Comparative, population-level analysis of social networks in organizations

by

Abigail Z. Jacobs

B.A., Northwestern University, 2011

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2017

This thesis entitled:
Comparative, population-level analysis of social networks in organizations
written by Abigail Z. Jacobs
has been approved for the Department of Computer Science

_____

Prof. Aaron Clauset

_____

Prof. Brian Keegan

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Jacobs, Abigail Z. (Ph.D., Computer Science)

Comparative, population-level analysis of social networks in organizations

Thesis directed by Prof. Aaron Clauset

As social behavior moves increasingly online, the study of social behavior has followed. Online traces of social systems, whether to study online behavior directly or the online traces of offline activity, have made possible previously unavailable empirical analyses of people, groups and organizations. However, practically observing any social system is nontrivial: even if we can directly instrument and measure the social constructs we wish to study, we will still observe this through the lens of the system itself. We inherit effects due to the design and history of the platform, the ecology of other online systems, the measurement tool and pre-processing of our data, and the assumptions of our models. At the same time, organizations represent a fundamental unit of human social behavior. Then, to understand social behavior, we must understand how the size, boundaries, and context of organizations impact social relationships within them. I focus on this boundary of online systems and offline activity in organizations. We exploit heterogeneities across populations of social networks to explore the boundary of online systems, online social behavior, and offline activity across different organizations. I discuss empirical work exploring how offline behavior is reflected in online systems, and conversely, how an online system relates to offline outcomes. We then turn to the relationship between the measurement of networks from online data and past work on network structure and evolution.

In this dissertation, I develop a comparative structural perspective to tease apart the roles of these exogenous and endogenous processes on network structure. Using populations of comparable networks, I explore the roles of individual social strategies, organizational environments, and network construction on network structure. First, I explore how the unique timing and setting of Facebook's initial expansion to universities afforded a natural experiment, revealing differences in social strategies and network growth, and we explore empirical network scaling in this population

of networks. We find that the social strategies employed by students who only interacted online differed from those who had interacted in the offline world. Second, I explore a vaunted tradition of organization theory—relating a firm's informal network structure to firm performance—using a novel email network data set across a population of large firms. In this setting, I explore the previously untested heterogeneity of firms and the relationships between organization size, organization context and social network structure. There, we find a surprising amount of heterogeneity across firm types, and a lack of relationship between network structure and firm performance. We find novel scaling results, including a lack of relationship between the size of a firm and an individual's number of contacts, but find that the formal geographic structure of an organization increases bottlenecks in communication across firms. Finally, reflecting on the challenges of working with social networks drawn from interaction data, I explore the connections between network construction and network evolution. To put these connections in perspective, I visit the theory of weak ties, network stability and network densification using this lens. We find evidence to confirm, reject, and suggest novel hypotheses in this literature. We find, for example, that network densification can appear as an artifact of total activity within the observed system.

The comparative approach is uncontroversial but novel in the empirical study of networks, organization theory, and computational social science. In this context, the comparative approach allows us to compare empirical scaling properties to results from random graph theory. Using networks bounded by organizations and platforms, we can leverage the boundaries of online systems to relate covariates at the platform-, organization-, or network-level to structure. This provides a novel empirical perspective into how offline behavior is reflected in online systems, and conversely, how an online system relates to offline outcomes. By working with populations of *comparable* networks, we can meaningfully characterize variation across empirical social networks and shed light on the ecological and organizational processes underlying online social systems. The comparative, population-level approach suggests a novel opportunity in network science and the computational study of social systems.

Dedication

For Judy and Israel S. (Jake) Jacobs, who paved the way,

and Sarah Jacobs, who kept me questioning

Acknowledgements

A potential conclusion from this dissertation is that networks do not matter. Anecdotally ($N = 1$), the strength and content of my relationships have had a major impact on me during this process. While I am sure this list is incomplete, I would like to thank my strong and weak ties who supported me and made this all possible.

First, and most formally, I have made it through this process due to the patience of my dissertation committee, chaired by Aaron Clauset, which included Jordan Boyd-Graber, Brian Keegan, Mike Mozer, and Hanna Wallach, as well as Jim Martin during earlier iterations.

I want to thank my advisor, Aaron Clauset. Aaron encouraged me into research before either of us had even reached Colorado, and then his commitment to interdisciplinary research and critical academic exploration opened many doors for me. I am particularly thankful to Aaron for letting me find my own way through grad school. My path was more meaningful having gone through these detours. I am grateful to Duncan Watts, with whom I collaborated for much of this work. Duncan supported me in carving a new path path through both New York and computational social science. I learned a tremendous amount from him over the last two years. My life changed in 2014, coincidentally at the Women in Machine Learning Workshop, when I met Hanna Wallach. Thank you for your mentorship, guidance, and friendship. I can not imagine this without you.

I visited many desks, academic labs, and reading groups and slept on even more couches, in the making of this dissertation. This generosity of these hosts, mentors, and friends across cities and institutions can not be overstated.

Microsoft Research NYC, and the Computational Social Science group in particular, served as

an academic home for much of the creation of this document. I particularly want to thank Shawndra Hill, Jake Hofman, Sid Suri, Ashton Anderson, Solon Barocas, Amit Sharma, and Ally Wharton. I benefited from an embarrassingly long list of other members of the lab—permanent and postdoc, visitors and interns—who informally served as invaluable mentors, teachers and friends. Thank you for sharing your space, time and energy. The Clauset Lab was my home at Colorado, from which I especially want to thank Chris Aicher, Nora Connor, Amir Ghasemian, Dan Larremore, Sears Merritt, Leto Peel, and Sam Way. Other formal collaborators and mentors contributed deeply to my progress in graduate school, including Jennifer Dunne, Winter Mason, Cristopher Moore, Cosma Shalizi, and Johan Ugander. A broader cast of individuals, agencies, and institutions including the NSF supported my Ph.D., and I thank the relevant agencies and colleagues at the end of each chapter. Despite the best efforts of my committee, my collaborators, and my mentors, any remaining errors, ambiguities, omissions and mistakes remaining in this document are mine alone.

In Boulder, Joey Azofeifa, Rebecca Hames, Ryan Langendorf, Liz Millikin and Kenny Underwood were patient friends and thoughtful supporters. I am also grateful to have gone through so many academic and life stages with Nora Connor. I shared projects, much coffee, excellent food and cocktails, multiple time zones, and even a defense day with Sam Way, and almost as much with Laura Norris. The pleasure was all mine. Nora, Sam, Laura, and Kenny, thank you especially for keeping up not one but many home bases for me. Stacey Reynolds was my first local friend and pillar of support, and she remains a shining light of Boulder. Stacey, you are celebrated and missed. Beyond Colorado, I am especially indebted to Ashton Anderson, Eleanor Brush, Hannah King, Adit Kumar, Aaron Schein, Nandita Seshadri, Caitlin Ting, Tim Vieira, Elaine Wah, and Amy Wesolowski.

Finally, I am grateful to my sisters Sarah and Sophie, my parents Cindi and Mike, and Dorothy. I barely noticed Cindi get her doctorate, a feat which seems even more unimaginable now. Sarah questioned my decisions while leading the way forward. Sophie, I'm excited to see you continue to grow into your own.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Social networks—specifically, social network data—are derived from largely unobserved social systems. This represents a challenge and an opportunity to understand how structure can be derived from such complex systems and what we can learn from that structure. Conversely, from a data-driven perspective, we can how to understand networks in the context of the social systems from which they are drawn. Network structure serves as a lens into the interactions in a social system, representing local and global constraints on relationships, information flow, and group formation. With an observed social network, we use metrics or models to characterize network structure. These tools operationalize theorized or observed social processes, such as centrality measures to represent status, or stochastic block models to infer community structure.

Social networks additionally encode structure that reflects endogenous and exogenous processes in the social system from which they are derived. For example, the Facebook friendship network reflects differences induced by system-endogenous design changes from within the platform itself (Malik and Pfeffer (2016) and Zignani et al. (2014a)) and shifting norms among users (boyd (2013) and Tufekci (2008)). The Facebook network structure also reflects its history from competing with other online social networks (boyd and Ellison (2007) and Kleineberg and Boguñá (2015, 2016)). However, given this mixture of processes, determining when and if the structure we discover is an accurate reflection of underlying social processes of interest is then a nontrivial task.

A fundamental unit of social systems is the organization. To understand social systems and individuals, we then also need to understand organizations and communities. We also look

to the specific setting of organizational social networks. The history of using social networks to understand organizations is long: in the foundations of management, Roethlisberger and Dickson (1939) argue for the importance of the informal relationships within work groups, and some of the earliest social network research focused on the optimal organization of work groups (Moreno 1934).[1]

However, this is typically done on single examples: we typically only observe one network for a setting, whether it be a single organization, such as a karate club (Zachary (1977)) in the "offline" world, a single firm, or a single online social network platform. A comparative perspective allows us to understand and measure the natural variation in these systems, and understand the degree to which the examples we observe are products of their environment or are atypical.

The comparative perspective is the norm in other domains: consider the demographer or biologist characterizing a population of a species, or the anthropologist or political scientist comparing societies and institutions. In the study of social networks, this perspective has largely been missing. This is partly due to a lack of data—there is only one Twitter—and so this gap is reflected in underdeveloped empirical results and open theoretical challenges in the network science literature. The lack of comparative work also extends to the organizational theory literature, which draws from

[1] Moreno (1934) develops sociometry as a tool to design "a social group which can function at the maximum efficiency and with the minimum of disruptive tendencies and processes." His book works through a range of applications on organizational networks, from the organization of work groups in a prison to predicting how network structure was related to which students ran away from a "school" for teenage girls convicted as delinquents. (Coincidentally, this was the New York State Training School for Girls in Hudson, NY. Moreno conducted most of his studies there in 1932; the singer Ella Fitzgerald was sent there as a teenager in April 1933, sentenced delinquent by the state. A judge wrote that she was "ungovernable and will not obey the just and lawful commands of her mother"; she ran away some time later that year (Immarigeon 2014).)

Despite the organizational focus of Moreno's work, the primary challenge he saw for survival was not in reference to the school retention rate or to organizational success, but a greater existential threat. Explicitly, "the meaning of the title of this book 'Who Shall Survive?' is the survival of creativity, of man's universe. *The survival of human existence itself is at stake*" (Moreno (1953), p. 600; emphasis his). Moreno described concern about humanity faced with "two threats, the aggression coming from man and the aggression coming from 'robots.' The answer to the first [is] *sociometry*" (Moreno (1953), p. 599).

In the 1934 edition, and even then "the destiny of man" (Moreno 1934, p. 366) was already at stake: "the weakest point in our present day universe is the incapacity of man to meet the machine, the cultural conserve, or the robot, otherwise through submission, actual destruction, and social revolution," Moreno (1934), p. 363. The early adoption of this language is notable; the stakes appear to be raised further in the later 1953 edition, which presents the introduction of the atomic bomb as relevant and discouraging in this progression.

Moreno characterizes his study of the structure of human relationships as "the discovery and demonstration of the social atom," and presents sociometry as one of the tools necessary for "the final situation of man and his survival" (Moreno 1934, p. 363). He also departs from this narrative to apparently advocate an additional approach, beyond the social and technical: the "eugenic doctrine" as a "promiser of extreme happiness to man" (Moreno 1934, p. 365; Moreno 1953, p. 597). Despite this horrifying early suggestion, the field of organizational social networks has not carried this suggestion forward, nor do we promote it here.

economic sociology to understand the structure and behavior of organizations. This field has historically relied heavily on single examples to understand the social system of the organization, and advanced a wide range of under-specified, and occasionally conflicting, hypotheses (Blau (1965), Carroll and Hannan (2000), Davis (2010, 2015a), Kimberly (1976), and Schwarz et al. (2007)). In 1965 the sociologist Peter Blau laid out a potential research agenda for "The Comparative Study of Organizations." Blau (1965) opens: "The comparative method, in the broadest sense of the term, underlies all scientific and scholarly theorizing." We proceed from this sentiment.

1.1 Problem setting

This dissertation focuses on social networks, representing people and the interactions among them. We adopt a comparative, population-level approach to explore heterogeneities across networks that are defined within the boundaries of organizations (universities; firms) and mediated by online platforms (early Facebook; email). The measurement of networks, the comparative approach, and the growth and comparison of online networks and organizations all invite unique challenges. We review the context and breadth of these challenges and approaches that are aligned with the direction we adopt here.

1.1.1 Networks

N = 1 network analysis. Network measures describe the *structure* of networks, that is, the local and global patterns of connections that ideally correspond to some social phenomenon of interest. These measures may capture some local description of the population of individuals in the network, such as degree, which might describe the number of contacts or friends in a social network, or the aggregate of local descriptions, such as average degree. These measures can also capture global properties of the networks, unobservable to individuals. For example, the greatest distance between any two individuals in the network, captured by the network diameter, or the degree to which communication must pass through a small number of bottlenecks to diffuse across the network (a measure of centralization: see Chapter 4). Our emphasis is on whole-network measures

that characterize the patterns across the whole network, either aggregates of local structure or global structure, rather than the specific position of an individual. White et al. (1976) were early advocates of this perspective, and used organizational social networks to argue for the efficacy of network-level analysis.

We focus on measures that are defined to describe some implied social process. The measures used throughout the dissertation are defined in Chapter 2. We refer the reader to Wasserman and Faust (1994) and Newman (2010) for more exhaustive references.

Given single network examples, it is also useful to establish means with which to infer whether or not large-scale structure in the observed network is meaningfully different than random chance. Probabilistic models have also been widely adopted as such an approach to understanding structure within single network examples. Generative models allow for sampling from distributions of networks: this yields a rigorous framework with which to detect large-scale patterns of structure, such as community structure (e.g., Aicher et al. (2015)) or hierarchical structure (e.g., Clauset et al. (2008)), and test whether or not that structure is meaningful.[2] Further discussion is beyond the scope of this dissertation but see Jacobs and Clauset (2014) for a relevant overview.

Regardless of the instrument, measurements of network structure vary across networks. Some sources of variation across networks are better understood: random graph models help us explore what network properties emerge as a function of network size (Newman (2010)). The degree to which these scaling properties appear empirically, or if different scaling properties apply, is still underdeveloped as empirical network studies are typically done on single ($N = 1$) network examples. Random graph models also help reveal how real world networks tend to empirically deviate from random: for example, social networks tend to have higher clustering (Watts and Strogatz (1998)). Other sources of variation across networks, due to behavioral norms, individual differences, or network- or organizational-level outcomes are less clear. For example: Is the Facebook

[2] Unfortunately, these models do not extend trivially to populations ($N > 1$) of networks. This is true whether they be exponential random graph models, which already suffer a range of degeneracies including lack of projectivity, or more traditional exchangeable generative models for network structure (D'Amour and Airoldi (2016) and Shalizi and Rinaldo (2013)). This is an open and exciting area for future methodological research.

network like Twitter? Are food webs like high school social networks?[3] Does online social network structure vary in a way that predicts their success or dissolution (Garcia et al. (2013))? Does organizational network structure predict performance, or correspond to industry differences (Chapter 4)? These questions underlie a range of methodological and social studies, but are inaccessible without analyzing populations of networks.

N > 1 network analysis. It is, however, reasonably common practice to compare a handful of different types of networks to demonstrate the robustness of an empirical phenomenon (e.g., citations among articles and patents, autonomous systems communications of the Internet, email networks, and movie actor-film relationships (Leskovec et al. (2007))) or of a modeling technique (e.g., hierarchical structure in a terrorist association network, a metabolic network, and a food web (Clauset et al. (2008))). Demonstrating a phenomenon across multiple diverse systems suggests robustness—never universality!—but do not provide systematic evidence of a particular phenomenon across a type of social systems.

Small sample sizes can suggest misleading empirical results (see, e.g., Button et al. (2013)), and messy social systems are no exception. Davis (1970) analyzes 742 (!) social networks from about 400 small social groups—using "sociometric data presented in punch card form (one card for each row of a sociomatrix)," no less—and fail to find support for Davis's own previous work on structural balance, which had found evidence from a population of 60 empirical networks (Davis and Leinhardt (1972); then in press). Uncontroversially, evidence for specific empirical social phenomena is more compelling when shown across multiple instances of comparable network types. For example, the role of status is robust across a population of high school social networks (Ball and Newman (2013)). Facebook networks decreased density and average shortest path, both at the country level and globally, over a period of five years (Backstrom et al. (2012)). Gender is better predictable as a node attribute in Facebook and high school networks using the distribution of attributes of two-hop neighbors (Altenburger and Ugander (2017)).

[3] Based on the strongly status-driven patterns of unreciprocated relationships—"aspirational friendships" (Ball and Newman (2013)) along a one-dimensional niche space—in high school friendship networks, comparing the high school social universe to the food chain is likely more apt than not.

Analyzing multiple networks quickly induces issues related to size: a single individual exchanging messages, for example, with 10% of an organization means something fundamentally different in a community of 100 vs. 100,000. The empirical relationship of network measures to network size is a fundamental question in networks-related research, as it immediately interferes with the comparison of measures across networks of different sizes (e.g., Dunne et al. (2002) and Faust and Skvoretz (2002)). The ecology community, for one, has been forced to reckon more directly with the role of size of networks is ecology. Ecologists are often interested in the stability or resilience of ecosystems, which may be of different sizes. Teasing apart the role of size from the questions of interest, which may otherwise be conflated with the role of ecological processes, is then of immediate interest. As a telling example, Dunne et al. (2013) find that food web research, where past work had tried to determine the impact of introducing species of parasites to food webs on ecological outcomes, conflated changes to network structure that were strictly due to increases in size with ecological impacts.

Understanding how network structure empirically varies with network and organizational properties, including size, brings us naturally to questions of comparison. Analysis across multiple networks is undoubtedly useful: to make a claim that some process occurs and to show that it occurs in multiple environments is a scientifically meaningful effort. But this is different than a $N > 1$ *comparative* or *population-level* approach, where we also can characterize variation at the individual and organizational level across a population of networks. We first consider the related context of studying online social systems (Chapter 1.1.2) and ecological approaches to studying organizations and networks (Chapter 1.1.3). Together these perspectives lend insights into the comparative approach, which we return to in Chapter 1.1.4.

1.1.2 Measuring online systems

Social dynamics, exogenous processes (such as platform competition or environmental constraints), and choices about network representation, implicit or explicit (including name generators, survey design, threshold setting), can obscure the patterns we intend to measure in online social

systems. We explore a range of these issues here and point to a set of related methods that have emerged as a result in Chapter 1.1.3.

Name generators In the study of social networks, name generators historically refer to the questions or definitions used to operationalize relationships between people (Campbell and Lee (1991)). In sociology, differences among name generators have been studied to capture different types of social structure. While it may not be surprising, it is still useful to understand that, e.g., advice networks differ in structure from the trust networks (Lazega and Pattison (1999)). While different name generators induce different structures, it is still difficult to make large-scale comparisons of network types without a meaningful baseline of heterogeneity in these systems to begin with. Data-driven comparison across networks of a single type can help quantify the diversity in these systems. If this natural heterogeneity is more or less significant than the differences between modes of measurement, this is revealing.

Within a single social system, dynamics and differences imbued by different network generators can obscure the patterns we intend to measure. Comparing the modes of defining an interaction or edge in a social network has been explored to some degree in online systems, e.g., email vs. in-person contact (Grippa et al. (2006) and Huberman and Adamic (2004)); maintained relationships, communication, and 'friendship' on Facebook (Marlow (2009)); within a single type of network, this has been explored through the strength of reciprocity in email networks (De Choudhury et al. (2010)) and the time window in proximity networks (Clauset and Eagle (2007)); see Chapter 5 for more details.

Network construction & measurement Sampling algorithms to construct networks from a more general system can lead to robust false discoveries of certain structural patterns (Lee et al. (2006)): for example, algorithms to sample subnetworks from large, difficult-to-measure networks can misleadingly suggest skewed degree distributions (Achlioptas et al. (2009)) and heuristic network measures detect degenerate community structure (Good et al. (2010)). Sampling an underlying network, such as when people encode their offline relationships on an online platform, can produce network densification, regardless of the structure of the underlying network (Pedarsani et

al. (2008) and Schoenebeck (2013)). Conversely, community structure in the underlying structure of a network can induce wide variance in estimates drawn from standard sampling approaches (Li and Rohe (2015) and Rohe (2015)). Network measures may also vary with network size—a theme we will return to in every chapter—and failing to take this into account can lead to incorrect inferences about the structure of networks (Dunne et al. (2013)). Assumptions used during modeling and inference may further introduce biases or obscure large-scale structure (Jacobs and Clauset (2014)), and algorithms to infer structure used in practice can vary dramatically in output and over-fitting (Ghasemian et al. (2017)).

These problems aside, inferring a social network from interaction data is a nontrivial task. Systems of social interactions can be measured in numerous ways, revealing different patterns related to the measurement tool itself. Constructing a network from a set of interactions involves nontrivial choices about representation that significantly impact the types of patterns detectable in a network (Clauset and Eagle (2007)): we explore the parameters that go into constructing a network from communication interactions in Chapter 5.

Boundaries of network data and of organizations. Measuring relationships in a social system requires choosing a boundary of that social system, that is, determining who belongs in our observed population. This challenge, related to sampling, is not novel: see, e.g., Laumann et al. (1989) for ways to characterize the boundaries of social networks by internally socially consistent or externally imposed social designations. This has also been recognized as a necessary challenge if the unit of interest is the full network structure (White et al. (1976)).

Fortunately, defining such a boundary should presumably be more straightforward in organizations (e.g., Wasserman and Faust (1994)), and even more so in online platforms that require explicit engagement or membership (Holme (2015))—and yet. First, this question is related to sampling, and asks what population is necessary to observe social and organizational processes of interest (Kilduff and Brass (2010)). Furthermore, even in traditional corporate settings, large firms act as systems of interacting smaller firms (Ghoshal and Bartlett (1990)), and relationships between firms transcend organizational boundaries (Granovetter (1994)), so even the boundaries

within traditional organizations are porous and non-obvious. In Chapter 4, we are forced to reconcile that the number of full-time employees—which excludes part time and temporary workers, as well as contractors—can differ widely from the number of organization-affiliated email users.

Platforms can automatically suggest a boundary: for example, those that are members of Facebook is a natural way to delimit the Facebook network. However, this is a nontrivial assumption. An example of immediate relevance is Facebook in its initial founding (Chapter 3), when one needed a specific university email address (e.g., @harvard.edu), which required having a sufficient relationship to the organization (as a student, professor, or staff); later Facebook opened to specific employers (and the concomitant problems of defining the boundaries of an organization), other student types, and the general public. Passive consumers of platforms, such as Reddit, or anonymous contributors immediately suggest that social activity on a platform may extend beyond the bounds of membership. Among members, failing to take into account cohort or level of user engagement can yield misleading aggregate assessments of user behavior (Barbosa et al. (2016)). Finally, interaction across and user migration between platforms yields a different view of the online social world (Chapter 1.1.3).

It is now unambiguous that social dynamics mediated on the Internet, including through online communication and communities, interact with social dynamics in the offline world (Wellman and Haythornthwaite (2008)).[4] While the boundary between online and offline worlds may be eroding, it is still useful to employ this dichotomy. For example, shocks and interventions in the "offline" world can induce and reveal shifts in social behavior, and this has been applied in settings as varied as hurricanes (Phan and Airoldi (2015)), changes in stock price (Romero et al. (2016)), FOIA requests (ben-Aaron et al. (2017)) and the implementation of censorship (Hobbs and Roberts (2016)). Differences among users of a given platform may interact differently online depending on their offline affiliation (Jacobs et al. (2015a), Kossinets and Watts (2009), and Zhu et al. (2014)). (Shocks that come from changes within the platform itself, or by interactions between platforms,

[4] How online communication, information dynamics, and social platforms on the Internet interact with broader economic and geopolitical systems, while a pressing and timely issue in 2017, is beyond the scope of this dissertation.

are of a different and endogenous flavor: see Chapter 1.1.3.) Regardless, explicitly considering the offline features of users and collections of users, including organizations, yields a useful perspective into patterns of online social behavior.

To briefly illustrate this perspective, we note that in Chapters 3, 4, and 5, we explicitly leverage the boundaries induced by organizations. In Chapter 3, we consider university-affiliated members of Facebook with active accounts at the beginning of September 2005 for a set of specific universities; using information about the platform growth and offline properties of those universities, we find that online network growth varies with the offline context of these networks. Chapter 4 uses employer-specific enterprise email to delimit membership in firms and account for firm size, and we ask if (well-theorized) offline properties of the organizations vary with network structure. In Chapter 5, we reckon directly with how the specification of networks derived from interaction data can induce differences in network structure, and how variation along these specifications can align with existing theory or, alternatively, build or erode trust in empirical results. However, we first discuss a range of perspectives that have emerged in response to a range of these challenges: these represent important directions for future research, and we draw on these perspectives in the work here.

1.1.3 Emerging perspectives in the study of online social systems

In response to these challenges in measuring social processes in online social systems, a number of theoretical and empirical approaches have emerged.

Platform effects: system design vs. user behavior Exploring the boundaries of network data, that is, understanding the way that it is shaped by and shapes the social system it is drawn from, is made more accessible in the comparative network setting. This can otherwise be difficult in typical $N = 1$ settings, such as online systems, where one instead must generally rely on hopefully revealing interventions, environmental and platform-induced natural experiments, and platform effects (Malik and Pfeffer (2016)). This can be brought into a causal inference context, using platform design changes (Malik and Pfeffer (2016), Oktay et al. (2010), and Su et al. (2016))

to infer how social networks change. Conversely, in settings where the platform design, such as recommender systems, are creating unknown effects on user behavior, exogenous changes to the system can be used to measure the effect of the design (Sharma et al. (2015) and Su et al. (2016)). Other system-level interventions may be relatively exogenous changes that impact user social behavior (such as reduced platform access through censorship (Hobbs and Roberts (2016)), or social structures responding differently to natural disasters, such as hurricanes (Phan and Airoldi (2015)) or designed as experiments (such as modified rewards systems (van de Rijt et al. (2014))).

