brynjolfsson & Van Alstyne (2006). Since cryptographic techniques were used to protect privacy, we observe unique tokens for every word in the email data and construct diffusion metrics based on the movement of words through the organization in email. Methods for analyzing the diffusion of email content are described in greater detail in § 3.2.

Table 1: Descriptive Statistics

Variable	Obs.	Mean	SD	Min	Max
Gender (Male = 1)	832419	.50	.49	0	1
Age Difference	562650	12.22	8.81	0	39
Gender Difference	832419	.50	.49	0	1
Education Difference	562650	1.38	1.26	0	6
Email Volume	809613	1474.65	1129.95	0	4496
Strength of Tie	832419	11.71	36.90	0	464
Path Length	832419	2.61	2.68	0	10
Geographic Proximity (Same Office = 1)	832419	.30	.46	0	1
Friends in Common	832419	6.70	5.75	0	35
Betweenness Centrality	809613	36.77	36.81	0	165.73
Constraint	809613	.213	.09	0	.51
Prior Project Co-Work	832419	.26	1.33	0	19
Industry Tenure Difference	562650	10.08	8.32	0	38
Same Area Specialty	832419	.10	.30	0	1
Managerial Level Difference	832419	.86	.71	0	2
Partner	832419	.36	.48	0	1
Consultant	832419	.40	.48	0	1
Researcher	832419	.22	.41	0	1

Survey questions were generated from a review of relevant literature and interviews with recruiters. Experts in survey methods at the Inter-University Consortium for Political and Social Science Research vetted the survey instrument, which was then pre-tested for comprehension and ease-of-use. Individual participants received \$25 for completed surveys and participation exceeded 85%.

Our data collection avoids several important limitations of data used in previous studies of networks and information transfer (e.g. Hansen 1999, 2002), and information diffusion in particular. First, as noted in Essay 2, traditional network studies tradeoff exists comprehensive observation of whole networks and the accuracy of respondents' recall. Respondents are shown to have difficulty recalling their networks (e.g. Bernard et. al 1981), especially when assessing network connections among individuals socially distant to themselves (Krackhardt & Kilduff 1999). The inaccuracy of respondent recall and the

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

bias associated with recall at social distance creates inaccurate estimates of networks (Kumbasar, Romney & Batchelder 1994). The estimation challenge posed by this approach is that network metrics are incredibly sensitive to the completeness of data (Marsden 1990). Observation of whole networks in their entirety is important for creating unbiased estimates of actors' topological positions. By capturing network data in email we not only address the issue of respondent recall, we are able to capture almost every employee in the firm with a high degree of accuracy. 87% of the recruiters eligible for our study agreed to participate in email data collection, and given that our inability to observe the remaining 13% is limited to messages between two employees who both opted out of the study, we have a relatively unbiased view of the communication network with nearly full coverage of the firm.

Second, estimation of diffusion processes from incomplete data can be especially problematic (Greve, Tuma & Strang 2001). As we record the time that emails were sent and received employing time stamps in email data, we avoid time aggregation bias – where event times are recorded imprecisely or with systematic error. Time aggregation occurs when observations of sequential events are recorded as having occurred simultaneously due to coarse grained observations over large blocks of time like days, weeks or months. Previous research has shown that time aggregation creates severe bias when time intervals are large relative to the average time to an event (Petersen 1991, Petersen & Koput 1992, Greve, Tuma & Strang 2001). In contagion models, time aggregation also creates bias in contagion updating. As previous adopters can influence future adoption, imprecise time aggregation can inaccurately characterize time dependent influences on actors. Also, without relatively complete coverage of the population, sparse sampling can create significant bias in estimates of the impact of relational variables such as social proximity. When the sampling probability is small, coefficient estimates can be bias downward to zero, and the extent of the bias increases as the sampling probability decreases (Greve, Tuma & Strang 2001). Our near complete coverage of the firm and the precision with which we record event times in email data help us avoid these sources of bias.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

Table 2: Correlation Matrix

Measure	1	2	3	4	5	6	7	8	9	10	11	12
1. Gender (Male = 1)	1.00											
2. Age Difference	.06	1.00										
3. Gender Difference	.27	.06	1.00									
4. Education Difference	.08	.09	.06	1.00								
5. Email Volume	-.11	-.09	-.04	-.04	1.00							
6. Strength of Tie	-.04	-.06	-.01	.01	.30	1.00						
7. Path Length	-.11	.00	.00	-.04	-.37	-.18	1.00					
8. Geographic Proximity	-.06	.09	-.01	-.03	.06	.17	-.06	1.00				
9. Friends in Common	.04	-.02	.03	.02	.50	.38	-.40	.08	1.00			
10. Betweenness Centrality	.06	.01	.01	-.07	.66	.22	-.31	.03	.48	1.00		
11. Constraint	-.18	-.05	-.05	.01	-.26	-.06	.54	-.05	-.34	-.33	1.00	
12. Prior Project Co-Work	.02	-.01	.01	-.01	.01	.37	-.09	.08	.15	.02	-.06	1.00
13. Industry Tenure Difference	.11	.50	.06	-.05	-.09	-.08	.03	.07	-.07	-.08	-.12	.05
14. Same Area Specialty	-.01	-.16	-.00	.01	.12	.40	-.12	.27	.19	.05	-.04	.33
15. Managerial Level Difference	.05	.52	.05	.11	.03	-.10	.01	-.03	.01	.02	-.07	.03
16. Partner	.21	.06	.06	.02	-.06	-.05	-.14	-.06	.07	-.03	-.31	.09
17. Consultant	-.12	-.07	-.03	-.04	-.31	-.09	.29	-.26	-.19	-.22	.19	-.01
18. Researcher	-.09	.01	-.03	.03	.40	.15	-.17	.35	.13	.27	.12	-.08
	13	14	15	16	17	18						
13. Industry Tenure Difference	1.00											
14. Same Area Specialty	-.12	1.00										
15. Managerial Level Difference	.50	-.21	1.00									
16. Partner	.26	-.05	.23	1.00								
17. Consultant	-.13	-.07	-.21	-.56	1.00							
18. Researcher	-.13	.14	-.01	-.44	-.51	1.00						

3.2. Identifying Heterogeneous Information Types

Our goal is to identify samples of words that exhibit different diffusion patterns and whose usage characteristics reflect those one would expect to find exhibited by event news and discussion topics. We defined ‘event news’ as simple, declarative, factual information that is likely triggered by an external event and of general interest to many people in the firm. Given these criteria, we assume event news is characterized by a spike in activity and a rapid pervasive diffusion to members of the organization, followed by a decline in use. More accurately, we assume that words that exhibit these usage characteristics are likely to be event news or information whose theoretical characteristics are similar to those described for event news. At the same time we are interested in identifying a sample of ‘discussion topics,’ which we define as more complex, specific to a group of people, containing more procedural information and in Von Hippel’s (1998) parlance “sticky.” We expect this information to exhibit more shallow diffusion, characterized by ‘back-and-forth’ conversation among smaller groups for more extended periods. In this section we describe our analytical method for identifying these two types of information.⁶⁰

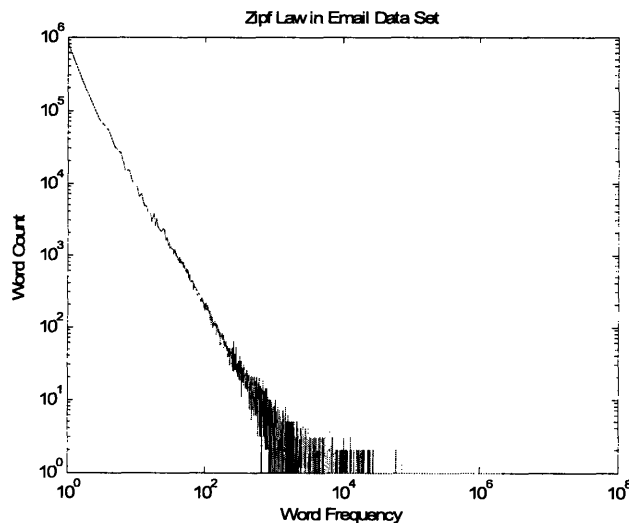


Figure 1. Distribution of Word Frequencies in Email Data

⁶⁰ We thank Tim Choe for his tireless coding efforts that extracted and manipulated the email data described in this section, and created the graphics depicting email discussions. This work was done as part of his research assistantship and Master’s thesis work under my supervision as well as that of Erik Brynjolfsson and Marshall Van Alstyne.

We began with a dataset consisting of approximately 1.5 million words whose frequencies were distributed according to the standard Zipf's Law distribution, with many highly infrequent words and smaller and smaller numbers of more frequent words, as shown in Figure 1. We initially eliminated words unlikely to cascade through the firm by culling the most infrequent words (term frequency < 11), words that are commonly used every week and are likely to be common language (words in at least one email in every week of the observation period), and words with low term frequency- cumulative inverse document frequency (tf-cidf), a common metric used to identify spikes in usage (Gruhl et. al. 2004).⁶¹ The tf-cidf constraint chooses words that record a spike in weekly usage greater than three times the previous weekly average, retaining words likely to cascade or diffuse. Together, these three methods reduced the number of emails under consideration to approximately 500,000, 495,000, and 120,000 words respectively. From these 120,000 candidates we sampled words likely to be event news and discussion topics.

In selecting event news, we sought words whose usage was characterized by a spike in activity and a rapid, pervasive diffusion to members of the organization, followed by a decline in use. To chose such a sample we chose words seen by more than 30 people with a coefficient of variation one standard deviation above the mean. We chose the 30 person threshold by first determining the percentage of the firm that used common words.

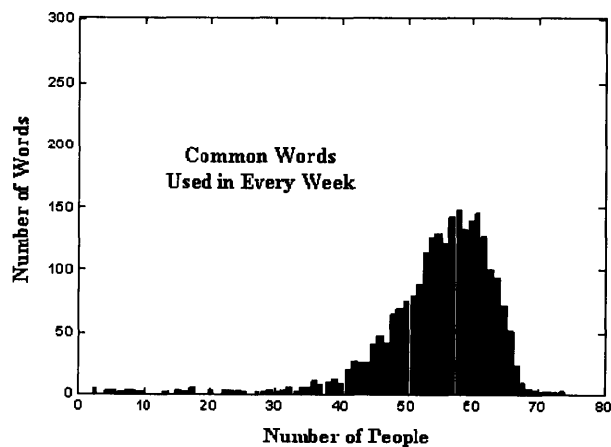


Figure 2. Distribution of Common Words over Employees

⁶¹ Our use of the cutoff of 11 produced similar results as cutoffs in the neighborhood of 11.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

The distribution of employees using common words provides a robust contextual proxy for information that is 'widely used' in the firm. By examining a histogram of the distribution of the number of common words over the number of people who used those words, we determined that most common words were used by between 30 and 70 people. To be conservative, we selected any word seen by more than 30 people as a potential observation of event news.

In order to select words likely to display rapid propagation, of those words that reached 30 people or more, we selected words with a high coefficient of variation of activity - words with bursts of activity in some weeks relative to others. The coefficient of variation has been used in previous work to identify spikes in topic frequency in blog posts (Gruhl et. al. 2004) and is a good measure of dispersion across data with heterogeneous mean values (Ancona & Caldwell 1992).⁶² Observations of a large number of people suddenly using a word much more frequently than usual are likely to indicate information triggered by some external event that is diffusing through the organization.⁶³ We therefore select words with a coefficient of variation one standard deviation above the mean. The result is a sample of 3275 words that are at first rarely used, and then suddenly are used much more frequently and by more than 30 people in the firm, followed by a decline in use. An example of an event news item is shown in Figure 3. The blue line represents the cumulative number of people to have seen the word in email, while the green line represents the frequency of use in email. As the figure shows, the word is rarely used and seen by less than 5 people in the first 80 days of the observation period, after which there is spike in activity accompanied by diffusion to nearly 60 people, followed by a decline in use. This example is descriptive of the words in our event news sample.

⁶² The coefficient of variation is simply the standard deviation of the number of emails per week that contain a given word divided by the mean number of emails per week that contain that word.

⁶³ While the argument could be made that a spike of activity is no guarantee of diffusion, occurrences of significant numbers of event driven spikes in usage that are not part of diffusion processes are only likely to downward bias estimates of the influence of relationship based metrics on the likelihood of seeing a given piece of information, making our estimates more conservative.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

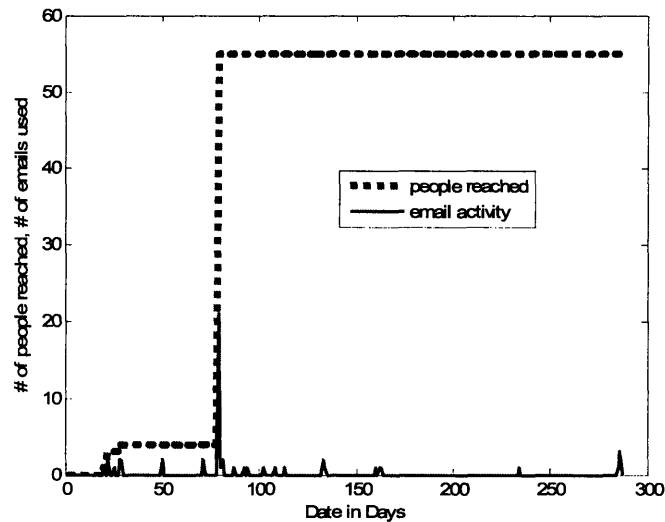


Figure 3. An Example Event News Item

We then selected a sample of discussion topic words. We began by identifying individual information cascades by tracing the flow of words through the organization. By ordering email time stamps and recording emails as they are sent and received, we constructed subgraphs of the network that traced a given word. As we were looking for discussion, we looked for words that were used in back and forth discussions via emails. We recorded a link if the receiver sent the word on within a small time window after receiving it to exclude words that were simply used later rather than received and sent on. This process created tree like subgraphs that recorded the path of a word diffusing from person to person. We characterized these subgraphs by their depth and their average depth – depth recording the number of people and average depth recording the sum of the depths of all connected subgraphs for a word divided by the total number of connected subgraphs. This process yielded approximately 3000 words with an average depth greater than 1.5. While this process identified words likely to be discussion topics, we were concerned that selecting our sample based on the same data that generated some of our social network graphs, would introduce endogeneity into our statistical specifications by ‘selecting on the dependent variable.’ We therefore chose a more parsimonious approach that did not use links between people in email as a selection criterion. We simply selected words where users both received and sent the word in

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

email. This simple criterion selected approximately 4100 words from the original candidate set. An example of the usage characteristics of discussion topic words is shown in Figure 4.