It is worth noting the concept of "platform effects" relies on two distinct interpretations of online social networks. The first mode, on which this dissertation rests heavily, is the concept of online social networks as social networks, representing connections between people. The second mode, which often employs the same language to refer to the platform operators, i.e., the organizations that host and design the system platform online: the social networking sites themselves (Weber et al. (2016)). Taking advantage of when the social networking site (organization) makes changes to the platform that effect the social network (structure) yields insight into social processes happening between users of the platform, which may appear in the social network structure. However, social networking sites (the latter mode) incorporate ideas and models from the social network literature (the former mode) into the design of systems, which then encourages behavior that follows the design, a process called performativity (Healy (2015)). For example, the concept of triadic closure describes the process of people with shared connections becoming connected—enacting the concept that "the friend of my friend is my friend"—has been well established in the study of networks (Rapoport (1953)) and in sociology (Granovetter (1973)). Zignani et al. (2014a) found that triadic closure increased suddenly in 2008, concurrent with when the Facebook platform introduced the "People You May Know" feature: that is, when the platform introduced a recommender system explicitly leveraging the concept of triadic closure (Malik and Pfeffer (2016)). The feedback loop induced between social network concepts and the design of social networking sites is as old as social networking sites themselves: the first modern social networking site—SixDegrees.com, founded in 1997 (boyd and Ellison (2007) and Weber et al. (2016))—encoded the popular network concept of

"six degrees of separation" in its name.[5]

Ecological approaches: organizations Organizational ecology represents a major paradigm of organization theory that emphasizes the diversity and sources of heterogeneity in populations of organizations (Baum and Shipilov (2006)). In organization theory, these ideas have been imported for several decades (Carroll (1984) and Hannan and Freeman (1993)), including population ecology and evolution (Baum and Singh (1994a)) and niche theory (Baum and Singh (1994b,c)). The tradition of organizational ecology is traceable to Hannan and Freeman (1977)—"The Population Ecology of Organizations"—where a core challenge in organization theory is understanding the heterogeneity of organizations. Paraphrasing the ecologist G. E. Hutchinson, Hannan and Freeman (1977) ask, "Why are there so many kinds of organizations?"[6] In Chapter 4, we suggest novel empirical support of this perspective in a population of firms.

As it has become more apparent that online communities and organizations face similar constraints to growth, change, and evolution to those as offline organizations (e.g., Kreiss et al. (2011), Shaw and Hill (2014), and Wang et al. (2013)), the (offline) organizational ecology perspective may also provide novel opportunity to understand online communities. While studies that cross multiple communities are still rare, this is a compelling open area for future research. Despite a prevalence of enterprise online systems, such as email, messaging, and task-based applications that serve multiple organizations, and multi-community online platforms, such as Reddit, Wikia

[5] This concept has a rich history in popular and network science (see Watts (2004) for a relevant overview), sociology (Milgram (1967)), literature (Guare (1990) and Karinthy (1929); see Backstrom et al. (2012)), and pop culture (see Schuessler (2017) for a recent reincarnation). The website followed only a few years after the 1993 film based on John Guare's 1990 play, "Six Degrees of Separation."

[6] Understanding the foundations and sources of diversity in natural systems is foundational in biology. The perspective of niches affords a way to characterize the space over which species reside, providing a general framework for understanding coexistence and competition. According to the research program led by Michael Hannan and John Freeman, these problems map naturally to organizations. Hannan and Freeman (1977) purposefully mirror a famous address by G. Evelyn Hutchinson, the founder of ecological niche theory: "Why are there so many kinds of animals?" (Hutchinson (1959)).

Here we focus on organizations and networks, which may occupy different niches in the context of *populations* of organizations and networks. Organizations, like all social and biological systems, function at multiple scales (Simon (1962)), and so niche modeling can also be applied within organizations. A thorough discussion of niches applied to other system levels is beyond the scope of this dissertation, but we note that niche models can be applied to understand roles *within* networks (see Jacobs and Clauset (2014), Jacobs et al. (2015b), and Williams and Purves (2011)). Liu et al. (2015) employ this perspective to examine roles within organizations, using informal social networks in a firm derived from an email communication network (cf. Chapter 4); they find that density and diversity within niches is related to employee performance.

and StackExchange (Hill and Shaw (2017) and Tan and Lee (2015)), this area remains underdeveloped. However, the ecological perspective on the rate, constraints, and sources of heterogeneity in populations of organizations and the structural mechanisms that support and impede change in organizations suggests novel hypotheses for the analysis of online communities.

This perspective has been applied in a handful of settings. Wang et al. (2013) find, for example, that online groups in the same niche suffer competition from shared members. More subtly, Zhu et al. (2014) uses ecological niches to describe communities within an online organizational communications tool. They find that competition among similar communities is stronger among communities with users with shared offline affiliations. They also find that similar communities are more successful when they do not have users with shared offline affiliations. This tradeoff—between competition driven by similar users vs. benefit to the community from having coexisting competitors—suggests a novel characterization of online communities. Tan and Lee (2015) characterizes how users traverse multiple communities in a niche space and show how these trajectories predict future activity. In a platform for petitions, TeBlunthuis et al. (2017) find evidence of density dependence theory, with an inverse U-shaped relationship between the density of the niche that a petition occupies and its success. However, they do not find evidence that specialized petitions do better, although this has been found for offline organizational forms (Carroll (1985)).

Finally, we note again that social networking sites are themselves a type of organization (a nontrivial observation: see Weber et al. (2016)). Then to the extent that users are shared resources, these concepts from organizational ecology map to the competition and coexistence of different platforms by expanding over different niches (Kleineberg and Boguñá (2015, 2016)). More deeply understanding the structure of users within platforms lends itself to more specific ecological concept of community assembly, which I describe next.

Ecological approaches: assembly We draw from the ecological notion of assembly, in which a community, possibly represented by a network, is formed in a way that depends on a number of complex factors: composition of the current community; ordering effects (which group arrives earliest may set constraints on who may join or set norms for behavior); competition within

and between systems; and natural limits on growth (due to local or global resources or current community size). Community assembly was originally framed using islands as the unit of analysis: a novel opportunity for ecologists to consider almost-independent model systems, and quantitatively describe how variation across these communities could be explained by processes within and outside of these units (MacArthur and Wilson (1963) and Warren et al. (2015)). Assembly specifically leverages the boundaries of the systems to delimit patterns of growth, competition, and evolution. We exploit the analogues to these concepts in online social networks and organizational networks and use *network assembly* as a frame, within which we can tease apart different social processes. Separating the complex social, behavioral, and engineered processes that mediate online and offline social networks is nontrivial but crucial to understanding how social networks encode and influence relationships.

In the context of social network data, the process of network assembly brings together the unknown mixture of online, offline, social and behavioral, structural and design-based mechanisms that are subject to constraints due to ordering effects, competition, technology, composition and context. This is a concept richer than that of network growth, and captures both the emergent construction of networks or groups as well as their explicit formation (Bascompte and Stouffer (2009), Contractor (2013), Lungeanu et al. (2014), May (2009), and Saavedra et al. (2008)).

These concepts have been explored in a range of empirical settings in online social networks. Shaw and Hill (2014) find the establishment of norms is set and entrenched by the early administrators of peer production systems: specifically, the initial settlers of the governance arm of wikis determine future behavior (providing support for the Iron Law of Oligarchy (Michels (1915)), analogous to founder effects in community assembly). Notably, Shaw and Hill (2014) find this across a population of peer production systems. On individual systems, Heaberlin and DeDeo (2016) find that the evolution of norms is related to those established by the earliest users on Wikipedia. Kooti et al. (2012) found the establishment of retweeting norms on Twitter diffused from core, active members; this empirical observation found that pre-existing network structure was related to this spread. Centola and Baronchelli (2015) found this experimentally, establishing that the

structure of a social network could determine the emergence of norms.

Beyond initial founders, Barbosa et al. (2016) found that the arrival timing of different cohorts revealed differences in activity and engagement with Reddit, an online community. Fire and Guestrin (2016) found that the distribution of arrival times across a population of Reddit networks was related to differences in network structure. In Chapter 3 (Jacobs et al. (2015a)), we find that differences in offline context of users, which varied by cohort, changed patterns of adoption and network structure; during a similar time window on Facebook, Lampe et al. (2006) found that offline networks informed online social network activity. As in ecological communities (and organizational ecology: see, e.g., Baum and Singh (1994b)), platforms can coexist by establishing different niches (Kleineberg and Boguñá (2015)). Different niches can be related to, or exacerbate, demographic differences across platforms (boyd (2013) and Hargittai (2007)). Turmoil within a platform can drive cross-platform migration, which in turn can drive differences in user behavior (Newell et al. (2016)). The spread of a platform across demographic niches, combined with platform design effects—such as a feature to highlight most popular users or a policy enforcing that profiles map to real identities, both on Friendster (boyd (2006))—can drive behavioral norms on a platform, as well as drive users away, leading to collapse.

1.1.4 Comparative & population-level approach

Comparative, population-level analyses of networks allows us to unite these perspectives under a common umbrella, find more compelling evidence of social processes across organizations, and take advantage of natural variation in the environment and engineered variation in platforms to meaningfully compare social systems.

At a structural level, we can begin by remarking that empirical networks are structurally different than those generated from random graph models. How properties of observed empirical social systems vary with network structure is of unambiguous interest in the study of networks (Newman (2010), Newman (2003), and Watts and Strogatz (1998)). For example: are social networks different than random graphs? (Yes. See, e.g., Newman and Park (2003) and Watts and

Strogatz (1998).) Are social networks different than biological and technological networks? (Maybe, although it depends on your choice of data set construction (Larremore et al. (2014)) or taxonomic classification (Onnela et al. (2012)).) Is Twitter like Facebook? (Great question. This could be answered by comparing their embeddings in some latent space (Asta and Shalizi (2014)), distribution of community structure (Onnela et al. (2012)); their position in the organizational niche space (Kleineberg and Boguñá (2015), Wang et al. (2013), and Weber et al. (2016)); or the "social resilience" (Garcia et al. (2013)) or "loyalty" of their users (Hamilton et al. (2017)).) Or, more simply, how does empirical network structure vary by size? Is scaling similar to what is predicted by random graph models? (It's complicated. But consider Chapters 3, 4, and 5.) Does organizational network structure predict their performance? (We suggest not (Chapter 4).) By adopting a comparative approach, we can begin to understand how variation among social systems—by their external environment (e.g., industry of an organization; location of an ecosystem); by their internal attributes (e.g., size, prior history of assembly); or by their antecedents or outcomes (e.g., funding of a platform, performance of an organization)—relates to network structure, and, conversely, whether variation network structure can reveal differences in social processes and outcomes.

Exogenous variation in social systems can yield insights into meaningful social processes. Along organizational dimensions, online communities with larger administrative (moderator) teams on Reddit were more likely to join a collective action protest (Matias (2016)). Traud et al. (2012) compare across one hundred university Facebook networks and find that attending the same high school matters more for network structure in larger universities than smaller colleges. Gee et al. (2017) find evidence of Granovetter (1973)'s paradox of weak ties, but that strong ties are more useful in countries with greater income equality. Recalling the idea of assembly in online systems, shocks can also provide sources of meaningful variation, for example by exogenous shocks (e.g., by hurricanesPhan and Airoldi (2015)) and shocks induced by platform design (e.g., of reward systems (van de Rijt et al. (2014))).

Cross-platform studies provide an opportunity to characterize the exchange of users (Newell et al. (2016) and Tan and Lee (2015)), information (Leskovec et al. (2009)), and connections across

communities (Hill and Shaw (2017)). This supports both an ecological approach (Kleineberg and Boguñá (2015)) and is analogous to prior work on the study of connections between organizations (interorganizational networks) in the organization theory literature (Provan et al. (2007) and Zaheer et al. (2010)). This has been theorized to drive outcomes at the single organization and community level (Kilduff and Brass (2010) and Provan et al. (2007)).

Within a single platform or medium, analyzing $N > 1$ comparable communities can mitigate platform effects. In contrast, cross-platform studies based on multiple $N = 1$ may still end up overfitting to environment-specific attributes: Facebook is different than Twitter for a lot of reasons.[7] Single platform studies can also define a population boundary, which can help define which communities are observed. This is crucial because it is impossible to fully characterize the diversity of online organizations, or correlates of successful communities, without also characterizing systems that did not become successful (Hill (2013)). Hill and Shaw (2017) persuasively argue for the comparative, population-level study of online communities across a single platform, medium or type. To paraphrase, Hill and Shaw isolate five benefits of such a perspective: generalizability of results across communities; the ability to study community- or organization-level attributes and outcomes (see Chapter 4); insight into diffusion between communities, e.g., across platforms or news media; insight into ecological dynamics, extending the organizational ecology approach to online systems (recall Chapter 1.1.3); and insight into multilevel processes, merging individual-level dynamics with understanding meso- and macro-level processes.

We note that this work is largely observational, and is likely to continue in that vein. Experiments across network structures are usually limited to artificial or virtual lab settings (see, e.g., Centola (2010) and Mao et al. (2016)). However, there have been experiments conducted across multiple network structures: to find effective distribution strategies of microfinance loans across villages, for example (Banerjee et al. (2013)). Experiments using virtual labs and the extension of

[7] An early (and well-cited) publication on Twitter was "What is Twitter, a Social Network or a News Media?" (Kwak et al. 2010), which considered the properties of only the Twitter network, with no comparison. In the present context, this title is evocative. Having sufficiently many $N > 1$ examples of networks from varied online platforms would be a step in this direction. But even given a taxonomy of networks (as in Onnela et al. (2012)), the external context of these platforms, differences in user bases, and platform design would all be necessarily relevant to understanding this space.

meaningful experimentation techniques to estimate the effect of social processes across networks (Eckles et al. (2016)) will contribute to future research in this space. The comparative perspective enables research questions that consider changes or effects of policies on platforms, how design changes can induce favorable (and unfavorable) shifts in user behavior, and how groups and organizations can encode their external environment or support productivity. Specifically, this emphasis on outcomes, design, and the ecology, competition, and interaction among systems lends itself towards "solution-oriented social science" (Watts (2017)), suggesting an underexplored opportunity for computational social science.

1.2 Contributions

This work unites and draws on a range of these perspectives. Each chapter considers an empirical setting where populations of instances of social systems can be meaningfully compared. With that comparative perspective, we can consider the role of social processes within and outside of the boundaries of the observed system. As this population-level view has been rare, the range and sources of heterogeneity in these systems is unknown *a priori*. By considering the context of these systems—for example, the educational trajectories and social opportunities of Facebook's earliest users, or the industrial differences and geographical constraints across a population of firms—and making explicit the construction of networks from interaction data, we can provide novel understanding into how these systems vary.

In Chapter 3, "Natural experiments in online social network assembly," we leverage the unique timing and growth strategy from the first two year's of Facebook's existence to reveal multiple sources of exogenous variation in the user base during that time. We apply this comparative, structural perspective to one hundred Facebook friendship networks, representing the individual (and previously distinct) university networks to which Facebook first expanded. Facebook's early design discouraged cross-university connections, and so we take advantage of the shared platform with nearly distinct networks to treat this as a *population* of online social networks. Furthermore, we draw on the ecological notion of assembly to characterize the initial establishment of online

communities. Here, Facebook's iterative expansion strategy created university networks of different ages, and differences in adoption led to systematic differences in structure across this population. Furthermore, the timing of this expansion and the timing of our data snapshot coincided with shifts in the student population, corresponding to students sharing a campus when they gained access to the network, before and after graduation or arrival on campus, respectively. We find heterogeneities across these networks corresponding to attributes of the underlying population (adoption), network size, and context (with respect to graduation timing and arrival on campus). This chapter is based on a previously published paper coauthored with Sam Way, Johan Ugander, and Aaron Clauset, "Assembling thefacebook: Using Heterogeneity to Understand Online Social Network Assembly" (Jacobs et al. (2015a)). It is included here in full, with minor modifications.

In Chapter 4, "A comparative study of informal social networks in firms," I introduce a novel type of data set from a population of large firms. Defining *formal* ties to be the set of hierarchical authority relationships within a company, we define the *informal social network* to be the set of social ties, which may or may not be aligned with the formal network. In organizations, the structure of these networks are believed to play a role in outcomes, and decades of literature from organizational theory, management, and economic sociology have used theory and case studies to characterize this relationship. Using traces of communication patterns within each firm, we empirically explore the structure and heterogeneity of a population of informal social networks, derived from a large email communication dataset, for this population of firms. We empirically explore the theorized relationships between network structure, organizational context, and outcomes. We find a surprising amount of heterogeneity across this population, as well as no evidence of an empirical relationship between the structure and performance of these firms. We do find that size is the primary meaningful variable to characterize structure across these firms, with two notable caveats. First, we find that an employee's average number of contacts does *not* vary with the size of the firm, and we note that this is meaningful both within the organization theory literature and quite broadly across the social networks literature. Second, the centralization of these firms does not vary with size, but it does vary with how firms are geographically dispersed. Overall, this

computational, empirical perspective reveals a large scale of heterogeneity and lack of meaningful correlations; this suggests that there is a diversity of communication structures with which firms can successfully accomplish complex tasks, but also suggests a potential challenge and opportunity for organization theory. This chapter is being prepared for submission for an organization theory audience and is coauthored with Duncan Watts.

Chapter 5, "Empirical network construction: computational perspectives on weak ties, stability, and densification," explores the construction of networks from communication and interaction data. Expanding from comparison across populations of static networks (Chapter 3 and 4) to dynamic networks, we empirically explore a range of hypotheses, empirical observations from single-network studies, and theoretical mechanisms from the social networks literature. This represents the second contribution from this organizational research program: representing another novel large-scale data set that I constructed, we examine the dynamics of the communication patterns from across a population large organizations. We make explicit three variables that are used to derive networks from interaction data. These variables can be implicitly or explicitly chosen, and potentially beyond control by a researcher. We describe how previous theoretical and empirical work has been drawn along these dimensions, and how this perspective can highlight new research questions. We demonstrate the utility of this framework by examining three phenomena, the theory of weak ties, network stability, and network densification, and we use a population of large communication networks of comparable origin to test and explore these ideas. First, considering network tie strength, we find that empirical structure varies in expected ways for weak and strong ties, but that very weak ties are qualitatively different. Turning to network stability, we show that despite global stability of network properties, individual properties vary rapidly over time, suggesting that cross-sectional analyses may be capturing dynamics other than those intended. Finally, we show that this perspective admits us to revisit the decorated concept of network densification. We find evidence for network densification, but we also show that this pattern emerges as an artifact of the level of activity in the online system. Together, this perspective unites past theoretical work and novel empirical results, and reveals a range of tools for exploring the foundations of social network

research from communication and interaction data. This chapter will be prepared for submission for a computer science audience and is coauthored with Duncan Watts.

In Chapter 6, I conclude with a brief discussion synthesizing this work and point to future directions for research in this space.

Chapter 2

Background

This dissertation focuses on empirical populations of networks. Here we define relevant notation that crosses multiple chapters and introduce the data sets employed in the dissertation.

2.1 Notation and measures

2.1.1 Notation: Population-level analysis of networks

Population size N. We describe population size as the number of instances observed and being compared. For example, a population of university networks ($N = 100$), a population of firms ($N = 65$). Following these examples, the organizations are themselves the unit of analysis and may vary in size (e.g., by number of employees).

Network $G = (V, E)$. We formalize networks as graphs $G = (V, E)$ where the vertex set (equivalently nodes) V and edge set (equivalently ties) $E \subset V \times V$ represents pairwise relationships between the elements of V. For our purposes, we will exclude self-edges, i.e., $(i, i) \notin E \; \forall \; i \in V$. We will primarily focus on undirected reciprocated edges, such that if $(i, j) \in E$ then $(j, i) \in E$, and the definitions below will focus on this case.

Size (of a network) S. $S = |V|$ is the number of vertices in an individual network. Networks are expected to exhibit changes in their structural properties based on their size S (e.g., Newman (2010)). We will also be interested in the number of edges, $|E|$, from which we calculate average degree.

Occasionally it may be useful to give $\mathcal{S} = \sum_G S_G$ as the total number of individuals across all

networks in the population. That is, we analyze and $\mathcal{S} = 1.2$ million Facebook users from $N = 100$ universities in Chapter 3 and $\mathcal{S} = 1.4$ million active senders from across $N = 65$ organizations in Chapters 4 and 5.

2.1.2 Network-level analysis: network measures

These chapters employ a range of traditional network measures that are calculated over static networks. Beyond these, there are chapter-specific measures that we define in context. We refer the reader to Newman (2010) for further details and more exhaustive definitions and histories. These definitions assume connected, undirected networks.

Degree and mean degree $\langle k \rangle$. For an individual $i \in V$, the degree of an individual k_i is the number of ties (edges) they are coincident with. This may be in-degree, the number of incoming directed edges $\{(j, i) \in E\}$ for $j \in V$ or out-degree, the number of outgoing directed edges $\{(i, j) \in E\}$ for $j \in V$. For undirected networks as we employ here, these are the same.

The mean degree $\langle k \rangle$ is simply defined as the average over all nodes in the network, that is: $\langle k \rangle = \frac{1}{|V|} \sum_{i \in V} k_i$. Note that this is equivalent to $2|E|/S$.

Density. This represents the fraction of edges that exist out of all possible edges, and is given by $\frac{2|E|}{S(S-1)}$.

Mean geodesic, or average shortest path length L. This represents the average shortest path distance between all pairs in the network. We define d_{ij} to be the pairwise path distance between two nodes in the network, i.e., the number of "hops" across edges to reach each other. For example, if i and j share an edge, i.e., $(i, j) \in E$, then $d_{ij} = 1$. If they do not share an edge, but they do share a mutual neighbor, i.e., $(i, j) \notin E,\ (i, k) \in E,\ (j, k) \in E,\ i \neq j \neq k$, then $d_{ij} = 2$. (Note that this is the *shortest* path, i.e., we always use the minimum length path between two nodes.) Trivially, d_{ii} would be zero. If G is not connected and there exists no path between i and j then $d_{ij} = \infty$. Then the average shortest path length, or mean geodesic, is given by $L = \frac{2}{S(S-1)} \sum_{(i,j) \in V \times V, i > j} d_{ij}$.

Diameter. The diameter of a graph is defined as the longest shortest path in the graph.

That is, the diameter is given by $\max_{(i,j)\in V\times V, i\neq j} d_{ij}$, as defined in the previous definition.

Clustering coefficient C. In this work we employ the global clustering coefficient, calculated as the number of closed triplets $(\{i,j,k\} : \ (i,j),\ (j,k),(k,i) \in E)$ divided by the number of connected triplets ("wedges": $(\{i,j,k\} : \ (i,j),\ (j,k) \in E)$).

2.2 Data

Chapter 3 focuses a new dataset from Facebook's initial founding, 2004–2005. Chapters 4 and 5 use a novel dataset that I developed; I briefly describe this process in Chapter 2.2.2.

2.2.1 Historical Facebook data

In Chapter 3, we introduce a novel data set that augments the "Facebook100" network data set. We received the Facebook100 network dataset with permission from Mason A. Porter and Eric Kelsic. Please refer to (Traud et al. (2011, 2012)) for further detail from their original source. The Facebook100 network dataset represents $N = 100$ disjoint networks, corresponding to the friendship edges within the first hundred universities that received access to the Facebook online social network. The networks are static and were taken from a single snapshot (in September 2005: see Jacobs et al. (2015a) or Chapter 3 for more details).

Specific data sets we constructed are available online[1] . These include:

- Table of inferred dates that each university gained access to Facebook, February 2004–September 2004.
- Table of inferred dates that students physically arrived on each university campus in the fall of 2005.
- University identities associating Facebook100 nicknames to official (IPEDS) data sources.

We additionally use data from the National Center for Education Statistics (Institute of Education Sciences, U.S. Department of Education) Integrated Postsecondary Education Data

[1] `https://azjacobs.com/fb100/`

System (IPEDS)[2] .

2.2.2 Organizational communication data

In Chapters 4 and 5, we use aggregate, anonymized email metadata for a set of 65 U.S.-based organizations from a large enterprise email platform, and we aggregate over a contiguous six month period. We also use organization-level data from the Dun & Bradstreet Hoover's database[3] and MSN Money[4] and our data is drawn from U.S.-based data and organizations. We analyze these patterns in aggregate.

We create a number of minimal requirements for inclusion to cope with noise in each of our data sources. The Hoover's D&B database has inconsistencies and missing data, particularly among smaller companies and branches (children) of the parent organizations. We strictly use data aggregated to the highest level parent organization and we consider the full networks under the highest level parent: we do not separate firms by, e.g., division. Note that we have no meaningful way to measure behavior, either structure or performance, below the parent level, and comparisons at the parent level are then strictly comparable across firms.

We make the following reasonable restrictions:

- Organizations must have at least 100 full-time employees across the entire organization, according to the Hoover's record. (This mitigates noise from within the Hoover's database. This bound also occurs naturally a consequence of choosing publicly traded companies.)

- The Hoover's database must report sales figures for the organization. (Helps manage noisy database: if this was missing, other entries were likely missing too.)

- The organization must be U.S.-based and publicly traded on the NASDAQ or NYSE. (The noisy database is mitigated by publicly available financial information; this increases the available productivity measures and secondary data.) We use only U.S.-based data.

[2] http://nces.ed.gov/ipeds/
[3] http://www.hoovers.com/
[4] https://www.msn.com/en-us/money

- The number of full-time employees, as listed in the Hoover's database, must be within a factor of two of the number of active email senders. (Helps manage merging of noisy database with noisy estimate: not all senders need to be considered full-time employees, and not all full-time employees might be active email senders or information workers.)

Additionally, we restrict our data to exclude companies that were acquired during this time. We also exclude companies that transitioned email coverage during this time period and had relatively stable numbers of active internal senders across the six month time period (moving average of active senders did not change by more than 30%).

For the few data cleaning tasks where a ground-truth organization was needed, we use data from a large company with high ground-truth adoption of the email service across the organization.

To ensure regular usage, we aggregate across all senders to get the average (day of the week, hour) sending volume. This produces a 168-length feature vector of average sending volume corresponding to every (day of the week, hour) pair (Figure 4.1). As a proxy for accuracy of individual coverage, we select only organizations with daily average behavior that varies over the course of the day and is highly correlated with the reference dataset. We use only communication from within firms, such that we can observe the complete interaction patterns and infer the informal networks specific to each firm. This represents about 86.1% of all messages sent (mean: 86.1, standard deviation: 7.3%) across all firms, drawn from 2.1 billion messages in total. Next, we describe how we derive a network from this email metadata.

Comparability and selection bias. We note that these steps allow for *systematic*, apples-to-apples comparisons across organizations. This data curation process necessarily induces bias, as does working within a single enterprise data source. However, by comparing large firms curated with strong restrictions to be of similar origin, we exchange a potentially arbitrarily diverse data set to a smaller data set of 65 firms that are directly comparable. As we are interested in measuring the empirical heterogeneity across these firms, this curation process represents a best effort in restricting the amount of heterogeneity due to the instrument (email communication networks),

rather than among the organizational networks themselves. We discuss the limitations of these choices in context in Chapter 4.5.

2.3 Network construction from email metadata

We describe the construction of the networks inferred from the observed communication data. As choices of threshold and modeling choice in the network inference process could plausibly lead to different outcomes (De Choudhury et al. (2010) and Hofman et al. (2017)), we test the robustness of our results across different choices (Appendix B.3).

We define active senders as those who have sent and received at least one email from within the organization in the six month period. While this is a very weak requirement for what constitutes a 'real' sender, it is sufficient to exclude some presumably-automated and inactive senders. we only include senders in the giant connected component. For network calculations, only senders in the giant connected component are used. On average, this was over 97% of active senders (mean fraction in giant connected component (GCC) in the observed networks: 0.974, median: 0.976, max: 0.991, min: 0.945). This high percentage also serves as validation of the efficacy of the initial data cleaning process.

For each network, we treat edges as bidirected and condition on each edge being reciprocated: i and j must have each received at least one email from each other in the six month time period. As for reciprocity assumptions, this is a very lightweight assumption: there is no restriction on how many recipients were on such a message. This would include, e.g., j responding to a large thread of which i is a participant at any point. Although this was already required by definition of an active sender, we note that this means all senders have minimum in- and out-degree of 1. In an organizational context, again, this restriction is meaningful: we should largely exclude, e.g., automatic distribution lists and automatically generated emails associated with software systems.

We annotate each person-person edge with weights. Following De Choudhury et al. (2010), we use the geometric mean of the number of messages exchanged between each pair, weighted by the number of recipients on each message. Specifically, for each pair of individuals (i, j), and

messages they exchange $I_{ij} = \{\iota_1, \iota_2, \ldots, \iota_{m_{ij}}\}$, aggregated over the full six month time period, we define:

- $m_{ij} = |I_{ij}|$ = total messages sent from i to j
- $m_{ji} = |I_{ji}|$ = total messages sent from j to i
- Reciprocity $\tau_{ij} = \tau_{ji} = \sqrt{\omega_{ij} * \omega_{ji}}$, where

$$\omega_{ij} = \sum_{\iota \in I_{ij}} \frac{1}{\text{number of recipients}(\iota)}$$

  and similarly for ω_{ji}. Note that $\tau_{ij} = 0$ when the link is unreciprocated.

For example, if i emails j directly twice, that will count as weight $w_{ij} = 2$, but if j emails i once when there are two recipients total and once when there are four, that will count as weight $w_{ji} = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$. Then the reciprocity between the two of them will be $\tau_{ij} = \sqrt{2 \times \frac{3}{4}} = 1.22$. Then, one can consider the full weighted social network, or one can modulate the strength of relationships implied by the reciprocity score (De Choudhury et al. (2010)). The strength of relationships observed will also vary with the size of the time window observed: ten emails exchanged in one day vs. ten emails total exchanged over the course of a year vs. ten exchanged daily most days of the year may all imply different types of relationships.

We focus on the network aggregated over all six months of data, although we explore the robustness of our results over different intervals of time and across different thresholds in Appendix B.3. The results here shown are on the networks with minimum reciprocity strength $\tau_{ij} = 1$ and validated on networks with minimum reciprocity strength $\tau_{ij} = 5$. Prior work (De Choudhury et al. (2010)) suggests a heuristic of 5–10 reciprocated emails per year to infer relevant social networks (twice the span of time as here). However, the varied types of relationships captured in an informal organizational network may vary, and the diversity of relationship strengths may take on different meaning in this setting. Other than robustness checks of our results, we leave further exploration of network structure by tie strength to future work.

2.4 Organizational network properties

The distribution of network properties from across the population of networks are given in Table 2.1. The representation of organizations across industries is included in Table 2.2.

Table 2.1: **Properties of the communication networks.**

	Mean	Median	Std. Dev.	Min.	Max.
Number of senders S	21,247	12,732	30,903	4,446	218,986
Average sender degree $\langle k \rangle$	26.9	27.0	8.1	10.6	53.4
Density	0.00261	0.00219	0.00187	0.000132	0.00756
Median sender degree	34.7	34.0	16.5	3.0	91.0
Clustering coefficient C	0.163	0.170	0.047	0.029	0.303
Average shortest path length L	3.17	3.16	0.384	2.37	4.46
Diameter	8	8.34	2.15	6	15
Small world quotient Q	107.9	76.6	90.4	21.7	500.3
Gini coef. of betweenness	0.841	0.841	0.041	0.749	0.938
Gini coef. of degree	0.554	0.540	0.072	0.423	0.783

2.4.0.1 Industry classifications

Industry classifications use the SIC industry code standard. The classification system as applied here is shown in Table 2.2.[5] We use the first two digits of the primary SIC code designation for each firm. Every classification except for Mining, Construction, and Public Administration is represented in our data set.

[5] `http://siccode.com/en/siccode/list/directory`; retrieved May 9, 2017

SIC prefix	No. of firms	Classification
01-09	1	Agriculture, Forestry, Fishing
10-14	0	Mining
15-17	0	Construction
20-39	27	Manufacturing
40-49	7	Transportation & Public Utilities (incl. Communication)
50-51	5	Wholesale Trade
52-59	3	Retail Trade
60-67	5	Finance, Insurance, Real Estate
70-89	17	Services (incl. Technology)
91-99	0	Public Administration

Table 2.2: **Industry classifications for SIC codes.** We use firms' SIC code designation to group firms by industry. The second column reports the number of firms included in this data set.

Chapter 3

Natural experiments in online social network assembly

Online social networks represent a popular and diverse class of social media systems. Despite this variety, each of these systems undergoes a general process of *online social network assembly*, which represents the complicated and heterogeneous changes that transform newly born systems into mature platforms. However, little is known about this process. For example, how much of a network's assembly is driven by simple growth? How does a network's structure change as it matures? How does network structure vary with adoption rates and user heterogeneity, and do these properties play different roles at different points in the assembly? We investigate these and other questions using a unique dataset of online connections among the roughly one million users at the first 100 colleges admitted to Facebook, captured just 20 months after its launch. We first show that different vintages and adoption rates across this population of networks reveal temporal dynamics of the assembly process, and that assembly is only loosely related to network growth. We then exploit natural experiments embedded in this dataset and complementary data obtained via Internet archaeology to show that different subnetworks matured at different rates toward similar end states. These results shed light on the processes and patterns of online social network assembly, and may facilitate more effective design for online social systems. [1]

[1] This text originally appeared in "Assembling thefacebook: Using Heterogeneity to Understand Online Social Network Assembly" by Abigail Z. Jacobs, Samuel F. Way, Johan Ugander, and Aaron Clauset. Originally printed in *WebSci* '15, June 28 – July 01, 2015, Oxford, United Kingdom. DOI: http://dx.doi.org/10.1145/2786451.2786477

3.1 Introduction

Since their emergence in the mid-1990s, online social networks have grown into a highly popular and diverse class of social media systems. This class includes now-defunct systems such as Friendster, tribe.net and Orkut, niche systems such as Academia.edu and HR.com, and large, more general systems such as Facebook and LinkedIn. In contrast to earlier online social communities such as newsgroups (Fisher et al. 2006) and weblogs (Marlow 2004), many modern systems tend to encourage users to transfer offline relationships onto an online setting. Despite the wide variety of these systems—professional vs. personal, contextual vs. general, virtual vs. anchored offline—all of these systems undergo a general process of *online social network assembly* that represents the complicated and heterogeneous changes by which newly born systems evolve into mature platforms.

Relatively little, however, is known about the central tendencies or variability of this process, while such understanding would shed considerable light on the effective design of new platforms. As a result, questions abound. How much of a network's assembly is driven by simple growth processes? How does a network's structure change as it matures? How does network structure vary with adoption rates and user heterogeneity, and do these properties play different roles at different points in the assembly? Are there distinct developmental "phases" to the assembly of these systems?

One reason we lack good answers to such questions is a lack of good data. Traditional online social network datasets fall short in two key ways. First, understanding the effects of different processes requires a network-population perspective, in which many parallel network instances can be examined in order to discern the natural variability of network structure. Second, in the rare situations where populations of networks have been available, such as the National Longitudinal Study of Adolescent Health (Resnick et al. 1997), the underlying social processes do not vary across network instances enough to identify and model different aspects of assembly. By analogy, in social networks recorded from survey questionnaires, it is well-known that different so-called *name generators* (Campbell and Lee 1991)—questions used to elicit social ties—lead to networks

with substantially different structure. As a broad generalization for online social networks, we are interested in the general consequences of variations in the circumstances under which social networks are assembled online.

To understand the structural impact of different assembly processes, we therefore need a population of networks that vary dependably in their assembly. The so-called *Facebook100* dataset (Traud et al. 2012), which is a snapshot of $N = 100$ within-college social networks on Facebook in September 2005, provides just such a population. These networks provide a unique perspective on the very early assembly of a major online social network platform. Crucial to our investigation, these networks vary somewhat in their sizes, characteristics, and history. Each network has a different "vintage," representing a different amount of time between when the college first adopted Facebook and when the snapshot was taken. These vintages, and differential adoption rates across colleges, effectively reveal temporal dynamics of the assembly processes, which we exploit. Finally, a series of natural experiments related to the academic calendar and college characteristics created sufficient heterogeneity at the user- and network-level, which in turn can reveal certain aspects of the underlying assembly processes.

As an example of a natural experiment we can exploit, we note that these 100 colleges joined Facebook sometime between its launch in February 2004 and the end of September 2004 (Fig. 3.1). Because this period spans the end of the 2003–2004 school year, students in some graduating classes of 2004 would have experienced Facebook only as alumni (colleges that joined after graduation) while others experienced it as students (colleges that joined earlier). Comparing the subnetworks of these two groups of students, who should otherwise be fairly similar, with each other and with students of earlier or later graduation years, will shed light on the importance of physical proximity and on-campus interactions in driving network assembly.

As an additional natural experiment, the networks were observed in early September 2005, during the beginnings of the 2005–2006 academic calendars, dates that again vary considerably in this population. As a result, students in the class of 2009 (incoming freshmen in 2005) enrolling at colleges with late start dates (late September) were observed before any significant offline inter-

actions could have taken place (excluding brief summer orientation programs and students from the same high schools). As the students in these classes largely lack any shared historical context, the networks corresponding to colleges with late start dates primarily represent assemblies of relationships formed online, rather than offline. In contrast, students in the class of 2009 enrolling at colleges with early start dates *will* have shared a real world context. This affords an opportunity to ask: how do online social networks encoding online interactions differ in structure from networks that are also encoding offline interactions? We address this by constrasting the classes of 2009 at these early- and late-starting colleges.

By complementing the Facebook100 dataset with the above dates (a modest Internet-archaeological effort[2]), as well as with basic statistics provided by the U.S. Department of Education, we provide a unique, discerning perspective into how online social network structures differ depending on (i) the presence or absence of an underlying offline social network (by studying the classes of 2009), and (ii) the presence or absence of present-time social interactions (by studying the classes of 2004). We also present broad analyses of the population-level variability of network statistics in a general assembly process observed at different vintages. These results shed new light on the general processes that shape social network assembly in online environments, and may facilitate more effective designs of online social systems that relate to the offline world.

3.2 Facebook in the age of Friendster

In 2015, Facebook is today a large and sophisticated social media system, claiming more than 1.44 billion monthly active users (as of March 2015). In 2005, however, Facebook was a very different kind of online social network, in a correspondingly different social media landscape (boyd and Ellison 2007).

Facebook launched at Harvard University on February 4th, 2004 under the name `thefacebook.com`, at a time when the dominant online social networks were Friendster and MySpace. A host of other online college "facebooks" such as CUcommunity, CampusNetwork, and CollegeFacebook were also

[2] These data supplements are available at `http://azjacobs.com/fb100/`

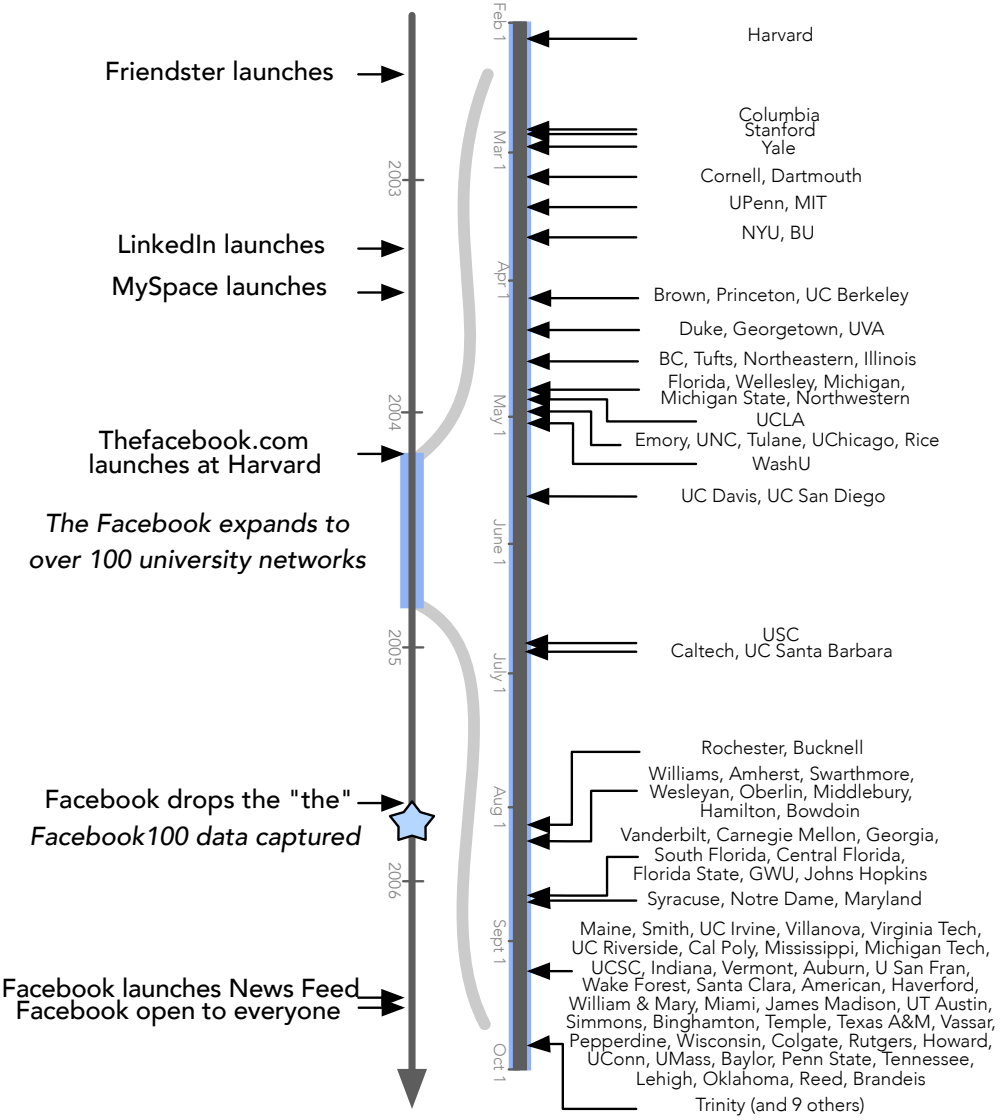

Figure 3.1: Key milestones in the early history of Facebook, including launch dates for the 100 colleges in the Facebook100 dataset.

emerging, in addition to efforts by individual universities to move their student directories onto the Web. Facebook initially limited registration to users affiliated with a sanctioned but growing list of colleges, starting with Harvard (Figs. 3.1 and 3.2). Facebook's popularity spread quickly[3] , and by

[3] *The Daily Northwestern* describes the first 48 hours of Facebook access at Northwestern University thusly: " 'It's an epidemic. . . my whole hall is infected,' said Erica Birnbaum, a Communication freshman. But it's not only one hall. After being available for only about 34 hours, 931 NU students already had registered as of 8 p.m. Monday . . . Such a large quantity of friend request and confirmation e-mails being sent from the Facebook caused Northwestern

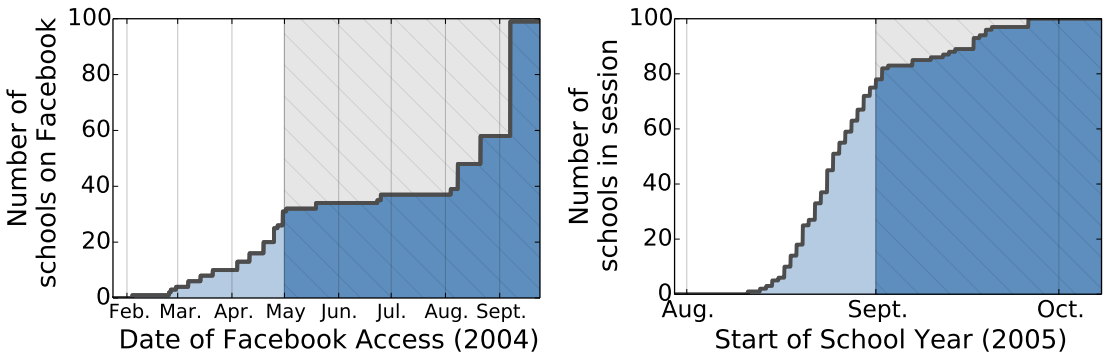

Figure 3.2: The cumulative distribution of schools in the Facebook100 dataset, by date added to Facebook during 2004 (left) and by start of the 2005–2006 school year (right). Shaded regions show how colleges are divided in terms of having received access to Facebook before or after the end of the 2003–2004 school year and whether or not the 2005–2006 school year had begun when the Facebook100 dataset was collected.

the time of the Facebook100 snapshot (September 2005), Facebook had dropped the "the" in its name, opened to over 800 colleges, and had just begun opening itself to high school students. By December 2005, Facebook's user base numbered 6 million, compared to 20 million for Friendster and over 22 million users for MySpace. In September 2006, Facebook opened to all persons over the age of 13.

Description of the network dataset The Facebook100 dataset (Traud et al. 2012) contains an anonymized snapshot of the friendship connections among $S = |V| = 1,208,316$ users affiliated with the first 100 colleges admitted to Facebook, all located in the United States. This comprises a total of $|E| = 93,969,074$ friendship edges (unweighted and undirected) between users within each separate college. Each vertex is associated with an array of social variables representing the person's status (undergraduate, graduate student, summer student, faculty, staff, or alumni), dorm (if any), major (if any), gender (M or F), and graduation year. Across all networks, only 0.03% of status values are missing. Other variables have slightly higher missing rates (gender: 5.6%; graduation year: 9.8%). Dorm and major have higher rates still, which is likely related to

University Information Technology to block all mail sent from the site Sunday night... 'It was viewed as an attack against the network.' " (26 April 2004)

off-campus living and undeclared majors. The completeness of these data reflects the pervading social norms surrounding data privacy expectations in 2005, and possibly a selective bias against users who disliked the default setting of sharing all information within the college network (Acquisti and Gross 2006; Tufekci 2008).

For nearly all colleges, alumni made up about 10–25% of users, a quantity that increased with the age of the network. Vertices labeled as faculty, staff or students who were not regular undergraduates (graduate students and summer students) made up on average 4.1% of each population.

Each college network includes an "index" variable that gives its ordinal position of when it joined Facebook: Harvard is 1 and Trinity College is 100 (Fig. 3.1). For each network, we acquired college-level variables (enrollment, public vs. private, semester vs. quarter calendar) from the Integrated Postsecondary Education Data System (IPEDS) provided by the U.S. Department of Education (National Center for Education Statistics 2014). Full-time undergraduate enrollment from 2007, the earliest date for which data are fully available, was used a proxy for 2005 enrollment.

By dividing the number of undergraduate accounts in each college network by reported enrollment, we can estimate the fraction of students in each network who were on Facebook, a measure of service adoption (Fig. 3.3). In some cases, the estimated ratio exceeds 1.0 as a result of either errors in our enrollment numbers, part-time students on Facebook who were not counted as "full-time enrolled," or multiple/fake accounts at the few colleges that allowed students to control multiple email aliases and circumvent Facebook's initial limits on access.

3.3 Online social network assembly

Online social network assembly is the process by which networks transform from initial creation to a mature online social network. Assembly processes are affected by the composition of the community, online and offline social and behavioral practices, limits on growth (e.g., needing an elite university email address), and competition from other systems, among other mechanisms. Assembly can in part be characterized by the sequence of structural changes that newly-born on-

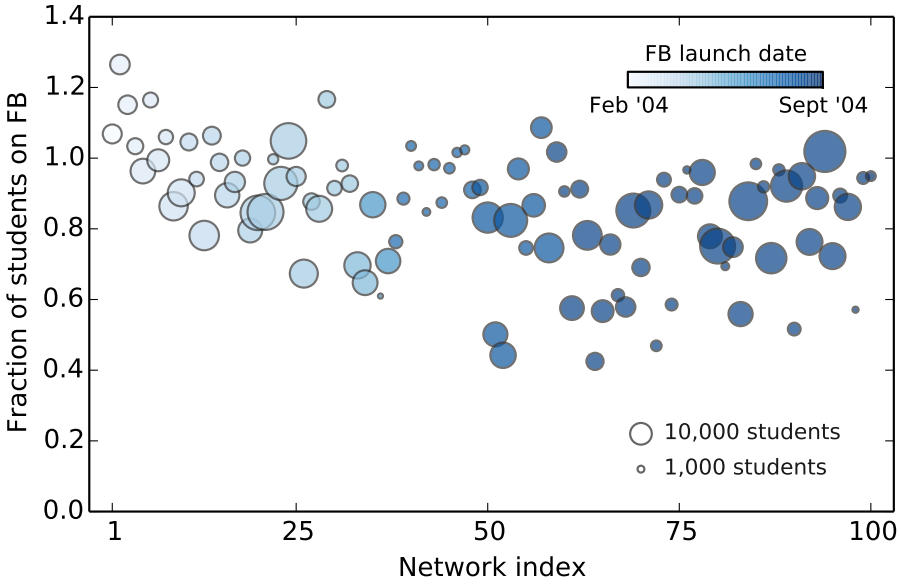

Figure 3.3: Fraction of undergraduates that adopted Facebook vs. network index. Vintage is visualized with network index, the order in which schools were given access to the site. Size corresponds to the size of the undergraduate population. Color indicates the date on which schools were opened to Facebook.

line social networks undergo as they mature. In particular, this area of study aims to identify and model the underlying social processes that guide assembly, and to identify the 'developmental' patterns that are common across different networks. Here, we focus on the role of network growth, user heterogeneity, adoption rate, and network 'vintage' in shaping these assembly patterns. We examine the impact of these elements on structural patterns in the networks, e.g., their degree distributions, clustering coefficients, diameters, and community structure (Ugander et al. 2011), as well as understanding how those patterns change under network growth (Backstrom et al. 2012), how they vary across subpopulations within the network, and what social processes govern these patterns and variations. We note, however, that reliably connecting observed patterns with the correct underlying processes can be complicated, as different processes can sometimes lead to similar, or even identical, structural patterns (Mitzenmacher 2004).

By comparing patterns across these networks, we aim to characterize the scale and sources of

natural variation. Here, several observable features of Facebook's early college networks—differing potential network sizes, ages, adoption and heterogeneity of context—play important roles in shedding light on its early assembly. First, its staged expansion among colleges during 2004 produced a population of online social networks of different vintages, at schools of different sizes, within which the service was adopted at different rates. Second, the graduation year annotations identify subpopulations that changed identity during the time observed, e.g., different classes that joined or left the campus environment.

Processes and models of assembly Online social network assembly is a special kind of network evolution. Most techniques and statistical models developed for analyzing the structure of temporal networks (Holme and Saramäki 2012), however, cannot be applied to the Facebook100 data because these networks are not snapshots of a single evolving system. Instead, we will exploit the several ways that temporal information is embedded within the observed network structures and represented in their covariates, e.g., vintage and adoption rates at the network level and graduation year at the vertex level.

The simplest model of assembly is network growth, in which the number of vertices and edges grow monotonically in some way. Several simple models of network growth exist, including many variations on preferential attachment (Kumar et al. 2010), in which new users join the network and create connections with existing users with probability proportional to those users' current degree; randomly grown networks (Callaway et al. 2001), which are related to classic random graph models; and the forest-fire model (Leskovec et al. 2005a, 2007), which is related to preferential attachment but produces both greater local clustering and a shrinking diameter. Crucially, these models assume that assembly is a homogeneous process, and thus network structure changes uniformly across all subsets of vertices (Schoenebeck 2013). In contrast, the assembly patterns of real online social networks are likely to be considerably more heterogeneous, both at the vertex level and at the network level. These models thus hold value primarily as theoretical reference points in our analysis.

Social surveys of early Facebook users provides some hints about the processes governing its assembly, and support our claim that assembly in real networks is unlikely to be simple or

homogeneous. One survey from 2006 found that students of different graduating years had different usage patterns, and that older students—those whose college careers were mostly over by the time Facebook arrived on their campus—were less likely to adopt the service (Tufekci 2008). Thus, local network structure is likely to vary by graduating year. Several surveys also found evidence that online connections on Facebook among current students generally reflect pre-existing offline relationships (Lampe et al. 2006; Mayer and Puller 2008). This implies that Facebook's early assembly should reflect the inhomogeneities of real-world social processes, which depend on factors like age, gender, and being on campus.

From a theoretical perspective, the social processes that seem likely to influence assembly in these networks can be divided into two major dichotomies: offline/online processes and contemporary/historical processes. In the first case, offline processes are those driven by relationships in the offline world that are then transferred to an online setting, while online processes are confined to mechanisms mediated by digital interactions alone. In the second case, contemporary processes are those that reflect social events in the present time, while historical processes are those where the formation of links in the online social network is driven by pre-existing relationships that are brought online.