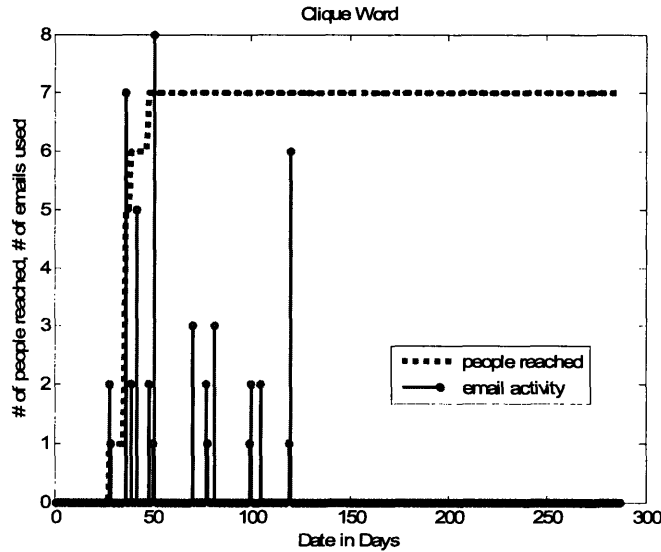


Figure 4. An Example Discussion Topic Item

Words in this sample display a lack of use, followed by a shallow diffusion to a limited number of people accompanied by an extended back and forth discussion, which in the case of the word show in Figure 4 lasts close to 3 months. These words are shared in back and forth conversation as demonstrated by the subgraphs of discussion topics shown in Figure 5.

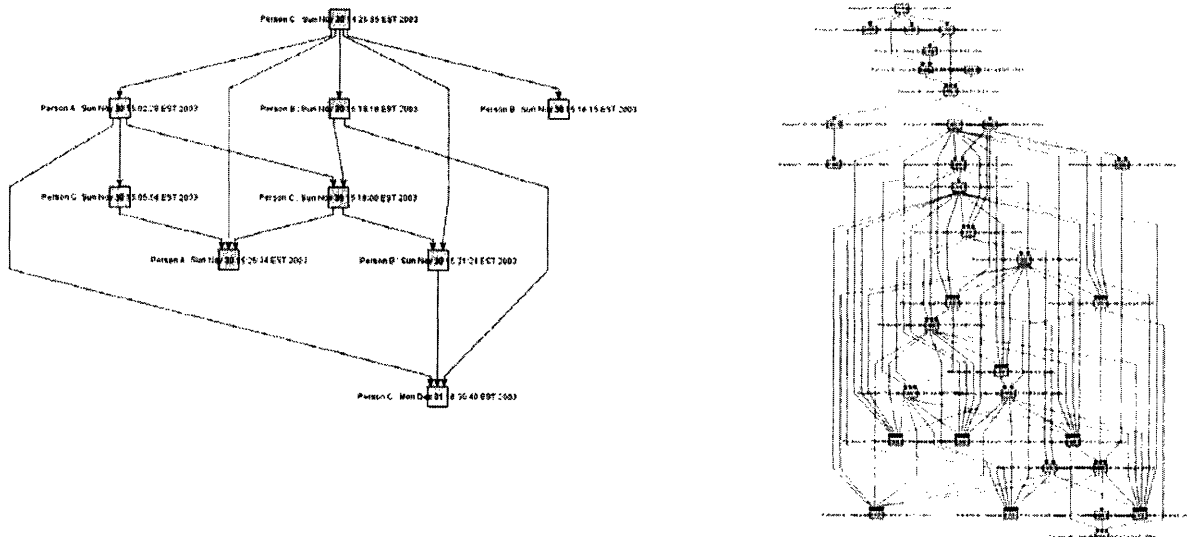


Figure 5. Discussion Paths in Discussion Topic Items

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

In the subgraph on the left, three people share a word back and forth in discussion. Person C sends the word to person A and person B, who each reply to person C, who in turn replies to them and subsequently receives a reply back from both B and C. The subgraph on the right shows a more complex back and forth discussion. This type of discussion, rather than exemplifying information that individuals disseminate widely among members of the firm, is likely specific to the individuals in this exchange. It is less likely to be declarative than procedural in the sense that those who receive the information react to it with a reply that triggers further ongoing discussion. In contrast to information that diffuses widely across the organization without back and forth discussion, we expect access to this type of information to be driven by functional relationships, strong ties and to move up and down the organizational hierarchy in line with the composition of teams which themselves are organized hierarchically.

After selecting these words based on their usage characteristics, we tested whether our information types exhibited significantly different usage characteristics and diffusion properties. As Leskovec, Singh & Kleinberg (2006) have noted, information cascades are typically shallow, but sometimes characterized by large bursts of wide propagation. We wanted to make sure we captured both these phenomena in our data. We therefore summarized the usage characteristics of words along several dimensions including the number of emails containing the word, the number of people who used the word, the coefficient of variation of use, the number of emails per person that contain the word, the total diffusion time divided by the total time in use (as a proxy for use beyond the diffusion to new users), and the maximum number of people who see the word for the first time in a given day (a proxy for the maximum spike in activity). We then tested whether words in each category differed significantly across these dimensions by conducting t-tests of the differences of means across information types and dimensions. Table 3 lists the mean usage characteristics and diffusion properties of both event news and discussion topics. Our t-tests demonstrate that these information types differ significantly across all dimensions of interest related to their use and diffusion. Results of t-tests of the difference of sample means for relevant dimensions are shown in Column 3.

Table 3: Mean Usage Characteristics and Diffusion Properties of Information Types

Information Type	News	Discussion	t-statistic
<i>Usage Characteristics & Diffusion Properties</i>			
Number of Words	3235	4168	-
Potential Diffusion Events	245280	320470	-
Realized Diffusion Events	65145	9344	-
Number of Emails	236.21	17.69	27.69***
Mean Diffusion Depth	36.31	2.48	213.28***
Coefficient of Variation	1.46	4.11	90.53***
Emails Per Person	6.10	7.47	1.105***
Diffusion Time / Total Use Time	.97	.48	66.36***
Maximum New Users Per Day	9.38	1.60	61.51***

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

3.3. Data Structure

We observe the diffusion of several thousand observations of each type of information from the original first use, which we define as the first occurrence of a given word in our data, to all employees in our sample. For each piece of information we observe whether a given employee received the word, the rank order in which they received the word relative to other employees, and the time between the first use of the word and the receipt of the word by each employee, constructed using the time stamps in email traffic. An observation, therefore is indexed by a word-recipient pair (one for each possible recipient in the firm) in which the first user is suppressed due to the one to one relationship between words and first users. For each word, our data log dyadic characteristics of each first user-recipient pair, such as the difference in their ages or industry tenures, for all potential recipients. Each observation also records individual characteristics of potential recipients, such as their gender, network position, or managerial level.