These classes represent different ways that social connections can be recorded in online networks, and are orthogonal to the social processes that drive link formation, such as homophily, social status, or strategic behavior. For instance, triadic closure—the event in which two people who have a mutual friend, but who are not themselves currently friends, become friends—can drive relationships in the past or present, because closing a triad can occur at any time, and can be mediated by either offline or online interactions. Different endogenous or exogenous forces can also shape the assembly of a particular online social network. For instance, features like Facebook's "People You May Know" module influence which links form by facilitating the transfer of offline relationships to the online network (Zignani et al. 2014b), while competition from other systems can impede or reverse link formation altogether (Ribeiro 2014). The systematic loss of links, and more generally the decay and disassembly of online social networks is a related but distinct re-

search domain, as disassembly processes are not simply assembly processes in reverse (Bascompte and Stouffer 2009; Garcia et al. 2013).

Here we focus on three distinct types of social processes in our data, and how they relate to the network assembly of early Facebook: (i) the transfer of offline historical friendships to the online environment (Ellison et al. 2007), (ii) the formation of connections that reflect present day and offline interactions in the college environment, and (iii) connections formed purely online. We expect to observe a mixture of these processes, and the corresponding patterns they induce, across our network population. Furthermore, because past work suggests that Facebook connections, from the very start, reflected offline social interactions (Lampe et al. 2006; Mayer and Puller 2008), we expect that networks further along in the assembly process will more closely resemble complex offline social structures. We expect strong differences in how quickly different Facebook subnetworks assemble, for instance between students and alumni, because students often live together, take classes together, socialize and work together and alumni generally do not.

Network growth due to accretion, in which existing users invite their friends to the network, and due to triadic closure mechanisms would tend to make the more mature subnetworks appear more dense, with higher mean degrees, and lower mean geodesic distances than less mature subnetworks. We expect the differences between subnetworks to decrease with older vintages. In addition, we expect different subnetworks to mature at different rates, unlike previous work that focuses on homogeneous processes (Schoenebeck 2013).

Finally, given Facebook's role in 2005 as a campus-oriented social network, we expect that adoption among undergraduates can be used as a proxy for maturity of the network. As the early design was to facilitate within-campus interactions, the college online social networks would grow by adding new users and increasing the connections among them. High adoption indicates the online social network would be nearing its effective finite limit for the undergraduate network.

3.4 Vintage, growth, and adoption in network assembly

To begin our analysis, we first test how changes in network structure are related to network size, network vintage, and service adoption. While the domain of study about network growth investigates the relationship of network properties to network size, it is an open question whether network assembly can be strictly explained by network size or network vintage, the relationship to which is not obvious *a priori*. We thus expect to see either no relationship between a particular measure of network structure and age—in the case that the corresponding network property is roughly stationary under the assembly process—or a simple relationship—in the case that the property is gradually modified with age. Alternatively, if assembly is equivalent to simple growth, as in traditional network models of growth, we expect to see certain specific relationships between network measures and network size. We evaluate these two competing hypotheses by examining the relationship of standard measures of network structure, such as mean degree, clustering coefficient, mean geodesic distance, and degree assortativity with network size S and vintage.[4]

We find that these networks as a population exhibit the classic "small world" pattern found in many social networks, with small pairwise distances and relatively high average clustering coefficients, capturing the frequency of triangles to length-two paths (Watts and Strogatz 1998). Specifically, the mean geodesic distance (average length of a shortest path) scales like $O(\log S)$ with network size S, while the clustering coefficient scales like $O(1/S)$,[5] seemingly towards a modest constant as an asymptotic end state (Fig. 3.4); in contrast, neither mean geodesic distance nor clustering coefficient varies clearly with vintage. The rising mean geodesic distance with S, and its independence of vintage, contrasts with the graph densification literature (Leskovec et al. 2005a), which predicts a falling distance with size or time, and it is instead consistent with basic theories for random graphs, which predicts a $O(\log S)$ behavior. The fact that a densification pattern is observed in Facebook several years later (Backstrom et al. 2012) suggests that online social net-

[4] For clarity, we visualize the schools by network *index*, corresponding to the order in which schools were added to Facebook. In these cases we overlay the color corresponding to the date added (Fig. 3.1), thereby vintage is monotonically increasing with index.

[5] In light of the $O(\log S/S)$ scaling we uncover in Chapter 4, future work should explore a broader range of hypotheses across this population of graphs.

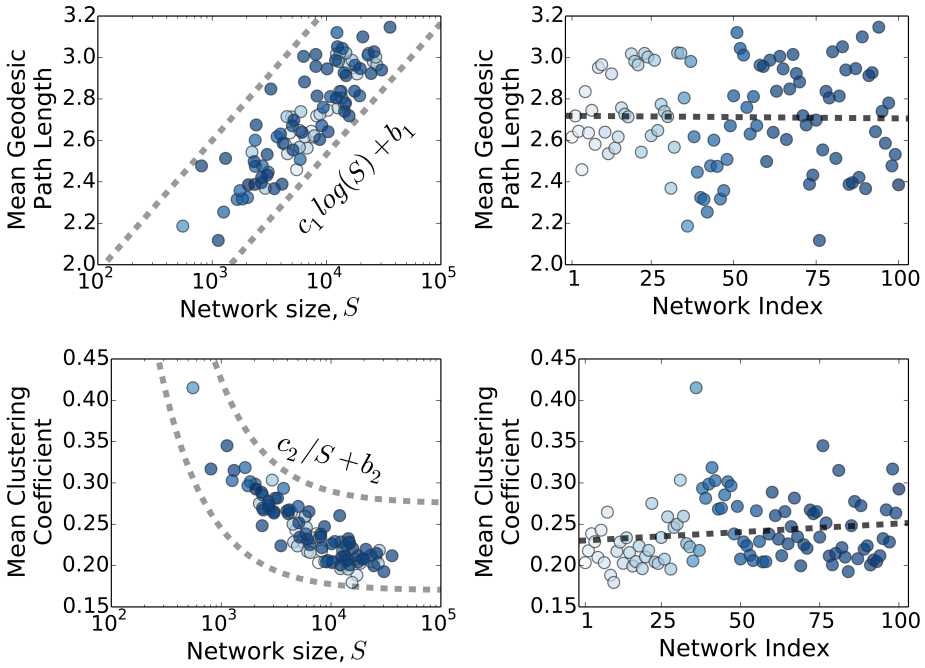

Figure 3.4: (top) Mean geodesic distance (shortest path length), and (bottom) mean clustering coefficient ordered by school size S and by network index. In agreement with results from random graph theory, the mean geodesic distance varies like $O(\log S)$ and the clustering coefficient varies like $1/S$. Color indicates the vintage of the network by date added. Dashed lines show an ordinary least squares fit to the data, demonstrating little to no trend between network features and vintage.

work assembly may go through distinct developmental phases, with an early phase of sparsification, resembling a growing random graph (Callaway et al. 2001), that is followed much later by densification. The falling clustering coefficient pattern observed here, which is expected in random graphs but not in social networks (Newman 2010), supports this hypothesis.

We examine several other measures of network structure, such as mean degree; assortativity on vertex degree (Pearson correlation of degrees between connected pairs); and modularity by gender or major. Modularity quantifies whether pairs share an attribute more than expected by random (positive) or less (negative) (Newman 2010). We find very weak or no correlation with network size or network vintage (Fig. 3.5). The lack of any clear relation with size and vintage for these measures supports the notion that the online social network assembly process for Facebook college networks is not uniquely explained by size and vintage. That is, assembly is more complex

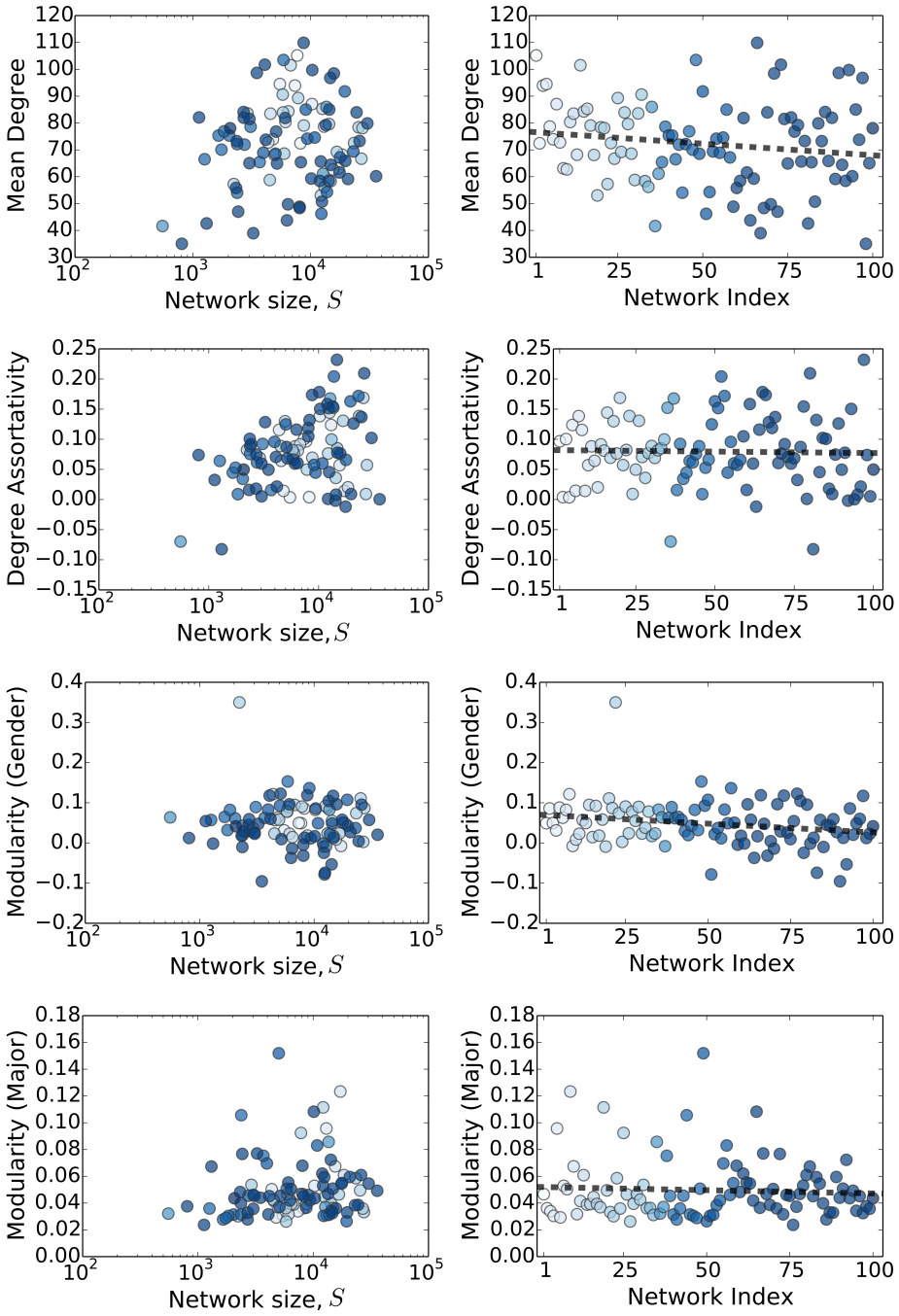

Figure 3.5: Relation of various network features to network size and network index. Colors indicate the vintage of the network by date added. Dashed lines show an ordinary least squares fit to the data, demonstrating little to no trend between network features and vintage.

than simple growth or network vintage.

As Facebook was introduced to different colleges, each school's online social network grew within a finite social space, limited by the size of the student population. The fraction of service adoption describes the relative growth in these populations and is therefore a plausible measure of the maturity of each network in this context. We expect to see more clear correlations between measures of network structure and the maturity of a networks's assembly process. (Because adoption levels are estimated only among students, we restrict these analyses to the induced subgraph among student vertices.) In Fig. 3.3 we find a relationship between vintage and adoption. We also find that as adoption increases, the normalized mean geodesic distance, i.e., the distance divided by the overall $O(\log S)$ pattern, tends to decrease slightly (Fig. 3.6). That is, the greater the level of adoption, the shorter the paths between a pair of individuals, controlling for network size (Fig. 3.4). Thus, adoption, rather than size, may be a better measure of the maturity of a network under assembly. Furthermore, this supports the two-phase developmental process, in which path lengths should grow during a sparse growth phase, and become on average shorter as the network densifies.

The degree distribution is a network description of great interest, with social networks frequently exhibiting heavy-tailed degree distributions. A consequence of this heavy-tailedness is the unequal distribution of mean neighbor degree to mean degree (Kooti et al. 2014). For regular graphs this ratio is one, while for all other degree distributions it is necessarily greater than one. We use this ratio as a proxy for the heavy-tailedness of the degree distribution, and find that degree distributions become less heavy-tailed as networks mature (Fig. 3.7). That is, even though the mean degree of a random neighbor of a vertex and the mean degree of a random vertex both tend to increase with adoption, the mean degree of a random vertex grows slightly faster as a network matures. This pattern is consistent with the two-phase developmental pattern suggested above, where an initial phase of sparse growth with many new vertices and comparatively few connections are added, and then followed by a densification phase, where new connections are mainly added between existing vertices.

Together, these results argue that network assembly is not simply network growth, or vintage,

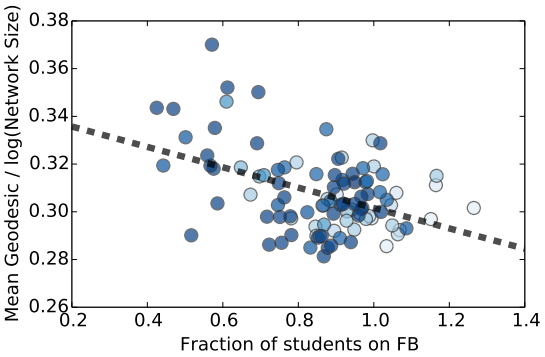

Figure 3.6: Even after controlling for size, the mean geodesic distance decreases with adoption in undergraduate networks. Color corresponds to the vintage of the network by date added.

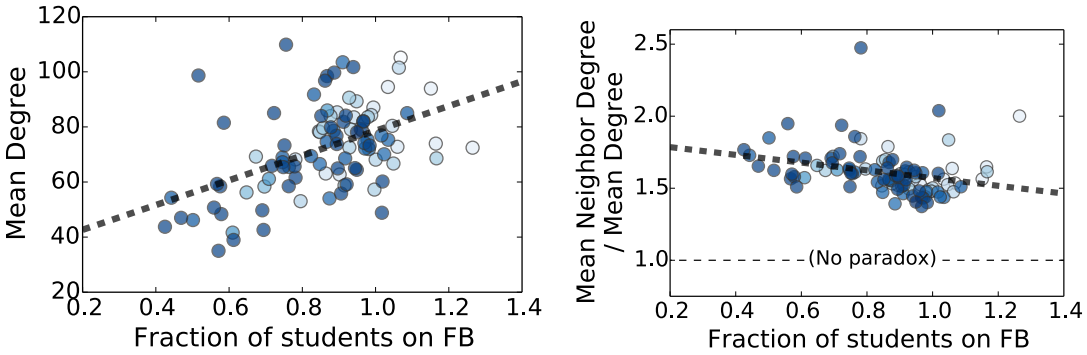

Figure 3.7: Mean degree increases and degree distributions become less skewed in more mature networks, shown here by adoption rate. Color corresponds to the vintage of the network by date added.

or adoption, and furthermore, that the Facebook100 networks are drawn from a single online social network assembly process. However, heterogeneity of the network assembly processes is induced by differences in network size and network adoption. The Facebook100 networks can provide useful insights into how these mechanisms interact, and heterogeneity within subpopulations of these networks can potentially reveal greater insight into the assembly mechanisms at play.

3.5 Heterogeneities from natural experiments

Accidents of history and the timing of our snapshot induced auspiciously observable heterogeneities in the online and offline assembly processes of our population of college networks. In this section, we examine these heterogeneities as natural experiments to explore the variability in online social network structure due to differing processes. These natural experiments are useful because they let us examine how different subpopulations of users differ in their connectivity, which lets us identify the detailed processes by which these networks assemble.

We begin by first examining basic differences among different subpopulations defined by graduating class year. We then use the timing of the arrival of freshmen on campus (in 2005, at the time of the snapshot) and the arrival of Facebook on campus (in 2004, either before and after the class of 2004 graduated) to investigate the maturity of the online social networks more precisely. Finally, we find that the subnetworks that had less time to mature (due to environmental and historical reasons) share broad structural patterns with the university networks that had lower adoption rates.

We first look at differences among the undergraduate population (Fig. 3.8). The classes of 2008 and 2009 arrived on campus as freshmen in the fall of 2004, at a similar time or after Facebook, and thus formed their offline and online social networks almost concurrently. Previous work found that classes with more established offline networks prior to Facebook's arrival had observable differences in behavior: survey research conducted within our sample showed that the classes of 2008 and 2009 were more likely than the classes of 2006 and 2007 to form offline friendships as a result of online friendships (Ellison et al. 2007). On the other hand, for the classes of 2006, 2007, and 2008, students had access to Facebook for a similar amount of time, so these networks should have had equal opportunity to assemble. Thus, we can investigate the roles of time and offline social context among these classes.

Between the classes of 2006, 2007, and 2008, we observe that the class of 2006 has notably lower mean degree, a more skewed degree distribution, and higher modularity by major. The lower

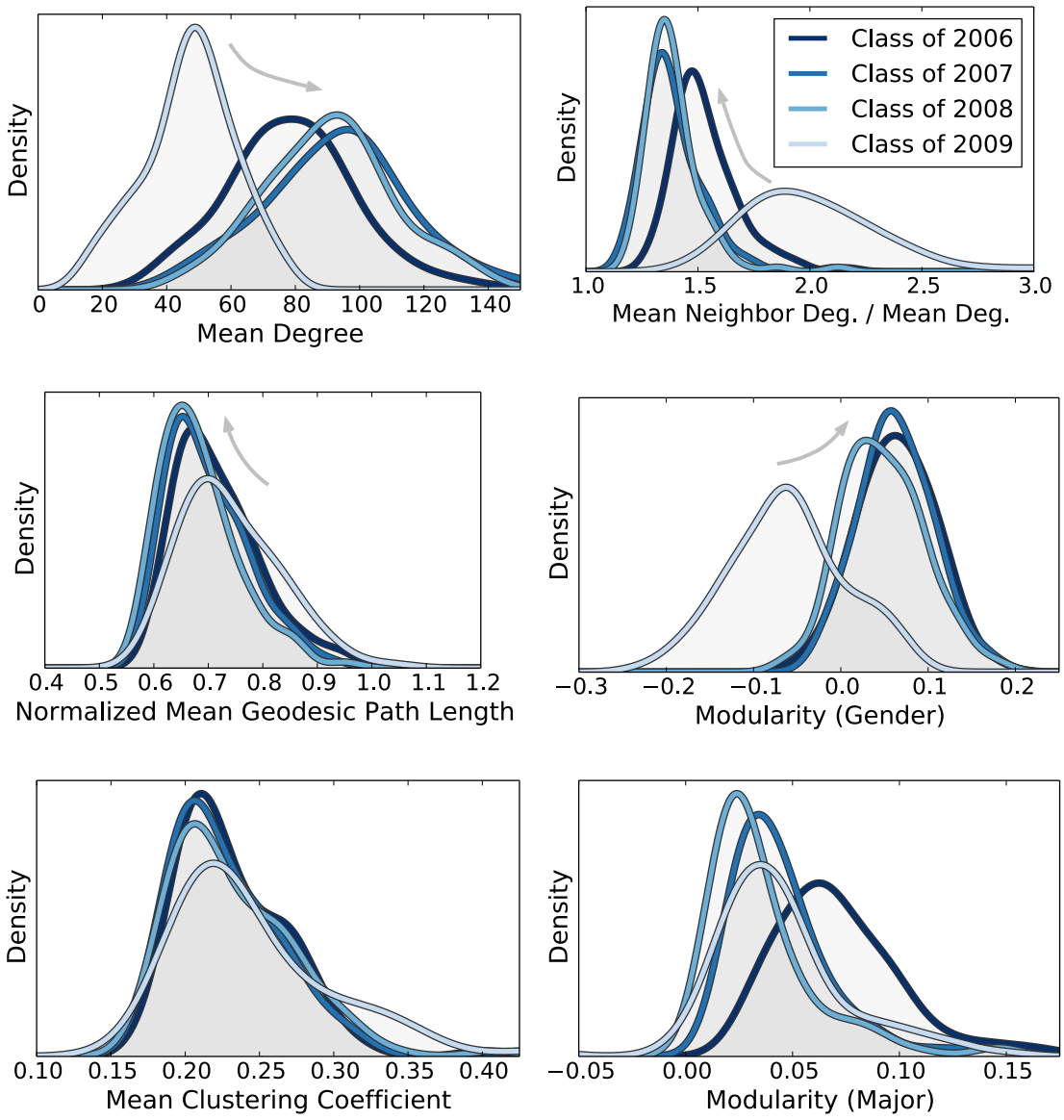

Figure 3.8: Distributions of undergraduate network features across the population of 100 schools, by graduating class. Distributions are visualized using kernel density estimation. Arrows move from class of 2009 to classes of 2007 and 2008, the classes with the highest adoption, when the difference between those distributions is statistically significant (two-sample KS test, $p < 0.01$).

mean degree and higher skew are consistent with a less mature network, possibly due to lower engagement (Tufekci 2008), while the higher modularity by major suggests that these upperclassmen simply mix less across majors.

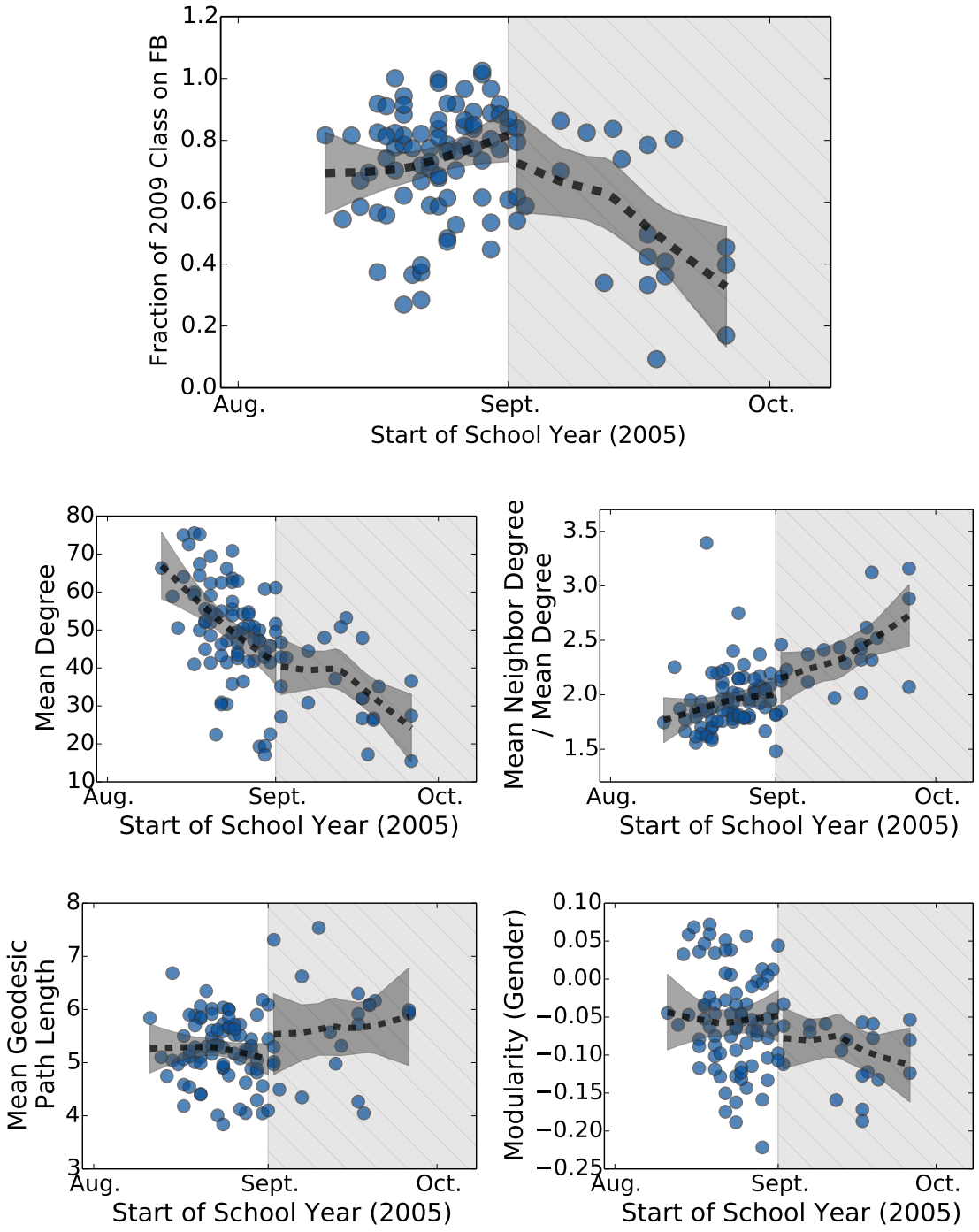

Figure 3.9: Network features ordered by date new students arrived on campus, August–September 2005. The snapshot was taken in early September 2005 (gray). The dashed lines are LOESS curves over schools that began before and after September 1, shown with 95% confidence intervals about the mean.

Class of 2009 natural experiment Across most statistics, the most strikingly different distributions are those that describe the class of 2009 networks (Fig. 3.8). The class of 2009 primarily began their undergraduate careers in the fall of 2005, when the snapshot of our data was taken. As these new students only recently gained university affiliations, the class of 2009 networks would have had the least time to develop. Notably, a fraction of these classes would have arrived on campus before the snapshot was taken, and those classes could have an offline basis for their online friendships.

Overall, the class of 2009 networks have lower average degree, more skewed degree distributions, and are disassortative by gender, whereas the older classes are assortative by gender. Studying these differences at the distributional level, it is not clear whether the differences we see in Fig. 3.8 are the result of the reduced vintage of these subnetworks, with students having only joined Facebook during the summer of 2005, or some difference of assembly connected to the principally online interactions that formed these networks. Enter the natural experiment.

Students enrolling in the fall of 2005 generally obtained access to Facebook during the summer of 2005, in conjunction with obtaining university email addresses. Activity on Facebook for students not yet on campus was essentially limited to online "social browsing" (Lampe et al. 2006), as they possessed no offline context yet to motivate "social searching." Through Internet-archaeological research, we gathered the calendar dates that incoming freshmen arrived on campus in 2005 at the 100 involved colleges to discover if and to what degree the observed differences in network structure could be connected to opportunities for offline interactions (Fig. 3.9). We first observe a strong relationship whereby the networks for new students who have spent more time on campus—but similar amounts of time socializing online—are more mature. Students that have spent more time on campus have higher mean degree, less skewed degree distributions, as well as higher adoption overall. Interestingly, we find strong evidence for a pattern of social browsing focusing on the opposite gender: students that have spent more time physically together, and thus are more actively engaging in social search, are more gender assortative than students that have primarily interacted online.

Controlling for the size of the freshman networks, there are three data points of particular interest: Northeastern, Caltech, and Tulane. At Northeastern, most undergraduates are enrolled in programs that are explicitly five-year programs: that is, students identify at the outset as having a five-year graduation date. (This is in contrast to most colleges, where students enter identifying with a four-year graduation date, despite potentially longer times to completion.) For the Northeastern networks, the class of 2009 shares properties well-aligned with the second year (sophomore) students at other schools; this should be expected, as most of the members of the Northeastern class of 2009 began college in Fall 2004, not 2005. Caltech, meanwhile, is known to have an exceptional social environment among the schools in the Facebook100 dataset, as was studied closely in earlier work (Traud et al. 2011, 2012). Caltech is an outlier on almost every network metric including clustering coefficient and modularity by dorm. The structure of Tulane's class of 2005 has at play unique external events, namely the massive disruption due to Hurricane Katrina, which hit New Orleans on August 29, 2005. Tulane freshmen ultimately spent very little time physically on campus, but may have coped with this significant event by connecting through the medium of Facebook during the early days following (Phan and Airoldi 2015).

Class of 2004 natural experiment Shifting our focus to the opposite temporal end of the dataset, the alumni in our sample reflect a diversity of social, spatial and cultural settings, and notably lacked the opportunity for closed mixing within university campuses. In Fig. 3.10, we consider three graduating classes of students: 2003, 2004, and 2005, which in sum comprise on average 84.4% of the alumni users at the time of the snapshot. (Less than 5% of alumni have observable earlier class years; 11.4% of the alumni do not report their class year.) We first investigate differences between these three classes, of which the class of 2005 spent almost a full year with Facebook while colocated on campus; some of the class of 2004 gained access to Facebook before graduating (Fig. 3.1), a distinction we will explore more deeply next; and the class of 2003 only having gained access to Facebook after graduation. We analyze the induced subgraphs of these alumni classes, and find that the more recent alumni networks are more mature, and furthermore that the class of 2004 network appears to represent a maturity level intermediate to the class of

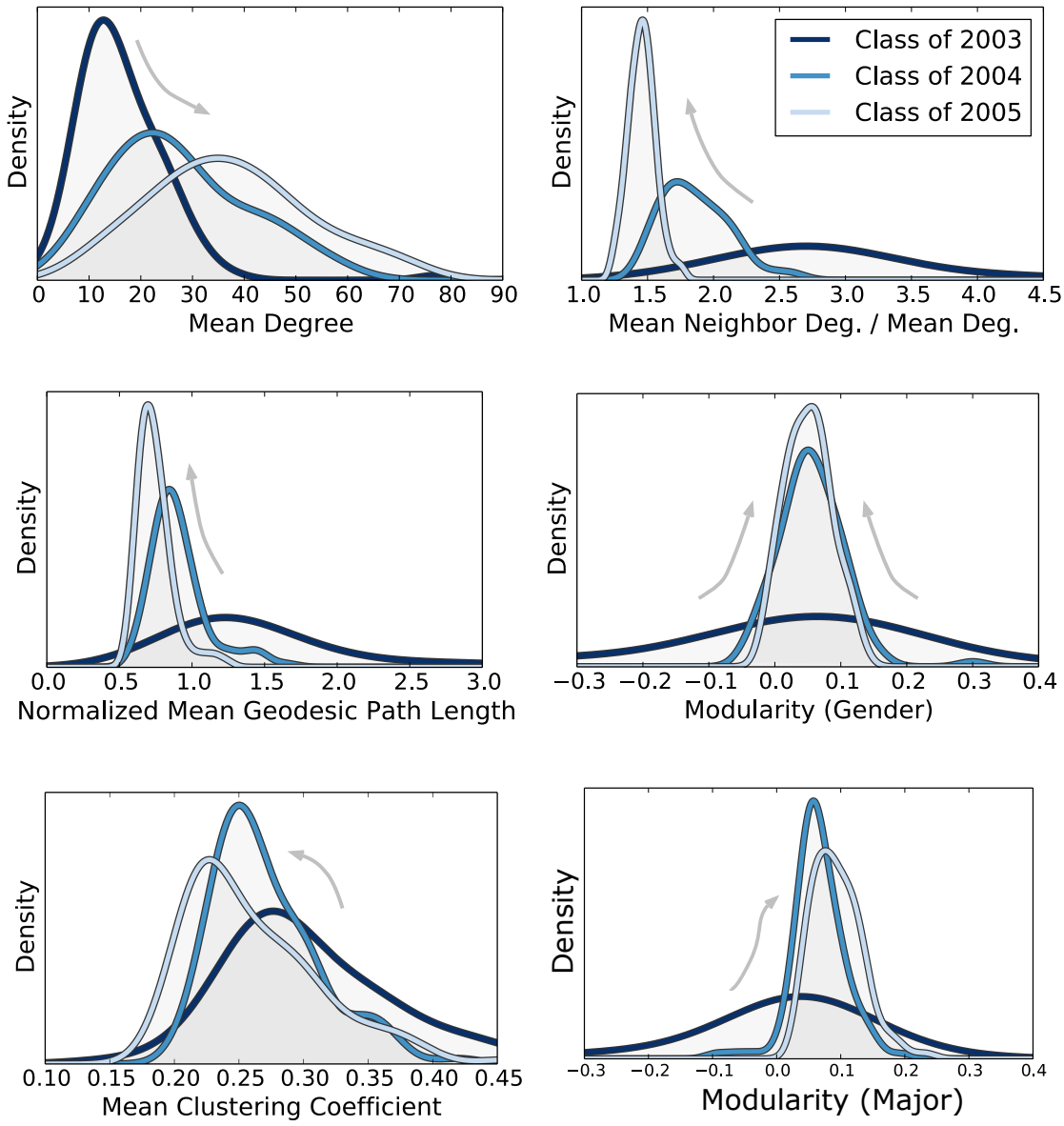

Figure 3.10: Distributions of alumni network features across the population of 100 schools, by graduating class. Distributions are visualized using kernel density estimation. Arrows move from the class of 2003 (lowest adoption) to the class of 2005 (highest), when the difference between those distributions is statistically significant (two-sample KS test, $p < 0.01$).

2003 and 2005. This smooth transition suggests that the university environment induces additional online assembly of the offline social networks being captured.

The graduating class of 2004 primarily finished their undergraduate careers during May and

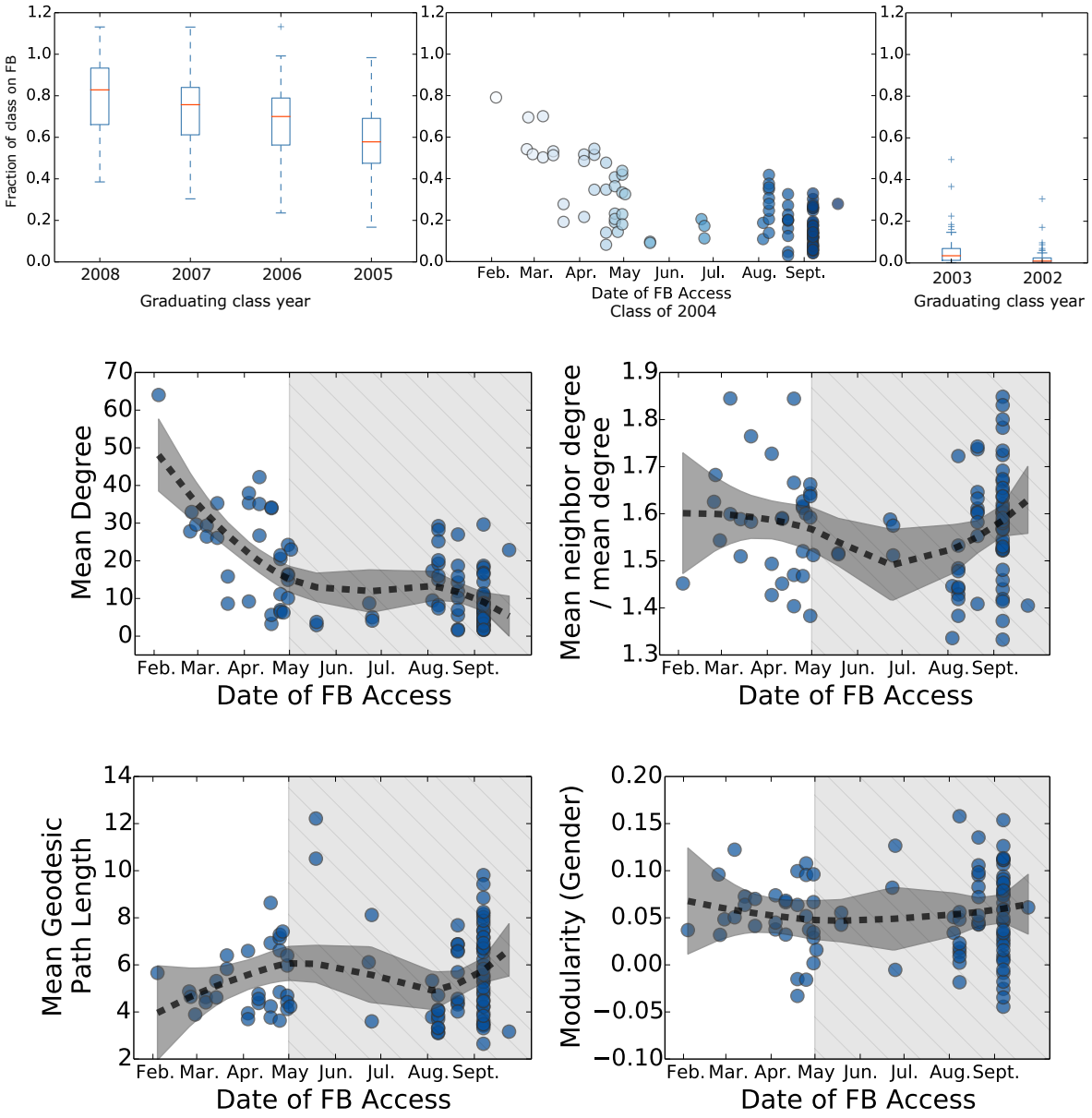

Figure 3.11: (top) Network adoption for different class years. The boxes are bound by the 25th and 75th percentiles, and the center line is the median. (top center) Network adoption for each university network by the class of 2004, ordered and shaded by date the university gained access to Facebook. (below) Network properties for the class of 2004 by date of access to Facebook. The shaded region separates classes that graduated prior to gaining access to Facebook, and the dashed lines are LOESS curves, shown with 95% confidence intervals about the mean.

June of 2004. Concurrently, Facebook was spreading to increasingly many campuses, with students at Harvard (id=1) graduating after several months on Facebook, and the University of California

San Diego (id=34) after just a few weeks. Of the 100 colleges in the sample, 66 did not gain access to Facebook until after the class of 2004 graduated, so those new alumni would no longer share the university environment when they joined.

Again using Internet archaeology—primarily via the Internet Archive, the Spring 2004 Media Kit from TheFacebook LLC, and student newspapers—we collected the dates that universities joined Facebook in order to tease apart the effects of the university environment on the early growth of the Facebook network. Across the different school networks, the class of 2004 student populations have approximately constant demographics, and the first 34 schools are comparable by size, public/private status, and geography compared to the remaining 66 (Figs. 3.1, 3.3). Thus, other things being equal, we can examine the impact of the arrival of Facebook on the network assembly of the class of 2004.

At the time of the snapshot, over a year past most students' graduation and granted access to Facebook, adoption still tracks strongly with the arrival time of Facebook (Fig. 3.11). We also find that mean degree correlates with arrival time, both of which suggest that the offline and cohesive social environment played a role in the rate at which these networks grew. Other variables did not exhibit a strong trend throughout this transition. This negative result suggests that the class of 2004 networks were of relatively constant maturity level. Arguably, this maturity level interpolates between the classes of 2003 (whose network assembly was almost exclusively outside of the college environment) and 2005 (whose graduating students were able to connect while on campus) (Fig. 3.10), whereas the size of the network was largely determined by the amount of time in a shared offline context. This suggests that the initial transition into alumni status realized a similar level of complexity of existing offline social structures, as opposed to the sharp transition exhibited among freshmen arriving on campus, with a discrete start time and novel social connections. This suggests that the type of shared offline context plays a significant role in the trajectory of network assemblies.

3.6 Discussion and conclusions

The large size, early rise, and storied history of Facebook make it a model system for studying the processes and patterns of online social network assembly, i.e., the complicated and heterogeneous changes these networks undergo as they mature. The Facebook100 networks capture a special part of this history—the first 20 months, the first 100 colleges, and the first one million users—which allows us to investigate the early stages of assembly. Our analysis sheds new light on the extent to which a network's assembly is driven by simple growth, how a network's structure changes as it matures, how network structure varies with adoption, and how the connectivity patterns of different groups of users tends to converge, at different rates, on similar end states.

Each of these results depended on our using a population of social graphs to measure distributions of structural statistics, which allowed us to better estimate the natural variability of network structure produced by the underlying social processes. In contrast, many other studies rely on a single network instance, which makes it difficult to identify whether some pattern reflects a general insight or a special case. Many questions and tasks in the analysis of networks would benefit from this kind of population approach.

Applied to the Facebook100 data, this approach revealed several novel insights into the assembly of online social networks. First, these graphs exhibit a clear $O(\log S)$ dependence for the mean geodesic distance (Fig. 3.4). This pattern agrees closely with conventional wisdom, which is largely drawn from classic results in random graph theory, but it defies recent claims about general "densification laws," which predict shrinking rather than growing distances. These results are not, in fact, contradictory, and instead suggest that online assembly proceeds through two distinct phases.

Initially, a network grows via sparsification, adding many new vertices from the extant population and a relatively smaller number of connections among them. For early Facebook, each time a new college joined, or a new class arrived on campus, this phase started anew within that population and proceeded as the adoption rate rose from zero. The second phase begins once the

network has expanded to include a large fraction of the available population. Then, assembly transitions into a densification pattern, adding many connections among existing vertices and a smaller number of completely new vertices. Of note, these two phases can be seen as corresponding to the growth and saturation phases of logistic growth within a finite population (Barrat et al. 2008).

Past work on distances in the large-scale Facebook social network (Backstrom et al. 2012) corroborates our finding: the mean geodesic distance between users peaked in 2008 and subsequently shrank, illustrating a transition into a densification pattern around that time. Between its opening to the general population in 2006 and 2008, Facebook was expanding rapidly into new populations, and our findings imply that its large-scale structure grew according to a sparsification pattern. The 2008 transition to densification implies that Facebook's expansion into new populations began to slow then, allowing continued link formation to begin to densify the network.

We find further evidence for this same two-phase pattern within the Facebook100 networks, distributed across different subpopulations, which experience network assembly at different rates but toward similar end states. By combining these networks with additional information about Facebook adoption rates, and college graduation and matriculation dates, we leveraged two natural experiments within these networks to show how structure varied between students on and off campus, between students of different graduating years, and between alumni and current students. Each of these analyses showed a consistent behavior: the longer a subpopulation had access to Facebook, especially for students on campus, the greater its level of adoption. As adoption increases we see distances shrink, degrees increase, and degree distributions becomes less heavy tailed.

This model would predict that just before Facebook opened up to the general population in 2006, the network structure within each of its college subnetworks was very mature, having reached high levels of adoption. Opening up to a wider range of users, however, moved the system as a whole back into the sparsification phase. As Facebook spread into this large and unadopted population, its diameter expanded and its degree distribution became more heavy-tailed, before transitioning back into the densification phase, as a greatly enlarged system, in 2008.

The specific processes by which online social networks assemble are also implicated by our

results, which sheds new light on several understudied questions about networks. The online assembly process described above tends to sample offline individuals and relations (Schoenebeck 2013), a pattern supported by social surveys of users at the time (Tufekci 2008). Online social networks that specifically reflect such offline relationships are thus different than those based on mainly online interactions. For instance, consider assortativity by gender among new students (Fig. 3.9): those who had not yet arrived on campus tended to connect with students of the opposite gender. In contrast, those on campus tended to connect with those of the same gender, which is the pattern observed among older students already on campus. That is, the former group did not have the offline social interactions to ground their behavior in reality, and thus treated Facebook very differently—apparently, like a dating website—than on campus students embroiled in the rich offline social milieu of college life.

Looking forward, it seems clear that designing or modifying online social networks is a task best done with a detailed understanding of how different social factors and processes influence the particular trajectory that assembly takes, both at the level of individual users and at the level of the entire network. That is, human behavior is not independent of the design of these systems, and designs are likely to be more effective and more useful if they are informed by an understanding of their impact on the long-term structure and function of these networks. The study of online social network assembly promises to shed new light on these tradeoffs.

Acknowledgements This work was completed with coauthors Sam Way, Johan Ugander and Aaron Clauset. This work was supported by the NSF Graduate Research Fellowship award no. DGE 1144083 (AZJ) and the Butcher Foundation (Sam Way). The authors thank Mason A. Porter and Eric Kelsic for providing the Facebook100 data and Leto Peel for useful discussions. The authors acknowledge the BioFrontiers Computing Core at the University of Colorado Boulder for providing High Performance Computing resources (NIH 1S10OD012300) supported by BioFrontiers IT.

Chapter 4

A comparative study of informal social networks in firms

> "Within a firm, informal networks are as an important a factor of production as its financial capital, buildings, and machinery and the human capital of its personnel."
> Flap et al. (1998)

Informal communication networks within firms represent the connections among workers across which information and organizational learning are transferred, status is exchanged, and social, mentoring, and administrative relationships are shared. Despite wide attention to the implications of the patterns of these relationships, little is known about the natural heterogeneity across communication networks in firms, and how the structure of those networks varies with attributes or performance of the organization. We present a comparative study of high resolution, within-firm communication network structure across 65 U.S. based, publicly traded firms of varying industries, sizes, and formal organization. We find a high level of heterogeneity across organizations, where within-type variation exceeds between-type variation. We largely find a lack of relationship between organizational type and network structure, as well as between network structure and organizational performance. The primary meaningful variable related to informal network structure is its size—although we find that the average individual's number of contacts, while varied, does not depend on the size of their organization. We find that centralization in firms does increase with geographic and organizational dispersion through increases in information bottlenecks, but surprisingly, we find no relationship to firm size or age. The scale of this heterogeneity and lack of meaningful correlations suggests that previous results based on case studies may reflect over-fitting. This novel empirical perspective suggests a potential challenge and opportunity for organizational theory.

4.1 Introduction

In "The Comparative Study of Organizations," Blau 1965 laid out the need for the "systematic comparison of a fairly large number of organizations in order to ... determine relationships between attributes of organizations." Such comparisons are necessary to advance the field of organization theory, he argued, and if undertaken, the initial empirical tasks would be straightforward: for example, measuring relationships between centralization and firm size, age, and industry. Such studies would allow the testing of existing organizational theory, as well as establish a basis from which to advance new theories. This motivation is both foundational and uncontroversial: in Blau's setting, "the comparative method, in the broadest sense of the term, underlies all scientific and scholarly theorizing," just as it does today.

Within organizations, patterns of communication reflect diverse types of relationships: status, power, knowledge transfer, collaboration, mentorship, and friendship all are encoded within these ties. The efficacy and success of organizations and individuals is widely thought to vary with the structure of these relationships within the organization. As one author said simply, "Social networks matter" (Kleinbaum 2017). The capabilities, responsiveness, and innovation of a firm are all thought to vary with the structure of these informal networks (Ahuja 2000; Kleinbaum and Stuart 2014; Krackhardt and Hanson 1993; Srivastava 2015). Others have gone so far to assert that, "within a firm, informal networks are as an important a factor of production as its financial capital, buildings, and machinery and the human capital of its personnel" (Flap et al. 1998). As a management tool, Krackhardt and Hanson 1993 assert that the informal network "is the central nervous system driving the collective thought processes, actions, and reactions of its business units" in a firm.

However, these organizational networks are, at best, difficult and expensive to observe at any level of analysis. The difficulty of observing communication patterns has meant that empirical studies are typically within a single organization, and frequently within a subset of that organization. Then even in this case study setting, we often still lack a direct test of communication network

structure across an organization. This is still useful—case studies provide a proof of the existence of social processes encoded by social structure—but they cannot reveal the range or sources of heterogeneity across firms. The field has lacked systematic, rigorous and direct tests of the relationships between informal social network structure in firms and organizational attributes, including performance, and this sentiment has been shared in organization theory from Blau 1965 through today (Davis 2015a). Although it has been recognized that new sources of private, firm-level data could make novel comparative analyses possible (Davis 2015b; George et al. 2014), such analysis has not yet been conducted with modern in-depth, firm-level data. The simple comparisons laid out by Blau to understand populations of firms, although well theorized, have not been empirically measured, or even measurable. The fundamental diversity of informal networks in firms is unknown.

Using a novel data set of email communication networks across a population of firms, we characterize the natural variation among informal social networks and the degree to which these communication patterns relate to organizational context and performance. The diversity of communication patterns within a single organization is thought to be significant for organizational outcomes, but it was previously impossible to directly characterize the degree to which communication patterns vary across organizations without a comparative, data-driven perspective. We examine the structure of communication patterns of organizations using email sending behavior from 65 U.S. based, publicly traded firms comprising 1.8 billion email exchanges among 1.4 million employees. We compare the internal sending patterns and structure across organizations of varying industry, level of productivity, and size, covering almost two orders of magnitude of firm size.

We find wide heterogeneity across firms. The primary meaningful variable for explaining network structure is firm size; however, we find that both average degree and centralization do not vary with size. Furthermore, once we account for firm size, this heterogeneity among firm networks is not explained by organizational context—industry, firm age, or dispersion—with one exception. We find that, as projected by theory, dispersion increases at a declining rate with firm size. Dispersion in this setting captures geographic dispersion and, by proxy, differentiation across the formal network structure. We find that network centralization increases with dispersion: that

is, although power does not become more concentrated with firm size or age, it does with dispersion. Finally, we find no relationship between network structure and firm performance.

Although our exploratory approach is unsuited for making causal claims (Button et al. 2013; Gelman and Loken 2014; Simmons et al. 2011), we argue that it has some important advantages for organization theory. First, by operationalizing and testing many possible relations, our approach provides a general template for empirical studies that seek to make claims about the causal effects of network structure on other organizational properties or the reverse. In place of vague or implicit assertions, that is, a comparative structural approach enforces explicit hypotheses about the relation between say, average path length (L) and size (S), even to the point of specifying a mathematically precise functional form (e.g. $L \propto S$ vs. $L \propto \log S$). Second, by making comparisons across many firms we avoid the mistake of generalizing a relation from a single case, or even a comparison of two cases. As we show in Section 4.4.2, simple heterogeneity across firms can easily yield misleading conclusions from small-N comparisons. Finally, our overall conclusion that firms exhibit large amounts of heterogeneity on almost any metric, and that very little of this observed heterogeneity is explainable in terms of any of the metrics that are commonly invoked in the literature, poses interesting challenges both to organization theory and to future empirical analyses.

4.2 Informal social networks in firms

Informal networks in firms are composed of the interpersonal social, communication, or trust relationships along and across the underlying hierarchy of formal roles in an organization. These informal networks—or emergent networks as they are sometimes called (Aldrich and Pfeffer 1976; Monge and Contractor 2001)—are defined in contrast to formal networks, which represent the hierarchy of relationships of formal authority. While there has been a rich history of understanding firms through formal network structure (March and Simon 1958; Weber 1947), this hierarchy does not represent all important relationships within a firm. (Simon 1962, in the process of arguing for the fundamental role of hierarchies in complex systems such as firms, still notes that "the formal hierarchy only exists on paper" (p. 468).) Communication and social interactions, and related

processes of power, information flow, and status, are informed by this hierarchy but not exclusively contained by it; these formal networks have been insufficient to explain the behavior and efficacy of firms (Monge and Contractor 2001). Moreover, empirical studies such as Rice 1994 confirm the overwhelming impression from lived experience that communication in organizations, while related to formal network structure, deviates from it in important ways. Communication networks represent a broad class of networks within organizations (Monge and Contractor 2001), and email networks, in particular, have been established to be a useful empirical tool to understand informal social networks in firms (Kleinbaum 2008).

In light of the impression that is conveyed by the literature on informal social networks that such networks contribute in consequential ways to important features of organizational behavior and performance, it is surprisingly difficult to identify in the literature any clear consensus on precisely how they should matter or with respect to what. In part, this lack of conceptual clarity derives from vaguely worded assertions that imply an effect without articulating a clearly testable hypothesis, and in part it derives from coexisting claims that appear to stand in contradiction with one another (e.g. denser networks predict higher/lower innovativeness), where these contradictions have not subsequently been addressed. In large, part, however, these problems themselves derive from the sheer diversity of network studies. Exploring the premise that network structure matters, we consider the inherent variability across the potential interpretations and hypotheses of such a statement, which vary with respect to at least three dimensions:

(1) The composition of organizational networks (e.g., individuals, teams, or organizations) and the relationships represented within them.

(2) The level of analysis (e.g., ego/individual, full network) at which structure is measured. The level determines what structural measures are used for analysis and to whom network effects are hypothesized to accrue.

(3) The variables of theoretical interest (e.g. performance, innovativeness, size, organizational type) and the direction of the proposed causal relation between the network and other

variables. The latter determines whether network structure is a predictor of some outcome variable of interest (e.g. performance) or is itself an outcome of some other variables (e.g. organizational type).

To clarify the space within which our contribution is situated, therefore, we next describe variation along these dimensions in more detail.

4.2.1 Organizational networks

The literature takes a number of interpretations of organizational network structure, varying by the unit of analysis (individual, team, organization) and scope of relationships (within or across organizations). The study of organizational networks broadly encompasses networks, for example, of individual shareholders; interorganizational networks of firms within an industry sector; intraorganizational networks composed of teams and team-level interactions, as well as intraorganizational networks composed of individuals and individual-level interactions; see Kilduff and Brass 2010 for a wider discussion. We contrast two types of analysis induced by whether relationships are within or between organizations. This choice of boundary frames the research questions available.