3.4. Statistical Specifications

We are interested in estimating the impact of hypothesized factors on the likelihood of seeing a strategic piece of information and seeing it sooner. Linear regression models are problematic when estimating temporal processes or likelihood outcomes for several reasons. Linear estimates of probabilistic outcomes create bias due to non-linearity near the upper and lower bounds of the likelihoods of discrete events, are not well suited to temporal processes in which outcome variables can be

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

conditioned on previous events (Strang & Tuma 1993), and produce biased estimates of longitudinal data in which right censoring is present and pervasive (Tuma & Hannan 1984). For these reasons we specify logistic regression and hazard rate models of the diffusion of different kinds of information in our data.

We first estimate the influence of independent variables on the likelihood of receiving a given piece of information using a standard logistic regression model formalized in equation 1.

$$\ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \alpha_i + \sum \beta_j X_i + \varepsilon_i \quad [1].$$

In this model, we estimate the impact of independent variables of interest on the likelihood of receiving a given piece of information by email. Since many of the variables of interest are time invariant characteristics of individuals and relationships and because our email data overlap project accounting data by ten months, we group observations in email and in accounting data into a pooled logistic regression. The coefficient estimates describe the impact of a given variable on the likelihood of receiving the word during the ten months of email observation.

The logistic regression model estimates the likelihood of receiving information; however, one limitation of this type of model is that they are not well suited to the estimation of dynamic processes in which the ordered timing of events matters. In particular, in diffusion studies cross sectional estimates may wash away temporal variation and allow later events to influence the estimates of earlier diffusion (Strang & Tuma 1993, Burt 1987). We therefore estimate the rate of receipt of different types of information conditional on having received the information, using a Cox proportional hazard rate model of the speed with which employees receive information:

$$R(t) = r(t)^b e^{\beta X} \quad [2].$$

where $R(t)$ represents the project completion rate, t is project time in the risk set, and $r(t)^b$ the baseline completion rate. The effects of independent variables are specified in the exponential power, where β is a vector of estimated coefficients and X is a vector of independent variables. The coefficients in this model have a straightforward interpretation: β represents the percent increase or decrease in the

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

rate at which information is seen associated with a one unit increase in the independent variable.

Coefficients greater than 1 represent an increase in the rate of information diffusing to the receiver (equal to $\beta - 1$); coefficients less than 1 represent a decrease (equal to $1 - \beta$). In testing the proportional hazards assumption we found no compelling evidence of duration dependence in any variables and proceeded with traditional estimations of the Cox model.

There has been some debate regarding appropriate estimation strategies for non-linear models with group specific parameters (e.g. Chamberlain 1980, Hausman et. al. 1984, Greene 2001, Green, Kim & Yoon 2001, Beck & Katz 2001). Group specific parameters arise in hierarchical data when multiple observations are sampled from the same group, or in longitudinal or time-series cross-section panel data with repeated observations of the same individual. For example, samples of students drawn randomly from several school districts will exhibit group specific effects at the district level, just as repeated observations of individuals will exhibit time invariant characteristics specific to the individual in time-series cross-section panel data or longitudinal event analyses. Our goal is to estimate a parameter vector common to all groups, while controlling for omitted variables that are consistent with a particular group (Chamberlain 1980).

In linear models, a conservative approach utilizes the fixed effects within estimator to control for time invariant characteristics of individuals or groups. However, fixed effects estimation is problematic in non-linear specifications such as ours. In linear models, using the group mean as a sufficient statistic to estimate the fixed effect allows consistent estimation of the parameter vector holding the group effect constant. In non-linear models, since the density of the observed random dependent variable is assumed to be fully defined, maximum likelihood is the most appropriate estimation strategy. However, in this framework, since the parameter vector β is estimated as a function of the group or fixed effect, and since the variance of the group effect does not converge to zero, “the maximum likelihood estimator of β is a function of a random variable which does not converge to a constant as $N \rightarrow \infty$ ” (Greene 2001: 7), making maximum likelihood estimation problematic. There are also issues of small sample bias. Hsiao

(1996) demonstrates that in short panels and small samples, the bias created by fixed effects specifications of non-linear models can reach 100%. Even in larger samples, the bias persists if the panel is short (Greene 2001), and remains significant in longer panels and samples as large as 100 unique observations (Heckman 1981, Greene 2001). In addition, in the case of binary outcome variables, such as the likelihood of receiving information, fixed effect models ignore characteristics of observations that never achieve the observed outcome, focusing estimation toward a comparison of independent variables among observations that do achieve the outcome (Beck & Katz 2001). This not only creates downward bias in estimates of infrequent events, it prevents estimation of variables of interest that do not change during the observation period, which make up a significant portion of the variation of interest in our models. Finally, the computational difficulty of maximizing functions with a large numbers of fixed parameters makes non-linear fixed effects models difficult in practice.⁶⁴

For these reasons, rather than employing fixed effects estimates, we address the group effect in our data by attempting to fully specify our models to control for the theoretically justifiable factors that could affect the flow of information. In doing so, we isolate estimates of our variables of interest holding constant most of the alternate explanations that may exist. We also cluster standard errors around individual recipients in order to maintain conservative estimates of the confidence intervals in light of the repeated observation of individuals. Although a host of hypotheses could be postulated about unobservable characteristics of individuals that make them more likely to receive information, our specifications include many of these possibilities. Given that we estimate many interesting independent variables that are constant (or nearly constant) over time and dyads, our conservative approach is likely to produce results robust to most alternate explanations of group effects without “throwing out the baby with the bath water” (Beck & Katz 2001).

4. Results

⁶⁴ Although brute force methods are becoming more feasible (see Greene (2001) for a discussion).

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

Our results demonstrate that demography, organizational hierarchy, tie and network characteristics and functional task characteristics all significantly influence the diffusion of information at our research site. However, different factors impact the diffusion of different types of information differently. The diffusion of event news is influenced by demographic and network factors but not by functional relationships (e.g. prior co-work, authority) or the strength of ties. In contrast, diffusion of discussion topics is heavily influenced by functional relationships and the strength of ties, as well as demographic and network factors. Discussion topics are more likely to diffuse vertically up and down the organizational hierarchy, across relationships with a prior working history, and across stronger ties, while news is more likely to diffuse laterally as well as vertically, and without regard to the strength or function of relationships.

4.1. Estimation of the Diffusion of Information

We first tested the diffusion of all types of information through the firm. Table 3 presents the results of logistic regression and hazard rate model estimates of the likelihood of receiving information and the rate at which different types of information diffuse to different users. Model 1 presents the results of the logistic regression estimating factors that influence the likelihood of receiving information. Although employment at the firm is balanced along gender lines, and controlling for the potential effect of an uneven distribution of men and women in higher level positions in the firm (partner and consultant dummies), men are 55% more likely than women to receive information of all types. Demographic dissimilarity between the originator of the information and the eventual recipient reduces the likelihood of receiving information by between 1% and 13%, with gender differences recording the largest magnitude impact and age differences the smallest. Higher communication volume makes one more likely to see information, although we use this as a control variable rather than a variable of interest simply because we only measure information diffusion in email. The strength of ties between originator and recipient increases the likelihood of receiving information. Ten additional emails sent between originator and recipient increases the likelihood that a diffusion process started by the originator reaches the recipient by

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

2%. Path length reduces the likelihood of receiving information, with each additional hop reducing the average likelihood of being involved diffusion by 29%. Having friends in common with the first user of a word seems to reduce the likelihood of receiving an information cascade that originates from that user – a strange result. However, having friends in common is positively correlated with email volume and the strength of ties. Holding these other variables constant, the positive effects of friends in common reduce and reverse. Betweenness centrality has a strong positive effect on the likelihood of receiving information, as do stronger project co-work relationships.