Interorganizational networks. Almost all network studies that consider multiple organizations are done at the interorganizational level. That is, the networks represent relationships between firms. For example, the networks of corporate stakeholders may predict how organizations respond to conflicting interests (Rowley 1997); patterns of ownership ties reveal the resilience of a national industry (Kogut and Walker 2001); and patterns of collaboration reveal industry dependence on strategic alliances for learning (Powell et al. 1996). While interorganizational networks are frequently used to characterize individual firm positions and outcomes, for example through strategic alliances or access to resources (Gulati et al. 2000), the network-level analysis of the structure of interfirm relationships can characterize the robustness of an industry, constraints on growth, or the structure of business groups (Granovetter 1994; Provan et al. 2007).

Intraorganizational networks. Firms contain underlying hierarchical structure of formal authority relationships. While these hierarchies are a key component of organizations and organizational communication, social interactions happen both along this hierarchy and across it (Monge and Contractor 2003). The patterns of relationships *within* individual organizations, i.e., the informal social networks, reveal how individuals are distributed across a firm, as well as how information, status and governance flow through the firm (Wellman and Berkowitz 1988). The capabilities and adaptability of a firm are then related to the structure of the informal network (Argote and Ingram 2000; Hansen 1999), suggesting a relationship to the performance of firms.

4.2.2 Level of analysis

Ego-level analysis. Network position of an individual can determine differences in social capital, status, or access to information (Burt 2000) and has been shown to be related to individual-level outcomes. The opportunities of an individual (ego) may be based on its relationships to its neighbors (alters), or more broadly to its neighbors' neighbors (Uzzi 1997). This perspective has been applied across types of networks: Balkundi and Harrison 2006 found, for example, that the centrality of both the ego position of a leader in a team and the position of a team within an organization are positively associated with team effectiveness; at the level of organizations and their inter-organizational relationships within industry sector, a moderate amount of embeddedness increases firms' chances of survival (Uzzi 1996).

Network-level analysis. In contrast, the global structure of social networks can reveal the underlying patterns of relationships within them (White et al. 1976). Measures such as the density, small-world structure or transitivity of a network can reveal the organization within a team, firm, or industry; centralization, for example, describes how power and status might be distributed across an organization. This can translate into system-level outcomes: Uzzi and Spiro 2005, for example, finds that the structure of collaboration networks among artists led to differences in innovativeness; Powell et al. 1996 describe how the interfirm network was related to the trajectory of the biotechnology; and Provan et al. 2007 summarize the literature on network-level analysis

in interorganizational networks. Within firms, Kleinbaum and Tushman 2007 find that social ties across organizational units are necessary for innovation across multi-divisional firms. Looking at individual organizations, this type of analysis has connected the structure of informal networks to organizational outcomes (Kleinbaum 2008).

4.2.3 Network structure as predictor and outcome

Networks in organizations are understood to matter: Cross et al. 2002, for example, argues that "critical informal networks often compete with and are fragmented by such aspects of organizations as formal structure, work processes, geographic dispersion, human resource practices, leadership style, and culture," but these networks are of "strategic and operational value." Networks appear here, and in the academic and management literature, in two ways: as an outcome of those organizational processes, and a predictor of value or success of the organization. As an outcome, communication network structure in firms is shaped by a range of mechanisms, including the formal organization, distribution of tasks, and access to communication technologies. As a predictor of outcomes, variations in network structures are tied to innovativeness or performance. We address work in both of these directions.

Network structure as dependent variable: unknown scale and sources of heterogeneity across firms

While network structure has been theorized to vary with organizational properties, such as centralization with size and dispersion, even the degree of variability across firms is unknown. DiMaggio and Powell 1983 famously counter Hannan and Freeman 1977's observation that there is a diversity of "internal structural arrangements" (specifically: "Why are there so many kinds of organizations?") by instead investigating the "startling homogeneity of organizational forms." Lacking empirical baselines, the pairwise comparison of firms cannot reveal the sources of heterogeneity or whether networks are meaningfully different. This is further reflected in choices made during analysis: for example, selecting thresholds for the presence of small-world structure across firms of different sizes (Baum et al. 2003).

Organization size as an organizational attribute is understood to have a role in the structure of organizational communication. This has been argued by Weber 1947 and in many forms since, but this requires teasing apart the concomitant role of size in organizational structure (Krackhardt 1994a) from function and performance (Child 1973). At the same time, potentially independent of any social process, network measurements are expected to vary with network size; these scaling patterns are well understood in random graph models but lack thorough empirical validation (Newman 2010).

Beyond size, Pugh et al. 1968 and colleagues argue that organizational context is an important predictor of the structure of communication (Child 1973; Pugh et al. 1969). They argue that although organizational context as "of primary importance in influencing the structure and functioning of organization[,] there have been few attempts, however, to relate these factors in a comparative systematic way to the characteristic aspects of structure" (p. 91). In the time since, differences in structure have been attributed to industry, geographic dispersion, and age (e.g., Cross et al. 2002; Hannan and Freeman 1984) or otherwise seek to mitigate differences by making intratypical comparisons (e.g., Blau and Schoenherr 1971). However, it is unclear the degree to which heterogeneity among firms is captured by these differences.

Network structure as independent variable: firm-level outcomes

There is a lively tradition of linking ego network structure to outcomes across all types of networks, including a range of firm-level financial performance outcomes (Shipilov and Li 2008). For linking network-level structure to network-level outcomes, there is tradition using interorganizational networks to measure the performance, productivity or robustness of an industry or state (Provan et al. 2007). For network-level studies of intraorganizational networks, the organization theory and strategy literature is rich with studies but sparse with direct tests of the relationship of between network structure and firm outcomes; Kleinbaum and Stuart 2014 lay out a compelling argument for the relationship established by the literature for the role of intraorganizational networks for firm performance.

Leana and Van Buren 1999, for example, argue that stable within-firm relationships can

improve firm outcomes, but can potentially impede the spread of innovation. Intraorganizational networks can mitigate the effects of geographic dispersion to improve innovation (Lahiri 2010). Alcácer and Zhao 2012 argue that firms with stronger internal linkages between research divisions yield competitive advantages in co-located R&D settings. These claims echo the theory that the efficacy of knowledge transfer within a firm is dependent on internal network structure, which in turn impacts productivity in large firms (Argote and Ingram 2000). Finally, where crises have induced changes to network structure (e.g., by increasing centralization), the structure of firm social networks and communication processes has been argued to impact decision-making with firm-level financial outcomes (Romero et al. 2016; Staw et al. 1981).

This perspective, linking networks within firms to organization-level outcomes, fits between the local level of individual outcomes and wide interfirm relationships. At this "meso" level, it has been argued that "there is clear evidence that individuals and groups substantially influence macro organizational phenomena" House et al. 1995. And yet, these claims have been limited by obvious "empirical obstacles to persuasive tests of the effect of intrafirm networks on firm-level performance," i.e., lacking "network data from many firms to test such a theory" (Kleinbaum and Stuart 2014). The conflict is then when, and how, organizational context is relevant to the large-scale patterns of individual communication, and to what degree the relationships among individuals is related to firm-level outcomes (Smith et al. 2006). Despite this, large, comparative in-depth studies across organizations are exceptionally rare.

This dichotomy of networks as consequent or antecedent makes the endogeneity of these tasks readily apparent. A shift in communication technologies, firm size, or geographic dispersion can induce changes in the informal network; however, changes in firm performance might encourage such a shift, for example, through a reduction in firm size through reorganization during a less productive year. Access to similar resources may induce homogenization across firms. There may be an additional loop, if the strategies implemented are based on the measurement of network structure to begin with (Healy 2015). While these concerns are not novel (cf. Davis 2010; Kleinbaum and Stuart 2014; March and Sutton 1997; Smith et al. 2006), this theoretical framework has lacked

direct empirical tests across firms.

4.2.4 The present work

To summarize, it is clear from the extensive literature on informal social networks in organizations that such networks are thought to matter both in theory and in practice. Somewhat more precisely, it is clear that organization theorists believe that informal network structure of an organization is related both to the context and to the performance of the organization (Ahuja et al. 2012; Burt 1997, 2004; Flap et al. 1998; Granovetter 2005; Ibarra et al. 2005; Kilduff and Brass 2010; Kilduff and Tsai 2003; Kleinbaum et al. 2013; Kogut 2000; Krackhardt 1994b; Krackhardt and Hanson 1993; McEvily et al. 2014; Nohria and Gulati 1994; Reagans and McEvily 2003; Reagans and Zuckerman 2001; Rice 1994; Romero et al. 2016; Williamson 1994). In spite of these evident beliefs, a broad reading of the literature on organizational networks does not yield clear and specific hypotheses that admit to testing in the type of a large-N comparative study of whole-organization networks that Blau 1965 called for. Instead, it surfaces a large number ambiguously-connected claims that operationalize key constructs in inconsistent ways, conflate qualitatively different units of analysis, select (and correspondingly exclude) variables of interest in an ad hoc manner, and take opposing positions on whether networks should be treated as inputs or outputs, and mix implicit with explicit statements of cause and effect[1] .

In the absence of any consensus on which specific features of network structure should be related to which other organizational properties, and in what manner, we instead adopt an exploratory, comparative approach. Specifically we leverage a unique dataset of intra-organizational

[1] We note that our assessment of the literature broadly echoes a series of previous assessments (Davis 2010, 2015a; Lewin and Minton 1986; Miner 1984; Schwarz et al. 2007; Smith et al. 2006) in which the authors have observed that although organization science has generated many individually interesting theories of organizations (Miner 1984 identified 110 distinct theories while Lewin and Minton 1986 identified 13 distinct bodies of theory), when viewed collectively it has failed to make clear and testable predictions about the relationship between organizational structure, context, and performance.

Nor are we alone in this specific frustration. Dalton et al. 1980 asserts"the literature on structure-performance relationships is among the most vexing and ambiguous in the field of management and organizational behavior." Thirty years later, in "Do Theories of Organizations Progress?" Davis 2010 finds "sloppy operationalizations" where "organizational researchers need to take measurement more seriously by making explicit the link between constructs and indicators, for both substantive and control variables."

email logs to conduct a comparative structural analysis of 65 US publicly traded firms to address two broad questions of general theoretical interest:

Q1: How does informal network structure vary as a function of organizational properties, such as size, age, industry, and dispersion?

Q2: How does an organization's performance vary as a function of informal network structure?

To make these questions precise and testable, we require clearly defined and quantifiable metrics that operationalize three key conceptual constructs: network structure, organizational properties, and organizational performance. For network structure, we draw from the networks literature a set of canonical network measures, representing communication diversity, average shortest path length, clustering, and centralization. For organizational properties, we refer to the organizations literature, where size, geographical dispersion, age and industry are commonly cited covariates of interest. And for performance, we refer to the industrial organization literature, in which return on equity, return on capital, % annual revenue growth, and revenue per capita are frequently invoked as metrics.

4.3 Data

Email communication data has been shown to reflect offline communication patterns (Kossinets and Watts 2006; Wellman and Haythornthwaite 2008). Specifically to firms, email communication networks encode the underlying formal structure as well as communication beyond the hierarchy (Adamic and Adar 2005; Kleinbaum and Stuart 2014; Monge and Contractor 2001; Rice 1994). These email sending patterns reveal traces of interpersonal, inter-group, and cross-organizational relational behaviors (e.g., Ducheneaut and Bellotti 2001; Fisher 2004; Tyler et al. 2005; Wuchty and Uzzi 2011).

Our goal is to comparatively study informal networks with respect to organizational attributes; however, the natural heterogeneity of informal networks in firms is unknown. Despite this interest in informal networks, we only observe email communication data. As part of our data collection process, we require that firms must have used email across the firm at comparably high

levels. This mitigates unknown biases in communication usage across firms, but as a result, our data set has far fewer companies than a more forgiving data collection process.

Using aggregated, anonymized email metadata from 1.8 billion messages and almost 1.4 million senders from a large enterprise email system, we derive the communication networks for 65 firms. We combine this with firm-level attributes from the Dun & Bradstreet Hoover's database. We collected properties of the firms as they were reported coincident with our communication data collection time period. Hoover's collects information about the full family tree of organizations within a firm; we restrict our data to the properties of the global parent of the organization.

To construct the networks, we first identify the set of organizations for which we have representative email communication data. This entailed a significant data cleaning effort, described in Section 4.3.1, to require that all firms had consistent, high email usage across the firm and across the time period observed. Lacking ground truth across most firms, we use a reference data set for which we have known high and consistent engagement across the firm. This effort prevents unintentionally capturing structural differences due to variation among email adoption and usage, rather than meaningful structural differences among informal networks. This cleaned subset of organizations comprises consistently active email users across the organization and across the window of observation (Section 4.3.1). We then describe how we construct networks from communication data (Section 4.3.2).

4.3.1 Dataset construction

We use aggregated, anonymized email metadata corresponding to U.S.-based firms, covering a six month time period. We hand-verified all identities of the organizations to align with the records in the Dun & Bradstreet Hoover's database, and restrict to only active, U.S.-headquartered, publicly traded firms.

We restrict our data to within-organization communication, such that we can observe the complete interaction patterns and construct the informal networks specific to each firm. This represents about 86.1% of all messages sent (mean: 86.1, standard deviation: 7.3%) across all

firms, drawn from 2.1 billion messages in total. Within this data set, we seek to restrict to only human-like active senders within active organizations (comprised of active human senders covering a reasonable amount of the organization). For these organizations, we characterize the distribution of daily and weekly sending patterns, the time series of daily behavior over the six month time window, and the intraorganizational email networks. We identify 65 large public companies for which we have active adoption and consistent use across the six month time period.

We include a number of restrictions to bound the noise from each data source and their combination. These restrictions also help insure that, without access to ground truth, the communication data we observe is reasonably representative of communication inside of a organization. They also insure that there is sufficient information to describe each firm's performance and organizational attributes. Specifically, we make the following requirements, with additional detail in Chapter 2.2.2.

- The Hoover's database must report sales figures for the organization.
- The highest level of the organization must be U.S.-based and publicly traded on the NASDAQ or NYSE.
- The number of full-time employees (as listed in the Hoover's database) must be within a factor of two of the number of active email senders.

The last requirement is due to merging two noisy sources of data. The number of active senders counts senders who have sent and received at least one internal email during the window of observation: this would exclude, for example, internal distribution lists and external marketing addresses, but also employees who use personal email accounts. Full-time employees is a well defined, but not necessarily representative, count of the number of active, email-sending employees in a firm. There are multiple sources of variation here. Organizations with a large number of contract or temporary workers may have a lower number of reported full-time employees than expected, but a high number of active email accounts. On the other hand, organizations with many retail or food service workers might have many official full-time employees but few active

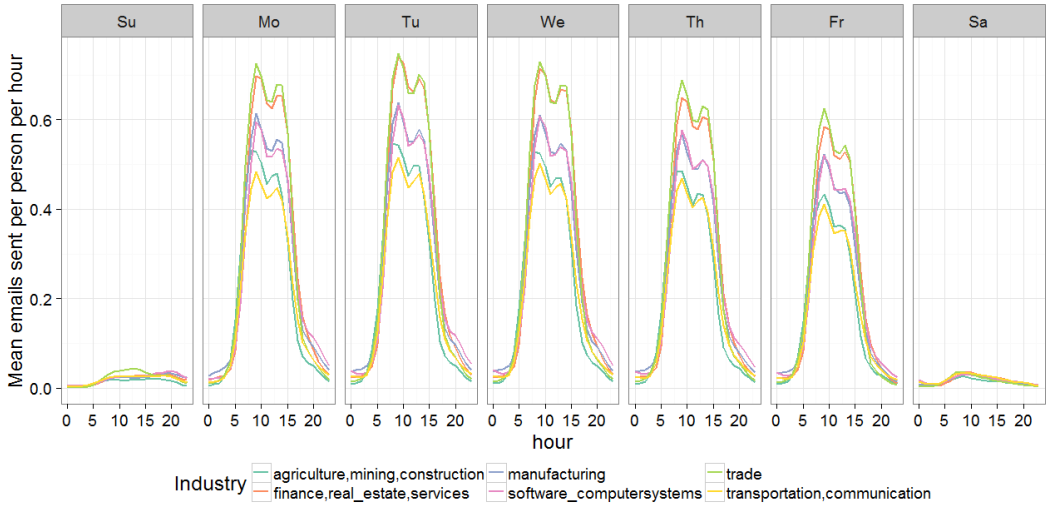

Figure 4.1: Average number of messages sent per hour by day of the week. Trade and services organizations send the most mail during peak times. In addition to the regular daily morning, lunch and afternoon pattern, the evening volumes are higher at the beginning of the week, and small peaks on Saturday morning and Sunday evening are common across industries.

emailers or information workers. In practice, organizations are fairly evenly distributed around this ratio and a lower tolerance does not dramatically change our viable set of organizations, but we are likely systematically excluding organizations of certain types, including retail firms.

Finally, we use distributions of individual behavior at a finer timescale than our eventual analysis to further identify organizations with active and consistent user behavior (Figure 4.1). As a proxy for how well we capture behavioral patterns, we select only organizations with daily average behavior that varies over the course of the day similarly to our reference data set (Chapter 2.2.2). This daily 'heartbeat' of activity is similar to other daily behavioral patterns observed in organizations, such as the physical proximity of acquaintances (Eagle and Pentland 2006) and activity in markets (Zaheer et al. 1999).

4.3.2 Network inference

From a record of interactions, such as communication, there is no unique network construction. Communication networks must be inferred from interaction data as a preprocessing step,

Table 4.1: Sizes of the communication networks. Total is taken as the sum over all 65 networks. Messages sent refers to the number of messages sent within the organization during the time period. Links and sender degree (i.e., number of contacts) are defined to be above the reciprocity threshold ($\tau \geq 1$); see Chapter 2.3 for more details.

	Mean	Minimum	Maximum	Std. Dev.	Total
Number of senders	21,247.2	4,446	218,986	30,903	1,381,065
Links between senders	58,4575	55,294	6,316,618	916,872	37,997,377
Unique messages sent	27,882,595	3,132,970	359,324,382	47,560,772	1,812,368,654
Average sender degree	26.9	10.6	53.4	8.1	–

the choices in which may change what patterns are detected (De Choudhury et al. 2010; Hofman et al. 2017; Kossinets and Watts 2006). In this exploratory analysis, we describe our results under one choice of network definition, and we show that our results are robust across a range of alternative choices of threshold and time window (Appendix B.3). Membership in the networks is restricted to within-organization email accounts that reflect active users. This boundary implies that the sizes of the networks are related to the sizes of the organizations. Each network represents within-organization aggregated over six months.

To construct each network, we define the nodes to be the active senders, and we annotate each person-person edge with weights. For each edge $(i.j)$ we assign the weight $\tau_{ij} = \tau_{ji}$ to be the geometric mean of the number of messages exchanged between each pair, weighted by the number of recipients on each message De Choudhury et al. 2010. For example, if i emails j directly five times, then this counts as $\frac{5}{1}$ messages, but if j emails i once directly and once again with four recipients total, that will count as weight $\frac{1}{1} + \frac{1}{4} = \frac{5}{4}$. Then the reciprocity between the two of them will be $\tau_{ij} = \tau_{ji} = \sqrt{5 \times \frac{5}{4}} = \frac{5}{2} = 2.5$.

We restrict the network to include only edges with weight $\tau_{ij} \geq 1$, and we validate our results on the network generated by other thresholds (see Appendix B.3). We then treat the network as undirected, having conditioned on each edge being reciprocated. We calculate network statistics only for senders in the giant connected component in the observed networks. This was on average 97.4% of identified users (maximum 99.1%, minimum 94.5%); this high percentage also serves as validation of the initial data cleaning process.

4.3.3 Network attributes

We calculate statistics over each network to characterize the structure. Averages are calculated within each network across all individuals (degree) or all pairs (shortest paths). We then compare these statistics across the population of firm networks.

Average degree. The average number of contacts one exchanges emails with. Degree is sometimes used as a proxy for social capital, potentially varying with status, power, or prestige; in communication networks, one's number of contacts may be limited by practical and cognitive constraints.

Average shortest path length. The shortest path (i.e., smallest number of hops) between two senders, averaged over all pairs of senders. In communication networks, this suggests how easily information could spread or innovations could be transferred across a large network, conditional on the links that already exist.

Clustering coefficient. A measure of transitivity of exchange, calculated by the number of closed triplets (triangles) by the number of connected triples. This can suggest collaboration and open communication, where there are multiple channels for sharing information, or as redundancy, where information is unnecessarily passed back and forth through additional ties. For example, consider two employees working together and with the same manager: a closed triangle of communication between all three parties could either represent collaboration or redundancy.

Centralization. A measure of inequality of access to information across the network, defined as the Gini coefficient of betweenness centrality. Betweenness centrality measures how many short paths will pass through an individual, so individuals with high betweenness will have access to diverse information and important roles in the network. If this is evenly distributed, then no individual has unique access to information, whereas if this is very unevenly distributed, then this suggests inequality in the distribution of information, and by proxy, power in the organization. Betweenness centrality tends to be unequally distributed in empirical social systems, appearing as right-skewed distributions (Newman 2010). The Gini coefficient measures how (un)evenly dis-

tributed this centrality is across the organization, from 0 (uniformly distributed) to 1 (power is concentrated in the top individual).

Small world quotient. We also include Walsh's small world quotient Q (Walsh 1999). This quotient is given by the ratio $Q = \frac{C/C_R}{L/L_R}$ for C, the clustering coefficient and L, the average shortest path length. These ratios are normalized by the expected values for an Erdős-Rényi random graph of the same size: $C_R = \frac{\langle k \rangle}{S}$ and $L_R = \log S$. This compares the degree to which a network's small world structure is different than random, and was also developed in Uzzi and Spiro 2005 to compare a population of empirical networks.

4.3.4 Organizational attributes

To define each organizational unit, we combine these entries to the highest level of parent organization. For example, Skype would be considered a part of Microsoft, and we would consider only Microsoft's organizational features.

Organizational context We use attributes of the firms to instrument different types of organizational contexts using the Dun & Bradstreet Hoover's database.

Firm demographics. We define the organization *age* to be the year of founding of the parent of the organization. We quantify the *size* of the organizations in two ways: the number of active senders, as defined by the communication networks (Sec. 4.3.2), and the number of employees listed in the D&B Hoover's database.

Dispersion. This counts the total number of distinct, active physical locations associated with the parent firm during 2015, according to the Hoover's database. This is a measure of the size of the organizational family tree, and is a proxy for geographic dispersion. This quantity will be related to differentiation in the firm and is our only measure that likely corresponds to the latent formal structure within these firms. Lacking precise measures of the formal hierarchies in each firm or any individual roles, this measure gives us insight into how the firm organizes itself, whether related to communication, function, industrial constraints, or management approach.

Industry. We use the primary SIC codes associated with these firms, according to the Hoover's

database. The first two digits are used to separate the six primary industry categories (Chapter 2.4.0.1).

Organizational performance We use publicly available financial data to create several measures of organizational performance, drawing from the D&B Hoover's database and MSN Money for firm attributes and corresponding industry averages for the end of 2015. We use four measures: log of revenue per employee; revenue quarterly growth rate (for the past year); return on assets (5 year average); and return on equity (5 year average). We also create a combined performance measure, by taking the average firm rank position of each measure compared to other firms in the dataset.

4.4 Results

Having constructed the informal networks for these firms, we can then measure the structure in these networks: specifically, to assess the underlying diversity in our data and how structural properties vary with firm attributes and outcomes. Recalling our original research questions Q1 and Q2, using the network construction and network attributes described in Section 4.3, we explore the relationships between the communication network structure and firm attributes. For the relationships between organizational attributes, performance and network structure, we compare different models of best fit and use AIC for model selection.

We find a considerable amount of heterogeneity among these networks. While some variability can be explained by the size of the firms (Section 4.4.1), we find within-type diversity of network structure typically exceeds any cross-type diversity. This large range of firm-level differences are not correlated with organizational characteristics (Section 4.4.2) or performance (Section 4.4.3). Finally, these results are robust: we find qualitatively similar results, including wide heterogeneity among networks, even within homogeneous industry sectors (Appendix B.1.4) and over different choices of network construction (Appendix B.1.3, B.3).

4.4.1 Firm size and informal network structure

A key feature of our data set is that these firms vary in size, over almost two orders of magnitude (Figure 4.2). Although this suggests generalizability of our results across large firms, measures of empirical networks are expected to vary with size, in a similar way to how properties of random networks scale (Newman 2003). So we must first tease apart what variability among network structure is a function of their size, rather than something inherent to the firms themselves.

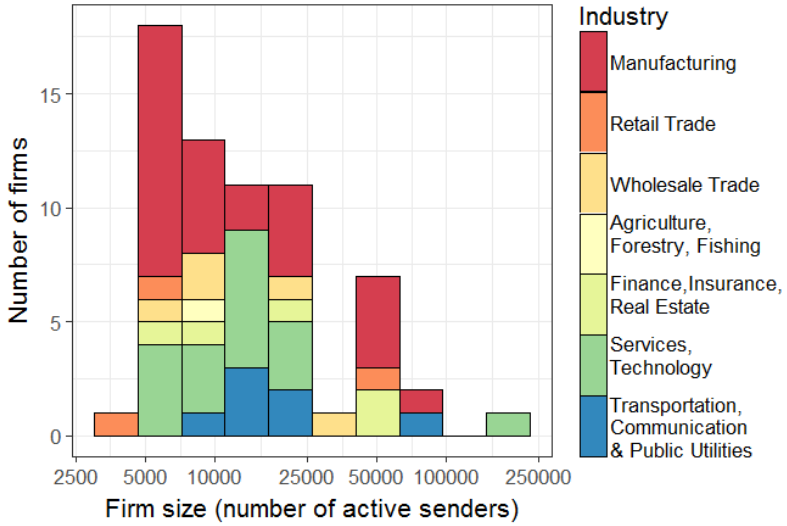

Figure 4.2: **Histogram of firm sizes.** Firm size is given by the number of active senders, and industry by top levels of SIC code.

The reasons to do this are twofold: first, understanding how firms vary by size is, on its own, interesting, and has been a foundational topic of organizational theory. (Dobrev and Carroll 2003 argue that "size is perhaps the most powerful explanatory organizational covariate in strategic analysis," a not uncommon sentiment.) The role of size is necessarily complicated, as firm size is endogenous to performance (Ahuja et al. 2012; March and Sutton 1997), as both an outcome and antecedent of productivity; unpacking causal effects of size on performance is already suspect. But, our second goal is direct: to the extent that our chosen measures vary with size, our interpretation of variability among firm networks mis-attributes known patterns to individual differences.