Table 3. Drivers of Access to Information		
	Model 1	Model 2
<i>Dependent Variable:</i>	Word Received	Rate of Receipt
<i>Specification (Coefficient Reported)</i>	<i>Logistic (Odds Ratio)</i>	<i>Hazard Model (Hazard Ratio)</i>
<i>Demography¹</i>		
Gender Dummy (Male = 1)	1.551 (.219)***	1.236 (.167)
<i>Demographic Distance</i>		
Age Difference	.986 (.004)***	.996 (.004)
Gender Difference	.869 (.014)***	1.009 (.010)
Education Difference	.906 (.023)***	.971 (.020)
<i>Tie & Network Characteristics</i>		
Communication Volume (Total Email)	1.0002 (.0002)**	1.000 (.000)
Strength of Tie	1.002 (.001)***	1.000 (.000)
Path Length	.711 (.047)***	.828 (.033)***
Geographic Proximity (Same Office = 1)	.857 (.088)	.865 (.078)
Friends in Common	.954 (.007)***	.992 (.005)
Betweenness Centrality	1.005 (.002)**	1.004 (.002)**
Constraint	.212 (.225)	.326 (.389)
<i>Task Characteristics</i>		
Prior Project Co-Work	1.042 (.016)***	1.031 (.012)**
Industry Tenure Difference	.996 (.006)	1.002 (.006)
Same Area Specialty	.883 (.080)	.983 (.067)
Managerial Level Difference	.951 (.038)	.997 (.033)
Partner Dummy	.933 (.188)	1.062 (.168)
Consultant Dummy	.870 (.184)	1.118 (.207)
<i>Word Type</i>		
Common Information	3.209 (.056)***	2.292 (.065)***
Discussion Topics	.081 (.008)***	.025 (.002)***
Log Pseudolikelihood	-234204.48	-1694852.4
Wald χ^2 (d.f.)	6264.80 (19)***	8878.76 (19)***
Pseudo R ²	.28	-
Observations	543308	462422
Notes: 1. Age, Edu, Industry Tenure N.S.		

The hazard rate model estimates of the drivers of the rate of information receipt reveal positive effects for project co-work and betweenness centrality, and a negative relationship between path length and

the rate at which information is received.⁶⁵ These results demonstrate the importance of demographic distance, network structure and project based working relationships on the likelihood of receiving information and the rate at which it is received.

4.2. Estimation of the Diffusion of Discussion Topics & Event News

Table 4 presents the results of logistic regression and hazard rate model estimates of the drivers of event news and discussion topic diffusion. Models 1 and 2 report the results of logistic regressions. The results demonstrate that demographic distance reduces the likelihood of receiving both news and discussion topics although with a slightly larger impact for news. The coefficient on the education difference parameter for instance indicates that one additional year of education difference between two individuals reduces the likelihood that news will diffuse between them by 7.5%, while the same one year difference in education reduces the likelihood of discussion topics diffusing between them by nearly 17%. Interestingly, men are over 50% more likely to see news information than women although gender has no effect on the likelihood of the diffusion of discussion topics. Overall demographic distance slows information diffusion.

Coefficients of tie and network characteristics also tell an interesting story. Strong ties are important predictors of the diffusion of discussion topics but not of news. News seems to diffuse pervasively throughout the organization without regard to the strength of ties – information of general interest is passed through relatively weak ties as well. The parameter estimate of the strength of ties on discussion diffusion in Model 2 indicate that ten additional emails exchanged between two people increases the likelihood that discussion topics will diffuse between them by 7% on average. Path length, which measures the number of nodes separating employees in the email network, reduces the likelihood of information diffusion, although the impact is much larger for discussion topics than for news. An

⁶⁵ The dummy variables for word type, which control for common information, news, and discussion topics show that common information is more pervasive and appears at a faster rate among employees than news (the omitted category), while discussion topics are much less likely to be seen and diffuse at a much slower rate.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

additional hop between individuals reduces the likelihood of information diffusion by 97%, indicating discussion topics almost always only diffuse to people who know each other directly, whereas news may travel across multiple hops to reach a receiver. Geographic proximity has no effect on diffusion although the parameter estimates are less than one, supporting the proposition that exchanges of information over email are less likely for co-located than geographically distant employees in our firm. Having friends in common with the first user of a word again reduces the likelihood of receiving information that originates from that user. Betweenness centrality is also positively associated with the likelihood of seeing both news and discussion topics, while constraint is not significant.

Table 4. Drivers of Access to Discussion Topics & Event News

	NEWS	DISCUSSION	NEWS	DISCUSSION
	Model 1	Model 2	Model 3	Model 4
<i>Dependent Variable:</i>	Word Received	Word Received	Rate of Receipt	Rate of Receipt
<i>Specification (Coefficient)</i>	<i>Logistic (Odds Ratio)</i>	<i>Logistic (Odds Ratio)</i>	<i>Hazard Model (Hazard Ratio)</i>	<i>Hazard Model (Hazard Ratio)</i>
<i>Demography¹</i>				
Gender (Male=1)	1.544 (.227)***	1.073 (.137)	1.332 (.228)*	1.075 (.162)
<i>Demographic Distance</i>				
Age Difference	.992 (.004)**	.981 (.007)***	.998 (.004)	.994 (.007)
Gender Difference	.902 (.017)***	.814 (.069)**	1.007 (.012)	1.092 (.110)
Education Difference	.925 (.022)***	.832 (.034)***	.966 (.024)	1.013 (.037)
<i>Tie & Network Characteristics</i>				
Email Volume	1.0001 (.00007)*	1.0001 (.0001)*	1.0001 (.000)	1.0001 (.000)**
Strength of Tie	1.000 (.000)	1.007 (.001)***	.999 (.000)	1.006 (.001)***
Path Length	.732 (.041)***	.029 (.005)***	.814 (.044)***	.310 (.045)***
Geographic Proximity	.883 (.090)	.929 (.106)	.879 (.097)	.993 (.115)
Friends in Common	.972 (.005)***	.877 (.012)***	.992 (.007)	.969 (.012)**
Betweenness Centrality	1.004 (.002)*	1.007 (.002)**	1.006 (.002)**	1.002 (.002)
Constraint	.186 (.213)	2.243 (2.651)	.282 (.410)	1.664 (1.698)
<i>Task Characteristics</i>				
Prior Project Co-Work	1.010 (.014)	1.080 (.0185)***	1.018 (.016)	1.066 (.018)***
Industry Tenure Difference	.996 (.006)	.978 (.008)**	.999 (.008)	.999 (.008)
Same Area Specialty	.933 (.073)	1.038 (.139)	.981 (.078)	1.795 (.252)***
<i>Organizational Hierarchy</i>				
Managerial Level Difference	.963 (.035)	1.138 (.079)*	.992 (.037)	1.097 (.089)
Partner Dummy	.856 (.186)	1.515 (.271)**	1.084 (.216)	1.411 (.232)**
Consultant Dummy	.798 (.177)	1.659 (.262)***	1.221 (.289)	1.749 (.288)***
Log Pseudolikelihood	-93273.148	-15167.79	-508288.77	-28166.432
Wald χ^2 (d.f.)	204.39 (17) ***	2816.61 (17)***	92.80 (17)***	762.33 (17)***
Pseudo R ²	.06	.54	-	-
Observations	163135	202500	120197	196541

Notes: ¹ We controlled for age and education, but these variables were never significant and did not contribute to the explanatory power of the model. Geographic Proximity: Same Office = 1; * p < .05; ** p < .01; *** p < .001.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

Perhaps most interesting are the effects of task characteristics and organizational hierarchy. Strong working relationships and similarity in industry tenure both have strong positive impacts on the likelihood of receiving discussion topics, but not on the diffusion of news. Each additional project that two people work on together increases the likelihood that discussion diffuses between them by 8%. The coefficient on managerial level difference indicates that discussion topics are more likely to diffuse vertically, up and down the organizational hierarchy rather than laterally across individuals of the same organizational rank. In our firm, since project teams and work flow are organized vertically, with one member from each organizational level represented on most teams and no teams of either three partners, three consultants or three researchers, discussion topics, which we propose represent more specific, procedural and complex information, typically move between organizational levels rather than across them. The coefficients on the partner and consultant dummy add more clarity to this result. As researcher is the omitted category, these strong positive estimates demonstrate that discussion topics are more likely to be seen by those higher in the organizational hierarchy, indicating that discussion is more likely to diffuse up than down the hierarchical structure of the firm. Our interviews revealed that consultants were responsible for taking care of most day to day work at the firm which is reflected in the fact that the parameter estimate for 'consultant' is larger and more significant than for partners. This result also provides some evidence for the argument that mid-level managers 'have their ear to the ground' and are possibly the most aware of informal information circulating in the firm.

Hazard rate analyses mirror the logistic regression results to a large extent. Men see news at a higher rate than women, although demographic differences do not seem to predict the rate at which individuals see either news or discussion topics. The strength of ties again has a strong positive impact on the hazard rate for discussion topics but has no effect for news, while greater path lengths consistently reduce the hazard rate across both types of information. We again see increases in the rate at which employees see discussion news with greater project co-work (6.6% increase per additional project), but this time we see that having the same area of expertise increases the rate while industry tenure differences have no effect. Although managerial level differences are insignificant, the partner and consult dummy

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

variables show that employees in the top two levels of the organization see information at a higher rate than researchers, while consultants again seem to be the most well informed.

5. Discussion & Conclusion

We develop theory on how different types of information diffuse through social groups in organizations. In order to build our understanding of organizational information dynamics, we examine factors that affect the diffusion of two different types of information – event news and discussion topics. We empirically test our hypotheses by observing several thousand diffusion processes of each type of information in real email communication data collected over ten months in a mid-sized executive recruiting firm. Our results reveal that the confluence of social structure and information content together determine the movement of information in our firm.

We demonstrate that demographic distance plays a significant role in reducing the diffusion of information across all information types, and that traditional social network characteristics such as the strength of ties and the path length between individuals increase and decrease the likelihood of receiving information respectively. We find that prior project co-work increases the likelihood of receiving information, demonstrating the influence of prior work history on the diffusion patterns of information in organizations. We also find gender to be a significant predictor of the flow of information in our firm, with men being 55% more likely than women to receive a given piece of information during its diffusion process.

We also find that different types of information diffuse differently. Upon examining the diffusion of event news and discussion topics we find striking differences in the factors associated with their movement. While demographic distance reduces the likelihood of seeing both types of information, task characteristics such as project co-work and industry tenure differences are more likely to reduce the likelihood of receiving discussion information than event news. Discussion topics are more likely to diffuse vertically up and down the organizational hierarchy, across relationships with a prior working

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

history, and across stronger ties, while news is more likely to diffuse laterally as well as vertically, and without regard to the strength or function of relationships.

These findings highlight the importance of considering both structure and content in information diffusion. The confluence of structural factors related to demographic, organizational and social relationships, and the content of the information diffusing through a social group determines its diffusion path and the likelihood of its diffusion to given individuals. The theory and results developed shed light on who is likely to become aware of a given piece of strategic information in an organization and who is likely to become aware of it sooner. The theory and evidence developed here addresses some of the fundamental assumptions underlying a variety of research streams from the diffusion of innovations, to dynamic trading behavior, to word of mouth viral marketing.

Acknowledgments

We are grateful to the National Science Foundation (Career Award IIS-9876233 and grant IIS-0085725), Cisco Systems and the MIT Center for Digital Business for generous funding. We thank Tim Choe and Jun Zhang for their tireless research assistance.

References

- Adamic, L., Rajan, L., Puniyani, A. & Huberman, B. 2001. "Search in power law networks." *Physical Review E*. 64, 046135.
- Ahuja, G. 2000. "Collaboration networks, structural holes and innovation: A longitudinal study." *Administrative Science Quarterly*, 45: 425-455.
- Allen, T. J. 1977. Managing the flow of technology. Cambridge, MA, MIT Press.
- Ancona, D.G. & Caldwell, D.F. 1992. Demography & Design: Predictors of new Product Team Performance. *Organization Science*, 3(3): 321-341.
- Aral, S., Brynjolfsson, E., & Van Alstyne, M. 2006. "Information, Technology and Information Worker Productivity: Task Level Evidence." *Proceedings of the 27th Annual International Conference on Information Systems*, Milwaukee, Wisconsin.
- Beck, N., & J. Katz. 2001. "Throwing out the baby with the bath water: A comment on Green, Kim and Yoon." *International Organization*, 55(2): 487-495.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