Average degree does not vary with size. One of the most basic network measures,

degree, considers the number of contacts one has in an organization. Empirically, individuals' degree may vary for functional reasons across an organization (an administrator may communicate with more than a specialized engineer), personality, and social constraints and mechanisms for information transfer. Bandwidth on total communication and structural patterns of an individual's network suggest tradeoffs for the size of their network (Aral and Van Alstyne 2011); in email organizational networks, the number of ties has been found to grow, causing networks to "densify" (Leskovec et al. 2007), corresponding to increasing average degree. On the other hand, degree may still be limited by cognitive constraints (Hill and Dunbar 2003), corresponding to constant average degree; Hill and Dunbar 2003 suggest that degree will vary with relationship type, whether strong ties or acquaintances, but will have low variance conditional on relationship type. Other conceptions of firm growth disagree as well: average degree in an organization might increase, due to greater availability of potential contacts (Krackhardt 1994a) or flatter managerial structures with wide span (Simon 1962), or decrease, due to increases in hierarchy and specialization (Blau and Schoenherr 1971).

Figure 4.3 shows that average degree varies widely across all firms (grand average is 26.9 reciprocated contacts, with standard deviation of 8.1 contacts (Table 4.1), but it does not vary with firm size—a conclusion that also holds for median degree[2] (Appendix B.1, Figure B.1.1) and is robust to network definition (Appendix B.3). In other words, this finding does not support any of the hypotheses suggested: Leskovec et al. 2007 would suggest that average degree increases, and Hill and Dunbar 2003 would suggest that cognitive constraints imply constant average degree, with little variability.

The average shortest path length has been of significant focus in the organizational theory and management literature (Uzzi et al. 2007). Network theory has also well described the scaling properties of average shortest path length in many random network models. In many of these models, average shortest path increases proportionally to $\log S / \log\langle k \rangle$ for the size of the network S

[2] The average median degree is higher, 34.7, suggesting that about half of typical senders in each organization have fully reciprocated email exchanges with about 35 people (Chapter 2.4; Figure B.1.1).

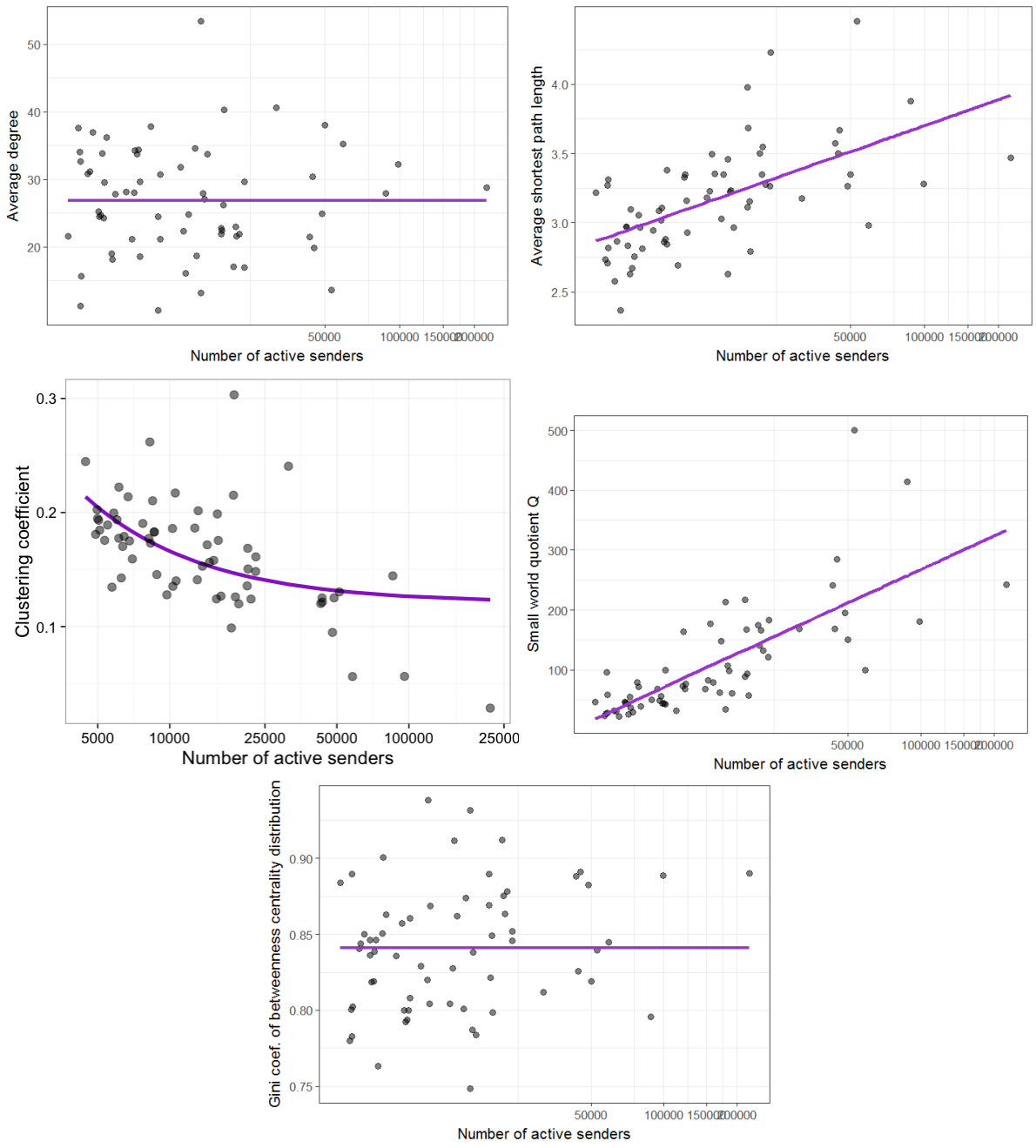

Figure 4.3: **Informal social networks exhibit wide heterogeneity, only some of which is explained by size.** We find three important results from these comparisons. (1) Average degree does not vary with size, which does not support a number of hypotheses from the literature. Conditional on degree, average shortest path varies in an expected way. (2) Clustering coefficient decreases as $\log S/S$, different than what has been modeled in the literature. The small world quotient varies with size, in an expected way conditional on average shortest path length and clustering coefficient. (3) Finally, centralization does not increase with size.

and average degree $\langle k \rangle$. This a pattern which has also been observed in empirical networks. This scaling pattern emerges naturally in systems that are well-approximated by an underlying hierarchy, which would have height proportional to $\log S$ given a fixed average degree. Many complex systems naturally have underlying hierarchical organization (Simon 1962), but this is explicitly a feature in large firms through the formal network of roles. Then not surprising, it is important nonetheless that we recover that as a function of network size S, average shortest path length is best fit by $\log S$; that is, we recover this logarithmic scaling in the firm networks (Figure 4.3).[3] Intuitively, we also find that additional variation in the average shortest path length is explained by average degree (Table B.2.1): in networks of equivalent size, additional ties would create more, potentially shorter, paths in the network. Then higher average degree suggests a lower average shortest path length, and we observe this in practice (Pearson's correlation -0.76). As average degree does not fluctuate with the size of the network, we do not observe any additional confounding scaling effects due to degree.

Clustering coefficient decreases like $\log S/S$. Reflecting near ubiquitous processes of homophily and transitivity, it is well known that clustering in social networks is generally much higher than what simple random graph models would predict (Watts and Strogatz 1998). We therefore expect that networks derived from communication networks in firms, will similarly reflect relatively high clustering: reflecting homophily, two members of the same team, unit or division are more likely to communicate on average than two members of different teams, units, etc.; and reflecting transitivity, team members who communicate with the same manager are also more likely to communicate with each other. Less clear is how clustering should be expected to vary with increasing size of a firm. Although generative models of graphs have not generally focused on the relationship between clustering and size, they have typically implied one of a few simple scaling relations. At one extreme, random graph models imply that the clustering coefficient scales as $O(\langle k \rangle / S)$, falling quickly with the size of the graph: in very simple models, such as the Erdős-

[3] In addition, we find that the diameter of the networks, defined as the longest shortest path in the network, also scales logarithmically with size (Appendix B.1, Figure B.1.3).

Rényi random graph, there is no basis for the common social mechanism that "the friend of your friend is your friend." At the other extreme, a number of "small-world" models (Holme and Kim 2002; Watts and Strogatz 1998) have shown that clustering can be kept nearly constant even as network size increases. Finally, in between these extremes, models of networks that assume some underlying hierarchical structure imply that an individual's transitivity will scale like $1/k_i$, although the scaling is sensitive to definition of global clustering coefficient (Ravasz and Barabási 2003; cf. discussion in Newman 2003).

Encouragingly, Fig. 4.3(C) shows that overall clustering coefficients are comparable to numerous previous studies of large-scale social networks (Kossinets and Watts 2006; Leskovec and Horvitz 2008; Ugander et al. 2011) which find clustering coefficients in the range $0.1 \leq C \leq 0.15$. With respect to scaling, the quantity and scale of our data is insufficient to reject a number of similar models; however, by observing that $C \times N$ scales like $O(\log S)$, we can assert that $C = O(\log S/S)$ (further details in Appendix B.1 and Fig. B.1.4). We do not find evidence to support constant or $1/S$ scaling, as suggested by the literature, but this may suggest agreement with Klemm and Eguiluz 2002a. As with degree, this result does not correspond with any of the dominant models of graphs; however it is not without precedent. For example, Klemm and Eguiluz 2002b describes a model where clustering falls quickly in small graphs, but for large graphs (size of over 100 nodes), clustering scales as $(\ln S)^2/S$; the authors expand on this to suggest two models: one with highly clustered, but constant, scale-free networks, or "random" scale-free networks with $(\ln S)^2/S$ clustering (Klemm and Eguiluz 2002a). Dunne et al. 2002, on the other hand, found empirically across a population of diverse networks that clustering coefficient does decrease proportionally to $1/S$.

Small-world quotient scales like $\log S$ The small world quotient, defined as $Q = \frac{C/C_R}{L/L_R}$, characterizes how far a network is from random. Interestingly, even though the measure ostensibly controls for size by comparing to a random network of the same size by definition, the quotient has been empirically observed to vary with size (Baum et al. 2003). As a result Gulati et al. 2012 specifically attempts to control for observed variation in the measure with respect to size, however they apply the same scaling properties as C_R and L_R, resulting in a measure that reduces back

to C/L. Adjusting for size would be a necessary step if C_R and L_R systematically depart from empirical scaling, i.e., if the scaling of random graphs is different than the empirically observed pattern; however, this pattern has been unknown. We find that the small world quotient scales logarithmically with the size of the network. This follows as a consequence of our observations about C and L. Recalling our earlier result, the dynamics of the denominator, $L/L_R = L/\log S$, appear to not contribute meaningfully here to the scaling: Figure 4.3 showed that L varied regularly around $\log S$ with average degree, which was unrelated to size. Then although the denominator captures the distance to what would be expected by a random network of the same size, the empirical networks have values both above and below this number, but within a small range of values. This suggests that the dynamics of the small world quotient are dominated by the effects of the clustering coefficient, in the numerator. The ratio $C/C_R = \frac{C}{\langle k \rangle / S}$ varies logarithmically with size, similarly to Q (Figs. B.1.4 and B.1.5), Appendix B.1). Recalling that the average degree, $\langle k \rangle$, was constant with size, this suggests that we are here primarily observing the dynamics due to the clustering coefficient: $O(S \times \frac{\log S}{S}) = O(\log S)$. While this does not suggest any immediate interpretation, this offers a way to interpret the role of size on analysis: past studies, including Baum et al. 2003; Davis et al. 2003 and Uzzi and Spiro 2005 use a size-independent threshold to determine whether or not a network has the small world property, which will be inappropriate for networks of significantly different sizes.

Centralization does not vary with firm size. We use the distribution of betweenness centrality as a proxy for how access to information is distributed across a company. We measure how unevenly distributed this centrality is using the Gini coefficient, where high values (near 1) suggest that very few individuals hold very highly central positions, and low values (near 0) suggest that centrality is uniformly distributed across the organization. We find that while inequality is typically high (mean: 0.84, standard deviation: 0.04), we find that there in no relationship between the distribution of inequality in firm networks to the size of the networks (Figure 4.3). We also observed no relationship to size in the inequality of the degree distribution (Appendix B.1, Figure B.1.2). The degree distributions are also less unevenly distributed (mean: 0.55, standard

	Network statistic				
	$\langle k \rangle$	L	C	Q	Gini
Size S	0.0	0.36	0.41	0.57	0.0
Industry	0.13	0.12	0.11	0.09	0.06
Age	0.0	0.0	0.0	0.0	0.0
Dispersion	0.0	0.18	0.09	0.15	0.13

Table 4.2: R^2 **for best-fitting models of network statistics.** R^2 captures the variance explained by the model of best fit for each relationship. For industry, this represents the variance explained by industry category, and these quantities are not significantly different than zero: see Appendix B.1.3 for more details.

deviation: 0.07). Degree may be less extremely distributed, as the number of contacts is still constrained by human activity: very few senders with extremely high degree, and many senders with very low degree, may not be representative of active email users in an organization over a long period of time, where we expect roles to be more evenly distributed across the organization and strength of reciprocity to limit the maximum degree feasible. High inequality that is invariant to size is compelling: as organizations get larger, we might expect that communication becomes more unequal. Instead, either as a consequence of the underlying hierarchy or despite it, firms do not appear to become more unequal with increased size. However, this still suggests that all firms have highly skewed, unequal distributions of power.

4.4.2 Firm context and informal network structure

We now consider what organizational attributes, beyond size, are related to network structure. Our firms vary widely in age, industry and geographic dispersion. We find that the heterogeneity of informal network structure in firms typically exceeds any heterogeneity explained by their characteristics. These results are summarized in Table 4.2.

Network heterogeneity is not explained by industry. While large public firms may have some common structures, representing hierarchy, administration and specialization, the needs of these firms may vary by industry. Manufacturing firms that produce physical products seem likely to have different online footprints than software companies or media companies, for example.

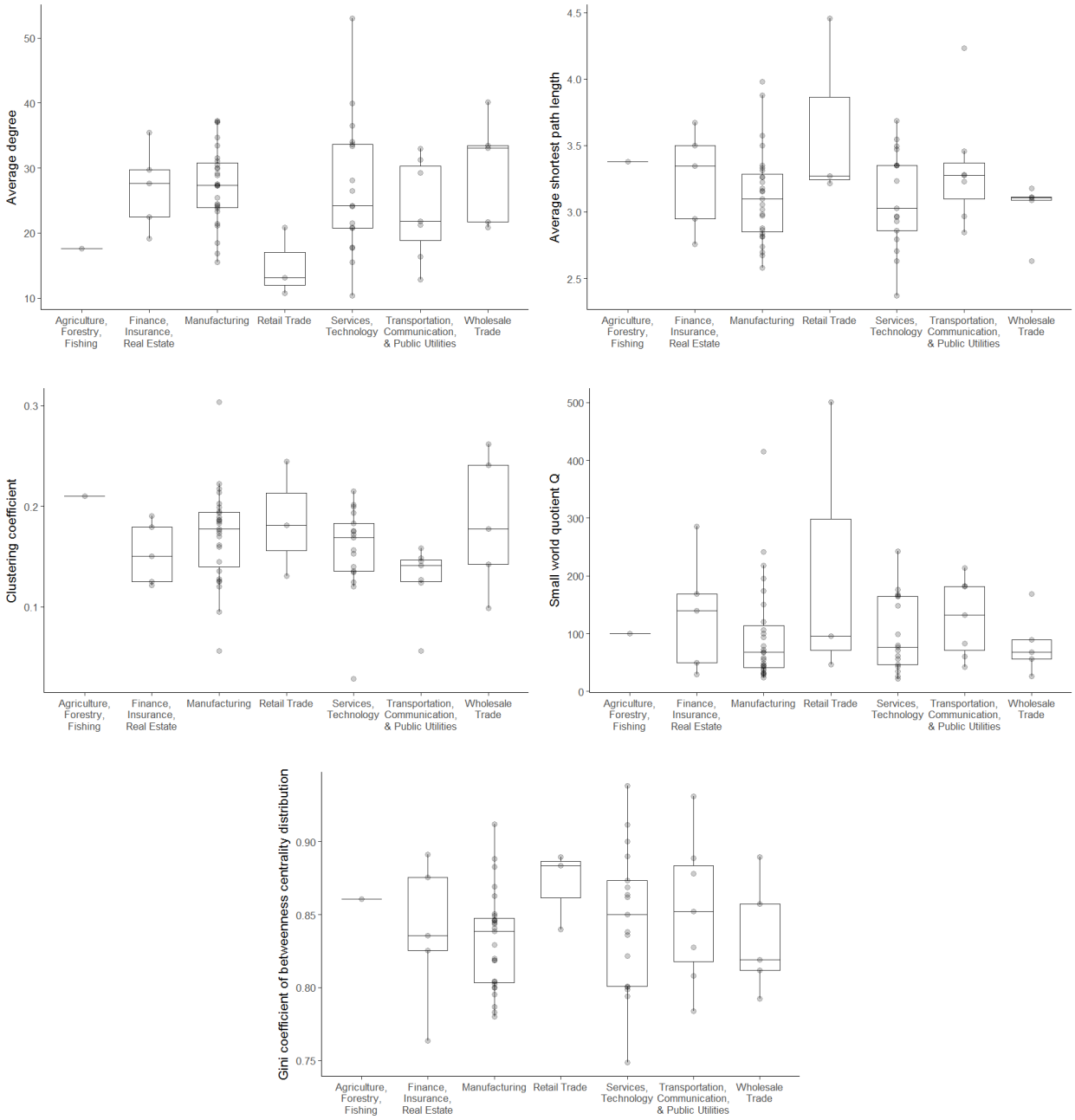

Figure 4.4: **Informal social network features are unrelated to industry.** Across all measures, we find that within-category variance exceeds between-category variance.

DiMaggio and Powell 1983 famously argue that firms with similar contacts or of similar types ought to become more similar over time by virtue of imitation and diffusion. Instead, we find that within-industry heterogeneity exceeds almost all between-industry heterogeneity. Across all five network measures, the scale of this heterogeneity is striking. In Figure 4.4, we separate out firms by each top-level SIC industry classification. For all network measures, there is significantly more within-industry variation than between. The scale and robustness of this diversity within industry suggests evidence against the "startling homogeneity" of firms (DiMaggio and Powell 1983).

Wide variation within industry is not unprecedented: Foster et al. 2008, for example, found broad differences even within similar product manufacturing, with productivity differences related to entry and exit of an industry. Larger public firms, as we observe here, may be more diversified, capturing a range of efficient (or inefficient) structures. But, here we still might be able to capture cultural differences around communication that vary by industry. When we condition on size, we do find moderate evidence for deviation in average shortest path length and small world quotient by the retail trade sector (Appendix B.1.3). However, we have insufficient data to validate this result, and we then leave this hypothesis to future confirmatory research, which will have to tease apart whether this is an effect of our measurement tool—if email captures behavior differently, potentially less well, in retail settings—or whether these informal networks actually vary meaningfully as a function of organizational context.

Network heterogeneity is not explained by firm age. Age is measured from the year of founding of the parent organization, subsuming any mergers and acquisitions. Older firms can suffer from organizational inertia (Hannan and Freeman 1977), and their informal networks may be subject to decades of path dependence. This may lead to the emergence of more concentrated power and increased centralization; Krackhardt 1994a argued that Michels's Iron Law of Oligarchy applied at least as well to firms. DiMaggio and Powell 1983 famously suggested that organizational forms should become more similar over time: it is possible then that we are observing firms at different stages in this process. However, we lack the statistical power to claim that heterogeneity is decreasing with age. We also find that if the small-worldness of the network is evolving (cf. Gulati

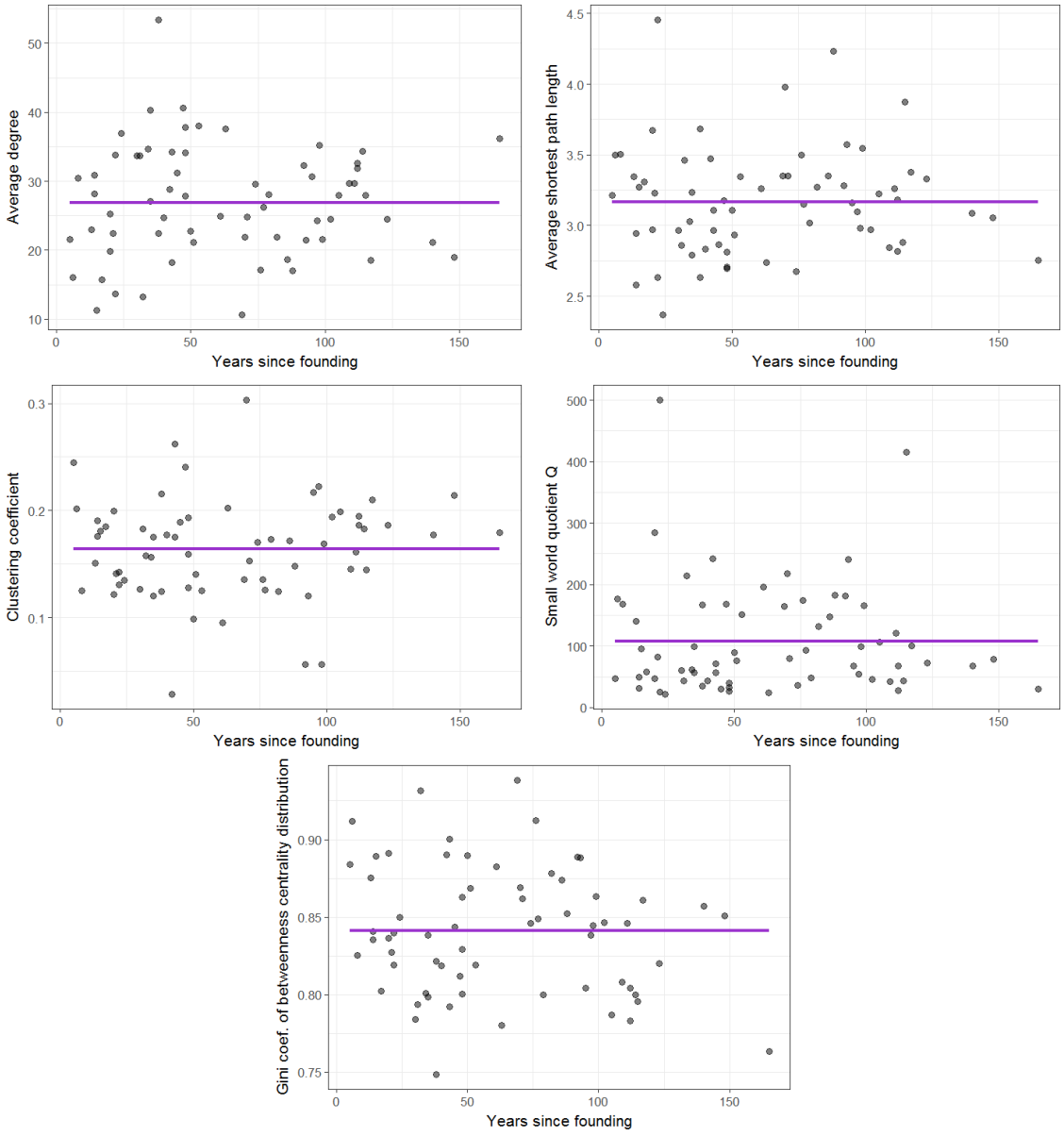

Figure 4.5: **Informal social network features are unrelated to firm age.** However, organizational network properties are very diverse across all ages.

et al. 2012), there is not evidence that this structure varies with the age of the firm. Once again, we find wide heterogeneity across all network measures (Figure 4.5), but we find no relationship between firm age and network properties. This is robust to network definition, that is, even on networks constructed from stronger pairwise relationships (Appendix B.3).

Degree does not increase with dispersion, but centralization does.

Average degree shows no relationship to dispersion (Figure 4.6)). This is suggestive that the constraint on total communication is somehow held constant regardless of organization size or arrangement. Clustering coefficient decreases logarithmically with firm dispersion, and the average shortest path length increases logarithmically and small world quotient $Q \propto \log\log d$. However, as predicted by theory (Blau 1970), dispersion increases at a declining rate with firm size (Figure B.1.7); this suggests that the clustering and average shortest path length results are an artifact of firm size (Figure B.1.8, Appendix B.1.2). The small world quotient does appear to vary meaningfully with dispersion: while clustering coefficient is not related to the amount of dispersion per person, it does appear to be more similar to random in firms that have more geographic units per person (Figure B.1.8). Centralization, on the other hand, grows logarithmically with firm dispersion, and appears to vary independently of size. This suggests that there is clearly an impact on how information *flows* through the network: with larger and more dispersed firms, individuals still can reach each other with relatively few hops but require longer paths through a less cohesive network. With increased centralization, these paths must also be less evenly distributed, creating more unequally distributed roles. This suggests that while organizations have mechanisms to keep individuals tied to each other across the organization, fewer individuals span these boundaries.