- Bernard, H.R., Killworth, P., & Sailor, L. 1981. "Summary of research on informant accuracy in network data and the reverse small world problem." *Connections*, (4:2): 11-25.
- Burt, R. 1987. "Social contagion and innovation: Cohesion versus structural equivalence." *American Journal of Sociology*, 92(6): 1287-1335.
- Burt, R. 1992. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA.
- Burt, R. 2000. "The network structure of social capital" In B. Staw, & Sutton, R. (Ed.), *Research in organizational behavior* (Vol. 22). New York, NY, JAI Press.
- Burt, R. 2002. "Bridge decay." *Social Networks*, 24(4): 333-363.
- Burt, R. 2004. "Structural holes & good ideas." *American Journal of Sociology*, (110): 349-99.
- Buskens, V. & K. Yamaguchi. 1999. "A new model for information diffusion in heterogeneous social networks." *Sociological Methodology*, 29: 281-325.
- Chaimberlin, G. 1980. "Analysis of covariance with qualitative data." *Review of Economic Studies*, 47(1): 225-238.
- Cohen, W. & D. Levinthal. 1990. "Absorptive Capacity: A new perspective on learning and innovation." *Administrative Science Quarterly*, 35: 128-152.
- Cohen, W. & P. Bacdayan. 1994. "Organizational routines are stored as procedural memory: Evidence from a laboratory study." *Organization Science*, 5(4): 554-568.
- Coleman, J.S. 1988. "Social Capital in the Creation of Human Capital" *American Journal of Sociology*, (94): S95-S120.
- Coleman, J., Katz, E., & H. Menzel. 1966. Medical innovation: A diffusion study, Bobbs-Merrill, New York.
- Cook, K.S., Emerson, R.M., Gilmore, M.R., & Yamagishi, T. 1983. "The distribution of power in exchange networks." *American Journal of Sociology*, 89: 275-305.
- Cramton, C.D. 2001. "The mutual knowledge problem and its consequences for dispersed collaboration." *Organization Science*, 12(3): 346-371.
- Cummings, J., & Cross, R. 2003. "Structural properties of work groups and their consequences for performance." *Social Networks*, 25(3):197-210.
- Cummings, J. 2004. "Work groups, structural diversity, and knowledge sharing in a global organization." *Management Science*, 50(3): 352-364.
- De Vany, A. & C. Lee. 2001. "Quality signals in information cascades and the dynamics of the distribution of motion picture box office revenues." *Journal of Economic Dynamics and Control*, 25(3-4): 593-614.
- Dellarocas, C. 2003. "The digitization of word of mouth: Promise and challenges of online feedback mechanisms." *Management Science* 49(10): 1407-1424.
- Dessein, W. 2002. "Authority and communication in organizations." *Review of Economic Studies*, 69(4): 811-838.
- Dewar, R. & J. Dutton. 1986. "The adoption of radical and incremental innovations: An empirical analysis." *Management Science* 32(11): 1422-1433.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

- Domingos, P., & M. Richardson. 2001. "Mining the network value of customers" *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA: 57-66.
- Emerson, R. 1962. "Power-Dependence Relations." *American Sociological Review*, 27: 31-41.
- Freeman, L. 1979. "Centrality in social networks: Conceptual clarification." *Social Networks*, 1(3): 215-234.
- Gargiulo, M., M. Benassi, 2000. "Trapped in your own net? Network cohesion, structural holes, and the adaptation of social capital." *Organization Science* (11:2): 183-196.
- Greve, H. Tuma, D. & N. Strang. 2001. "Estimation of diffusion models from incomplete data." *Sociological Methods & Research*, 29: 435.
- Granovetter, M. 1973. "The strength of weak ties." *American Journal of Sociology* (78):1360-80.
- Granovetter, M. 1978. "Threshold models of collective behavior." *American Journal of Sociology* 83(6):1420-1443.
- Green, D., Kim, S.Y., & D. Yoon. 2001. "Dirty pool." *International Organization*, 55(2): 441-468.
- Greene, W. 2001. "Fixed and random effects in nonlinear models." *NYU Department of Economics Working Paper*. New York, NY.
- Grilliches, Z. 1957. "Hybrid Corn: An exploration of the economics of technical change." *Econometrica* 25(4): 501-522.
- Gruhl, et al., 2004. "Information diffusion through blogspace", in *Proceedings of the 13th international conference on World Wide Web*. New York, NY.
- Hansen, M. 1999. "The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits." *Administrative Science Quarterly* (44:1):82-111.
- Hansen, M. 2002. "Knowledge networks: Explaining effective knowledge sharing in multiunit companies." *Organization Science* (13:3): 232-248.
- Hargadon, A. & R, Sutton. 1997. "Technology brokering and innovation in a product development firm." *Administrative Science Quarterly*, (42): 716-49.
- Hausman, J., Hall, B., & Z., Grilliches. 1984. "Econometric models for count data with an application to the patents-R&D relationship." *Econometrica*, 52(4): 909-938.
- Heckman, J. 1981. "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process." In Manski, C. & D. McFadden (eds) Structural Estimation of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA, 114-178.
- Hirshleifer, D., Subrahmanyam, A., & T., Sheridan. 1994. "Security Analysis and Trading Patterns when Some Investors Receive Information Before Others" *Journal of Finance*, 49(5): 1665-1698.
- Hsiao, C. 1996. "Logit & probit models." In Matyas, L. & P. Severstre (eds), The econometrics of panel data: Handbook of theory and applications, Kulwer Academic Publishers, Dordrecht.
- Kempe, D., Kleinberg, J., & E. Tardos. 2003. "Maximizing the spread of influence through a social network" *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, D.C.: 137-146.
- Kleinberg, J. 1999. "The small-world phenomenon: An algorithmic perspective." Cornell Computer Science Technical Report 99-1776 (October).

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

- Kleinberg, J. 2001. "The small-world phenomenon and the dynamics of information." *Advances in Neural Information Processing Systems (NIPS)* 14.
- Krackhardt, D. & Kilduff, M. 1999. "Whether close or far: Social distance effects on perceived balance in friendship networks." *Journal of personality and social psychology* (76) 770-82.
- Kumbasar, E., Romney, A.K., and Batchelder, W.H. 1994. Systematic biases in social perception. *American Journal of Sociology*, (100): 477-505.
- Lescovec, J., Singh, A., & J. Kleinberg. 2006. "Patterns of influence in a recommendation network." *Pacific-Asia Conference on Knowledge Discovery & Data Mining (PAKDD)*.
- Marschak, J. & Radner, R. 1972. *Economic Theory of Teams*, Yale University Press, CT.
- Marsden, P. 1990. "Network Data & Measurement." *Annual Review of Sociology* (16): 435-463.
- McPherson, M., L. Smith-Lovin & J. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415-444.
- Meyer, J., & B. Rowan. 1977. "Institutionalized organizations: Formal structure as myth and ceremony." *American Journal of Sociology*, 83(2): 340-363.
- Mintzberg, H. 1979. *The Structuring of Organizations*, Prentice-Hall, Englewood Cliffs, NJ.
- Nelson, R. "What is public and what is private about technology?" Working Paper 9-90, Center for Research in Management, University of California Berkeley.
- Newman, M., Forrest, S. & J. Balthrop. 2002 "Email networks and the spread of a computer virus." *Physical Review E.*, 66, 035101.
- Orlikowski, W.J. 2002. "Knowing in practice: Enacting a collective capability in distributed organizing." *Organization Science*, 13(3): 249-273.
- Petersen, T. 1991. "Time aggregation bias in continuous-time hazard-rate models." In *Sociological Methodology*, edited by Peter Marsden, Basil-Blackwell, Cambridge, MA: 263-290.
- Petersen, T. & K. Koput. 1992. "Time aggregation bias in hazard-rate models with covariates." *Sociological Methods & Research*, 21:25-51.
- Pfeffer, J. 1983. "Organizational Demography," in Larry L. Cummings and Barry M. Staw (eds.), *Research in Organizational Behavior*, 5: 299-257. JAI Press, Greenwich, CT.
- Podolny, J., Baron, J. 1997. "Resources and relationships: Social networks and mobility in the workplace." *American Sociological Review* (62:5): 673-693.
- Reagans, R. & McEvily, B. 2003. "Network Structure & Knowledge Transfer: The Effects of Cohesion & Range." *Administrative Science Quarterly*, (48): 240-67.
- Reagans, R. & Zuckerman, E. 2001. "Networks, diversity, and productivity: The social capital of corporate R&D teams." *Organization Science* (12:4): 502-517.
- Reagans, R. & Zuckerman, E. 2006. "Why Knowledge Does Not Equal Power: The Network Redundancy Tradeoff" *Working Paper Sloan School of Management* 2006, pp. 1-67.
- Rodgers, E. 1995. *The Diffusion of Innovations*. The Free Press, New York.
- Rosenberg, N. 1982. *Inside the black box: Technology and economics*. Cambridge University Press, New York.
- Ryan, B. & N.C. Gross 1943. "The diffusion of hybrid seed corn in two Iowa communities." *Rural Sociology* 8: 15-24.

Essay 3: Organizational Information Dynamics: Drivers of Information Diffusion in Organizations

- Salton, G., Wong, A., & Yang, C. S. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM*, 18(11): 613-620.
- Schelling, T.C. *Micromotives & Macrobehavior*. George J. McLeod Ltd. Toronto.
- Sparrowe, R., Liden, R., Wayne, S., & Kraimer, M. 2001. "Social networks and the performance of individuals and groups." *Academy of Management Journal*, 44(2): 316-325.
- Strang, D. & N. Tuma. 1993. "Spatial and temporal heterogeneity in diffusion" *American Journal of Sociology*, 99(3): 614.
- Szulanski, G. 1996. "Exploring internal stickiness: Impediments to the transfer of best practice within the firm." *Strategic Management Journal* (17): 27-43.
- Tuma, N.B., & Hannan, M.T. 1984. Social Dynamics: Models and Methods. Academic Press, New York.
- Uzzi, B. 1996. "The sources and consequences of embeddedness for the economic performance of organizations: The network effect" *American Sociological Review*, (61):674-98.
- Uzzi, B. 1997. "Social structure and competition in interfirm networks: The paradox of embeddedness." *Administrative Science Quarterly*, 42: 35-67.
- Van Alstyne, M. & Zhang, J. 2003. "EmailNet: A system for automatically mining social networks from organizational email communication," NAACSOS.
- Von Hippel, E. 1998. "Economics of Product Development by Users: The Impact of "Sticky" Local Information" *Management Science* (44:5): 629-644.
- Watts, D. 1999. "Networks, dynamics and the small-world phenomenon." *American Journal of Sociology*, 105(2): 493-527.
- Watts, D. & S. Strogatz. 1998. "Collective dynamics of small-world networks." *Nature*, 393: 440-442.
- Wejnert, B. 2002. "Integrating models of diffusion of innovations: A conceptual framework." *Annual Review of Sociology*, 28: 297-227.
- White, H. 1980. "A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity." *Econometrica* (48:4): 817-838.
- Wu, F. Huberman, B., Adamic, L., & J. Tyler. 2004. "Information flow in social groups." *Physica A: Statistical and Theoretical Physics*, 337(1-2): 327-335.
- Yamaguchi, K. 1994. "The flow of information through social networks: Diagonal-free measures of inefficiency and structural determinants of inefficiency." *Social Networks*, 16: 57-86.

Conclusion

Conclusion

While production in developed economies is rapidly shifting from traditional manufacturing sectors toward information and knowledge based work, economic models and organizational frameworks explaining production, productivity and performance in these sectors remain underdeveloped. To develop our understanding of production in the information age, this thesis explores the role of information and technology in the productivity and performance of information workers. The three essays develop analytical models of specific aspects of information worker production processes and estimate the contributions of information technology use and skills, and the flow of information to productivity and performance. The methodologies, theory and evidence developed here attempt to further our understanding of the role of information in productivity, and contribute a series of replicable tools and techniques for studying these and other related phenomena in new settings. Several important findings emerge from this work.

First, while information technology (IT) use and skills contribute to productivity and performance, they do so by changing how people work rather than by simply speeding up traditional work processes. While a common assumption is that IT speeds work to the 'speed of thought,' I find, in simple models that IT use is associated with longer project duration on average. In more complete models, I find that information workers use IT more in order to support more effective multitasking behavior, and that multitasking has a strong positive, but non-linear relationship with revenue generation and project completion. Controlling for multitasking behavior, IT use does speed work demonstrating a more nuanced explanation for how IT helps improve productivity: by enabling efficient and effective execution of more simultaneous projects. I also find an inverted-U shaped relationship between multitasking and output, such that beyond an optimum, more multitasking is associated with declining project completion rates and revenue generation. The use of IT shifts this inverted-U shaped production frontier out at all levels of multitasking by enabling workers to complete their simultaneous projects faster at any level of multitasking. This increase in productivity is enabled in part by the asynchronous nature of information seeking and sharing made possible by IT. I find that asynchronous modes of communication (email) and

Conclusion

information seeking (database use) are highly correlated with multitasking behavior. These findings underscore the fine grained and complex ways in which IT changes work processes and ultimately influences the performance of information workers.

Second, I find that the topological structure of workers' communication networks significantly impacts their productivity and performance. An individual's location in the topology of communications in the firm in part determines their access to information, which in turn enables better and faster decision making, enhancing effectiveness. In particular, I find that diverse network structures enable more effective multitasking and give employees access to more novel and non-redundant information, which drives greater revenue generation and faster project completion. Novel non-redundant information is valuable due to its local scarcity. Actors with scarce, novel information in a given network neighborhood are better positioned to broker opportunities, use information as a commodity, or apply information to problems that are intractable given local knowledge. In essence, network structure enables an 'information advantage' by influencing the types of information to which information workers have access. In addition, employees with diverse network structures and those that are 'in the thick' of the flow of information in the firm (measured by the degree to which they are 'in between' others in the communication topology) are more effective multitaskers and are able to take on more simultaneous projects than the average. The results also demonstrate that there exist additional benefits to network structure beyond those conferred through information advantage. These benefits may include power, negotiating leverage or access to non-information based resources such as capital or reciprocity.

Third, network structures impact the diffusion of information in organizations and help determine who sees strategic information and who sees it sooner. I find however that different types of information diffuse differently. The results demonstrate that the diffusion of news, characterized by a spike in communication and rapid, pervasive diffusion through the organization, is influenced by demographic and network factors but not by functional relationships (e.g. prior co-work, authority) or the strength of ties. In contrast, diffusion of discussion topics, which exhibit more shallow diffusion characterized by 'back-and-forth' conversation, is heavily influenced by functional relationships and the strength of ties, as well as

Conclusion

demographic and network factors. Discussion topics are more likely to diffuse vertically up and down the organizational hierarchy, across relationships with a prior working history, and across stronger ties, while news is more likely to diffuse laterally as well as vertically, and without regard to the strength or function of relationships. These findings highlight the importance of simultaneous considerations of structure and content in information diffusion studies and demonstrate the importance of organizational and social factors in generating access to strategic information in organizations.

Finally, the methods, tools and techniques developed in this thesis are replicable and can be applied in other settings to further our understanding of the role of information in economic processes. Several important methods and tools are developed, including (a) the use of email data to characterize social networks (although this approach is not necessarily novel, I discuss several methodological difficulties associated with this approach and how I attempt to overcome them), (b) the analysis of communication content (in this case using vector space models) to analyze relationships between network structure and information content, and (c) the development of methods for identifying and measuring the diffusion characteristics of different types of information in organizational communication. I hope these analyses and methods will contribute to furthering our knowledge of social networks and the role of information and technology in the productivity and performance of information workers and information intensive organizations.