4.4.3 Informal network structure and firm performance

Informal network structure reflects the capabilities, adaptability and performance of the firm. Hansen 1999 and Argote and Ingram 2000, for instance, describe how ties between project teams and among employees mediate knowledge search and transfer, with implications for project success and innovation, translating into firm-level performance outcomes. Uzzi and Spiro 2005

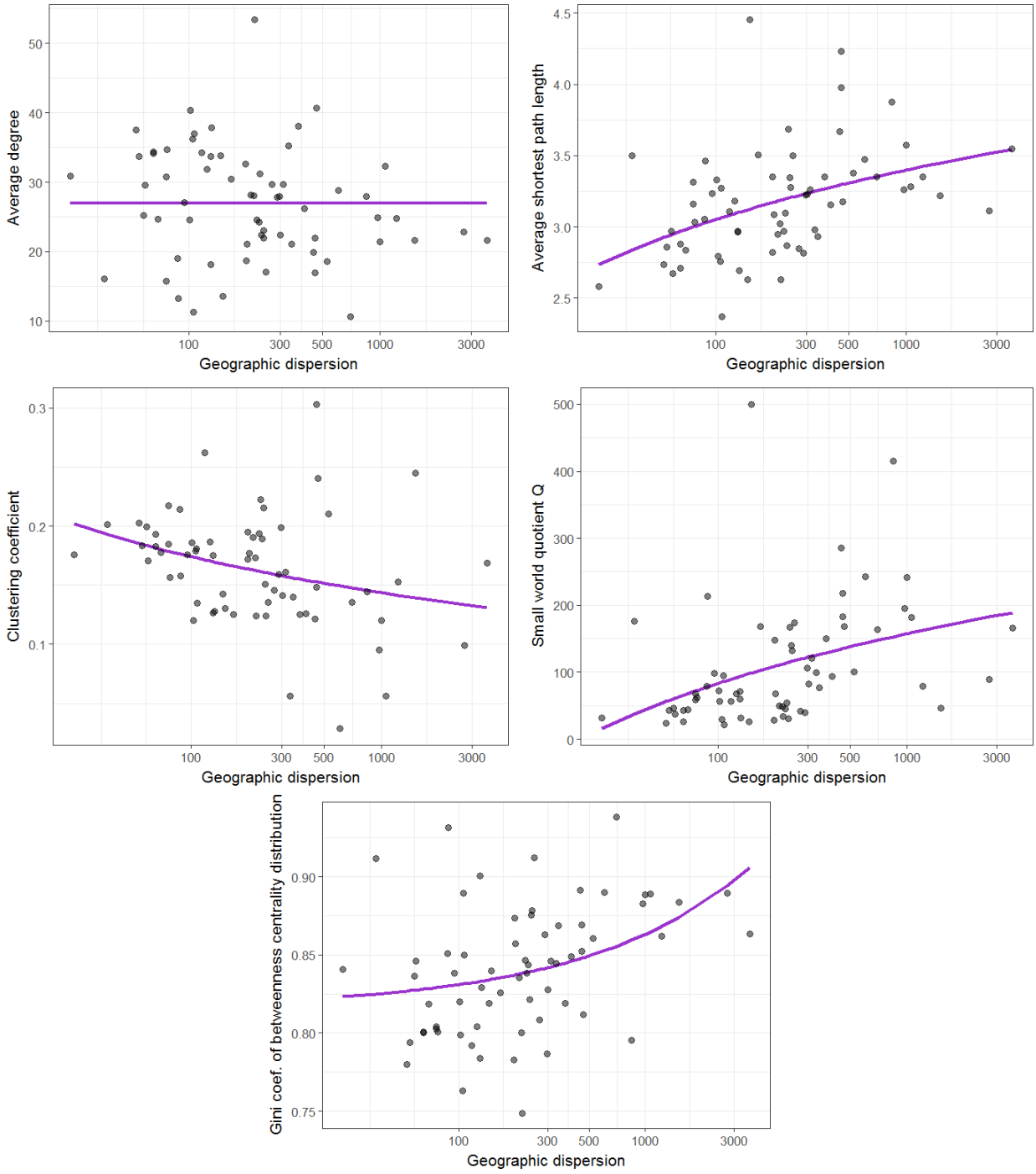

Figure 4.6: **Centralization increases with firm dispersion, but average degree does not.** Bottom panel, centralization increases as dispersion $d^{0.5}$(note log scale on the x-axis). These other measures, L, C and Q vary as an artifact of network size: see Figures B.1.6 and B.1.8.

found that variation in the small world properties among networks of artists were related to success and creativity. Leana and Van Buren 1999 argue that relationships within an organization impact organization-level outcomes, such as flexibility and productivity, by aligning collective goals and increasing trust; Lahiri 2010 and Alcácer and Zhao 2012 find that the structure of internal networks between teams increases innovativeness despite geographic pressures.

Dalton et al. 1980 also argues that while there have been a range of attempts to measure the relationship between structure and performance, the functional forms or even the sign of relationships is not clear from the literature. They find, e.g., negative and null relationships between centralization and performance, although those lacked rigorous quantitative measures of firm productivity and output (such as financial data), and also no systematic evidence for a relationship between size and firm performance (this established also by meta-analysis in Gooding and Wagner 1985). While the forms of these relationships are not clear, the theme that there is a relationship between structure and performance is consistent. Kleinbaum and Stuart 2014 present a cogent modern argument that there is a strong theoretical basis for intraorganizational networks to be a determinant of firm performance. Of course, "it ultimately will be necessary but challenging to assemble network data from many firms to test such a theory" (Kleinbaum and Stuart 2014): despite the theoretical prior in the literature, the empirical relationship between variation in informal network structure and firm outcomes is unknown.

We explore several potential types of relationships between network structure and organizational productivity. We use a range of outcome variables describing firm performance: income per employee, return on equity, return on assets, and revenue quarterly growth rate, and a combined rank measure based on the combination of these variables. We first look for evidence of pairwise relationships between network structure and performance. We then consider nested regression models to model more complex relationships between multiple types of network structure and performance. Finally, we create a classification task to predict whether firms are high or low performers and use a flexible machine learning method, random forests, to detect any relationship between network

structure and performance.[4] We find that performance is not predictable or related to observable differences in informal network structure.

In the first task, we look for simple correlations between network structure and productivity. To illustrate what types of patterns we find, we show average degree, average shortest path length, clustering and centralization—common variables of interest in the literature, e.g., Ahuja et al. 2012; Granovetter 2005; Reagans and Zuckerman 2001; Rice 1994—in Figure 4.7. Across different performance outcomes (see Section B.2.1 for additional details on variables), and confirmed across other network definitions (Appendix B.3), we find a stunning lack of signal across effectively every variable and outcome. For all pairwise relationships, we find $R^2 = 0$ for the best-fitting model across all pairs of informal social network structure ($\langle k \rangle$, L, C, Q, and centralization) and the performance outcomes (Income per Employee, Return on Assets, Return on Equity, revenue quarterly growth rate, and the combined ranking).

We next look to a regression task to predict outcomes. We incorporate dummy variables to condition on performance by industry sector. We exclude the single firm from the agriculture, mining and forestry industry sector. We then compare a series of nested linear regression models, where in addition to the industry dummy variables, we progressively incorporate the number of employees, adoption (ratio of senders to employees), and the network measures ($\langle k \rangle$, L, C, Q, and centralization) into the model. Appendix B.2.1 shows the explicit model constructions and regression coefficients. We find that performance is not predictable or related to observable differences in informal network structure. The lack of meaningful relationships was also robust to different network definitions (Appendix B.2.1).

Finally, we carried out a number of other prediction tasks: using the network structural measures as features, we used decision trees and random forests across different tasks of varying difficulty. These tasks were predicting whether a firm is above or below median industry performance; or, presumably easier, predicting whether or not an organization was in the top 10% or the bottom 10% for performance compared to others in the same industry. In preliminary work, these

[4] Results using random forests will be included in the future version of this paper submitted for publication.

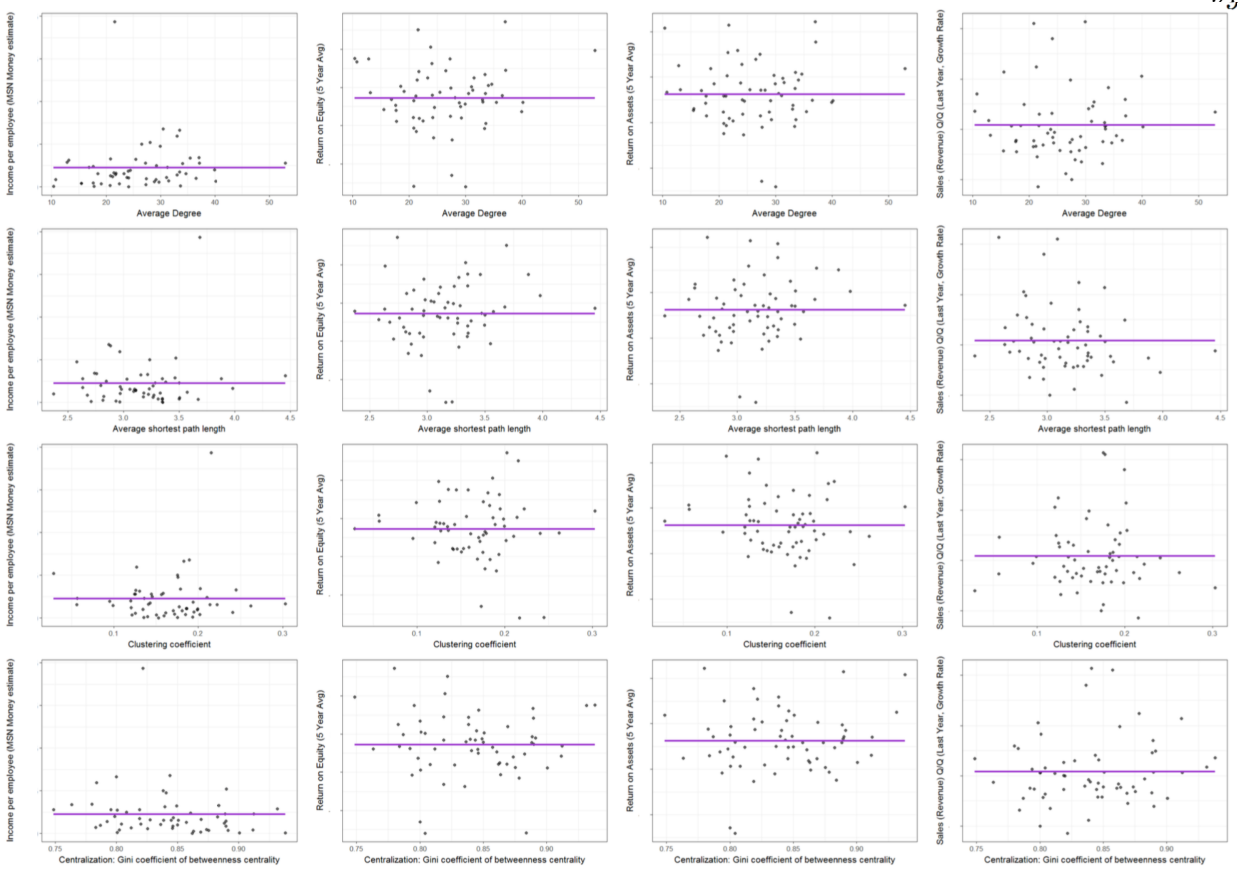

Figure 4.7: **Productivity measures are unrelated to network measures in informal social networks.** We compare network measures pairwise to different outcome variables. From top to bottom, each row shows average degree, average shortest path length, clustering coefficient, and centralization against the performance variables. From left to right, the performance variables are Income per Employee, Return on Equity, Return on Assets, and revenue quarterly growth rate. We find no statistically meaningful relationship between any of these measures, nor the measures not shown.

exercises, with many network features or very few curated features, revealed no signal, with predictions no better than random (AUC not statistically significantly different than 0.5 for classification tasks with balanced classes); we leave the complete results to future work (see Appendix B.2.2).

Overall, we find that financial performance in firms is not observably related to informal network structure. This holds true across industries, prediction tasks, and construction of the data set. While this setting is strictly exploratory—to discover potential relationships between network structure and performance—had we found some meaningful correlation, it would still most likely

be spurious. We discuss some of the potential avenues for future research, as well as the limitations of this study, in the closing section.

4.5 Discussion

Groups of people working together, from small teams to companies and governments of hundreds of thousands, accomplish complex tasks using a diversity of communication patterns. The extent of this diversity has been previously difficult to quantify, as the scale of this comparative data-driven perspective—particularly of structural communication patterns in firms (Kilduff and Brass 2010; Kleinbaum and Stuart 2014)—has been previously unavailable to the academic community. We developed a large, high-resolution data set to investigate communication patterns across a population of highly comparable large firms, ranging from a few thousand to a few hundred thousand employees, covering 1.4 million employees total from six industries.

We find that there is a wide amount of heterogeneity in the structure of informal networks in firms. Furthermore, we find that this heterogeneity is both large and largely unexplained by organizational context, despite what has been previously suggested in the organizational theory literature. These findings are robust across network definitions and timescales, as well as among homogeneous organizations. The scale and robustness of this heterogeneity suggests that patterns hypothesized from case studies may be easily subject to over-fitting: one might be able to equally validate competing hypotheses given only a handful of examples (cf. Davis 2010).

The salient association between the size of the organizational network and network structure is compelling. Krackhardt 1994a described how organizations would be naturally constrained by communication networks from growth: firm growth would lead to increases in degree that are potentially unsustainable. We find that network structure does vary with size, but not with productivity; this is seemingly in alignment with results from organizational theory that size does *not* matter for organizational function and productivity (Gooding and Wagner 1985). (Of course, disagreement remains: cf. Dobrev and Carroll 2003, "size is perhaps the most powerful explanatory organizational covariate in strategic analysis.")

We find that the average number of contacts that people have is constant with respect to the organization size, and that this is robust to network definition. This result, and the robustness of this result, appeals to a number of different academic communities. The study of the social brain suggests we might find constant average degree, although without wide heterogeneity (Hill and Dunbar 2003); degree has been suggested to be related to innovation in organizations (Reagans and Zuckerman 2001, e.g.); and the study of graph evolution and scaling has centered on models for which degree increases with network size (Leskovec et al. 2007).

Through organizational inertia or the iron law of oligarchy, we would expect to see centralization increase with the size and age of the firm. Instead we find that centralization of the network is unrelated to firm size and age. However, we find that centralization is positively related to the dispersion of the firm. This is not obvious: Monge and Contractor 2003, for example, describe how online communication across geographically disperse organizations "can lead to greater centralization in some organizations, but also undermines centralization in others," Furthermore, this could potentially relate to outcomes: Jansen et al. 2006 found that less centralized formal networks were associated with increased innovation. Furthermore, as centralization is not a function of size, it is not a managerial outcome that large firms are associated with information bottlenecks, but instead a correlate of the physical dispersion of senders. Geographic dispersion is likely related to shifts in the communication network: similar to in-person, off-line contact, the likelihood of email contact has been shown to drop off with physical distance (Krackhardt 1994a; Monge and Contractor 2003), and Adamic and Adar 2005 find that the probability of connection drops as inverse the distance between two people. Furthermore, centralization may be associated with geographic dispersion as an emergent organizational outcome or by managerial design. While greater geographic dispersion does not change the number of contacts an individual has, on average, our results suggest that the barrier of separate physical units is crossed by fewer individuals.

Despite a strong theoretical prior from the literature, we found no relationship between this heterogeneity and firm performance, and this is true across a range of productivity measures. While predicting firm performance is objectively hard, we use a range of tasks that vary in difficulty, task

and instrument: if there were a relationship, we ought to find it. While it is infeasible to prove that there could be no relationship, these tests suggest strongly that it is very unlikely that there is no relationship between network structure and performance. Data mining even more aggressively could plausibly discover a correlation, as a matter of uncovering statistical artifacts, rather than a meaningful association.

Limitations This empirical perspective offers a new lens across a population of complete intraorganizational networks; however, we do not yet fully understand the empirical and theoretical limitations in this new territory. A number of these open challenges are empirical, and left to future work. For theoretical challenges, we are constrained by what data is available and observable.

The constraints used to construct the data set potentially introduce several sources of bias. First, we are necessarily restricted to firms that are Microsoft customers. Beyond that, we have introduced a range of constraints as a best effort to find firms of similar email usage and data quality. Error has been necessarily introduced by merging a noisy database of firm attribute information (Dun & Bradstreet) with a noisy measurement tool (email, where senders need not be full-time employees). We potentially introduce endogeneity by restricting our data set to publicly traded firms: requiring that firms be large and publicly traded already implies that these are firms that have historically been successful, and we are then conditioning on the outcome (Davis 2015a).

One major source of bias for our analysis is due to the communication network data. This has a number of potential sources of bias. First, the patterns of relationships reflected by email communication data are necessarily noisy observations of true social relationships. Grippa et al. 2006 have shown that in a small organization, email can fail to represent interactions among closely co-located individuals. However, Grippa et al. 2006 also argue that email *underestimates* the roles of communication gatekeepers, compared to other types of network construction, which may suggest that our observed high levels of centralization may even be an underestimate. On the other hand, Quintane and Kleinbaum 2011 suggest that self-reported relationships will overstate high status relationships and therefore misrepresent information flow, compared to email networks.

Furthermore, different choices of network construction from email communication data will

lead to differences in network measures. De Choudhury et al. 2010 found that networks defined by 5–10 reciprocated emails per year were most predictive of future interaction; for our six month data set, this would correspond to half as many, but it may be that in the full organizational network setting, other levels of relationship strengths—weaker, stronger, or both—may better represent channels of information flow and communication structure relevant to organizational function. Different network differences could result in different inferences drawn from our data (De Choudhury et al. 2010; Hofman et al. 2017); this motivated our robustness checks across networks derived using different relationship strengths.

Construct validity is difficult to assess in studying organizations (Davis 2010). The concept of firm size is poorly defined—Kimberly 1976's criticism is still relevant today—and this is salient here, where the number of active company email senders is different than the number of full-time employees. (This is both a boundary specification problem and a construct validity problem.) Many studies that have looked at communication networks in organizations have been restricted to only a subset of the firm, for example, only the R&D branch (e.g., Rice 1994). One attempt to compare networks across firms only analyzed the top three levels of the firm (Nelson 2001). The infamous and well-studied Enron email network data set covers 158 out of about 20,000 total employees, primarily including senior management, the executive level and the board (Diesner et al. 2005). It is unclear what the correct boundaries are to understand an organization.

Finally, we are limited by what data is observable. We lack the content of the messages exchanged, and so we can not observe, for example, the qualities of relationships between senders or the types of information being shared. We cannot measure the content or sentiment attached to the email data, nor directly observe the cultural differences around email usage in these firms. We do not know the gender, generation, or job function of the senders in our data set. Furthermore, we cannot directly observe whether or not relationships are within or across teams. This limits what we understand about dispersion, in particular. For dispersion, we do not observe the true formal network, nor do we know how the firm distributes roles across geographically distinct locations. Dispersion within division, for example, may be more relevant to outcomes than across the firm:

Lahiri 2010 and Gibson and Gibbs 2006 examine the impact of geographic dispersion on innovation in R&D teams. Without knowing the formal network of roles, we cannot measure how well aligned, or similar, the informal network is to the formal network: it is possible that this network responsiveness is most related to firm context and outcomes (Kleinbaum and Stuart 2014). Soda and Zaheer 2012, for example, found that the alignment of informal and formal networks lead to increased performance, where this alignment creates a trade-off between reduced coordination and increased access to information.

Future directions

Organizational dynamics, reflected in their network structures, are important but still not well studied (Ahuja et al. 2012). This requires meaningful theory, or motivating practical challenges, to connect dynamics to organizational understanding. (Simply finding a relationship would be insufficient in this context: the existence of a correlation between some high resolution financial measure and some dynamic network measure would say far more about our many researcher degrees of freedom than anything about communication and structure in firms.) Understanding how information flows through organizations is crucial to understand organizational function, learning, and efficiency. This will be a problem that requires understanding what information is being transmitted and how—observing content, as well as roles—and may require looking more broadly across organizational types.

Global communication network structure has previously been found to be largely stable, but exhibits rapid dynamics at the local level (Kossinets and Watts 2006). The rate at which we observe a network compared to its underlying dynamics will then change the inferences we make from it (Clauset and Eagle 2007; De Choudhury et al. 2010; Hofman et al. 2017). Understanding how sampling effects the observation of communication networks would have implications for understanding how to do robust inference in dynamic networks. Empirical populations of organizational networks would then help us tease apart network change, at the individual and organizational level, to better understand when shifts in network structure are meaningful. This has implications for understanding individual networks (e.g., Burt 2002, 2004), consequences of organizational

change and network responsiveness (Kleinbaum 2017; Kleinbaum and Stuart 2014), and social and organizational reactions to shocks (Romero et al. 2016; Srivastava 2015).

The wide heterogeneity of observed organizational forms, even among firms within the same specialized subfield, suggests that there is a wide range of communication patterns that through which people can achieve complex, large-scale tasks. Given the interplay between social and communication structures at the individual, team, and inter-team level, mixed methods approaches may better explain how to interpret organizational level communication (Ibarra et al. 2005).

Conclusion We introduced a novel data set, comprising 1.4 million senders and 1.8 billion email exchanges from a diverse population of 65 large, publicly traded firms. We emphasize that the comparability of these networks is a key component of the construction of this data set: lacking any prior empirical baseline, and an unknown variability across informal networks, we made a range of restrictions to explore and assess the natural heterogeneity across these firms. While our data set is restricted to comparable firms with similar and consistent email adoption and usage, we discover a rich diversity of network structure. Variability in these informal networks is only partly described by network size (a function of the size of the firm), and in large part unrelated to the context or performance of the organization. And yet, organizations of all sizes, these included, are effective at accomplishing complex tasks. Understanding when, and how, informal networks are related to firms, given this broad diversity, then suggests a challenge for organization theory. Exploratory research of this nature can push us towards meaningful confirmatory research to understand heterogeneity in organizational structure (Davis 2015a; Hofman et al. 2017), however this may require revisiting the questions appropriate and testable for understanding organizations, including what aspects of performance are relevant in different contexts. We look forward to the questions that will emerge in this intersection.

Acknowledgements This work was done in collaboration with Duncan Watts, and was supported by NSF Graduate Research Fellowship award no. DGE 1144083 and Microsoft Research. The authors thank Aaron Clauset, Jake Hofman, and Amit Sharma for useful discussions and feedback.

Chapter 5

Empirical network construction: computational perspectives on weak ties, stability, and densification

While social relationships are not directly observable, we can observe and measure interactions between people. Empirically observed interactions then present a mode with which to infer the existence and strength of relationships. To define a network from relational data—such as the patterns of emails sent, phone calls exchanged, collaborations, or trades between individuals—we explicitly or implicitly select some instrument that detects these relationships. The resulting network encodes these ties into networks, a low-dimensional representation of the social world that is revealing nonetheless. However, empirical tests of social theories rely on having measured these networks in some reasonable way, and evidence of theoretical claims ought to be robust to the settings of that instrument. Here we make explicit the settings used to construct empirical network snapshots in the context of relational interaction and communication data. By unifying this representation and clarifying the space of these often implicit researcher choices, we reveal that a range of traditional social network problems fall strictly within this construction, and that these problems may not be apparent in implicitly defined and $N = 1$ network settings. To reveal the utility of this perspective, we empirically explore along each of these dimensions: relationship strength, window size, and timespan. Using a population of networks derived from email communication patterns, we explore the roles of these settings for understanding the theory of weak ties, stability, and densification, respectively. By emphasizing the precise dimensions across which networks are derived, this reveals a precise view of the literature where dynamic processes and structure have been previously

conflated. We motivate tools for robustness, find evidence that the lack of stability in networks suggests concern for traditional cross-sectional analysis, and find that network densification in organizational email networks is confounded by overall levels of activity within the system. The range of problems explored here suggest that these dimensions must be made precise and explicit in order to do meaningful comparative, population-level analysis in networks, and that the population-level view allows a novel opportunity to test a breadth of hypotheses from the networks literature.

5.1 Introduction

Interpersonal social relationships are unobservable. However, as researchers we often employ interaction data, communication data, online traces of social interactions, and surveys as instruments to determine the presence or absence of social relationships. Using these data sources, we then construct social networks from these inferred relationships (Golder and Macy 2014). For example, Eagle et al. (2010) investigate the well-theorized role of network diversity on economic success by comparing social networks inferred from cell phone communication data to economic development. This process—using inferred social networks, taken from different communication or interaction media—allows us easily observable and quantifiable relational data to test hypotheses motivated by social theories and empirical observation.

The sources of data to describe relationships between people may be explicitly designed (as in surveys) or found (as in online traces, call records, or email metadata) (Salganik 2017). We focus on the latter, and in particular on the patterns of interactions between people using email communication metadata, where patterns of timing, direction, and total volume of interactions are known, but where content, quality, and sentiment are unknown. While online behavior may differ from offline behavior, it would be a fallacy, as demonstrated by the title of Grippa et al. (2006)—"Email may not reflect the social network"—to suggest that networks generated from online data are necessarily less true than those generated by "ground-truth" survey. Survey methods can reveal how observed data might differ from experimentally designed data (Burke and Kraut 2014; Salganik 2017), but survey-generated networks suffer their own biases. In sociology, the method used to elicit

relationships is defined as the name generator. Social networks among the same individuals will vary across different (survey-based) name generators (Campbell and Lee 1991[1]), just as online networks may differ from those generated from surveys (Grippa et al. 2006; Wuchty and Uzzi 2011) and online networks will vary depending on tie definition (De Choudhury et al. 2010; Marlow 2009). Despite these challenges, interaction and communication data derived from online systems provide a useful tool with which to infer social relationships.

Unfortunately, having observed an interaction or communication between a pair does not uniquely define a relationship. Network construction requires defining some mapping between observed interactions and inferred relationships. This must implicitly or explicitly answer questions of the form: does one email per week, or two phone calls per year suggest the presence of a relationship? What if the calls are unreciprocated and never returned? That is, given a record of interaction events between pairs, one must still infer the presence or absence of a latent relationship—a task distinct from, but related to, predicting future interaction events.[2]

Data availability creates additional constraints on observable networks. Mode of interaction matters—for example, Facebook friendship relationships provide a different view of a network than the relationships implied by frequent message exchange (Marlow 2009) or shared photos (Kahanda and Neville 2009)—although only one mode of interaction may be publicly available or available to academics (such as Facebook friendships, as in Chapter 3, Traud et al. (2011) and Traud et al. (2012)). Beyond this construction, practically, API and data access restrictions might limit the total observed time. Limited windows of observations limits researchers' ability to observe meaningful network evolution; observing networks after their initial formation subjects them to constraints due to left-censoring; and failure to account for cohort effects can yield misleading aggregate behavior (Barbosa et al. 2016). Computational constraints may further impact the size of time windows

[1] In the spirit of this dissertation, we note that Campbell and Lee (1991) uses a population of comparable neighborhood networks to reveal the differences in networks induced by different name generators.

[2] Note that here we do not treat this as an inference task in the *probabilistic* sense, but in the literal sense of drawing conclusions (about the presence or strength of relationships) from (interaction) data. In this chapter, we focus on the structural consequences for network structure across different deterministic functions of relationship structure. This could be easily extended to a more general predictive model, but we leave that task to future work. See Chapter 5.5 for further discussion.

over which this data is aggregated, and this interacts with sampling rate of measurement tools, which themselves may be faster or slower than the rate of social interactions of interest (Clauset and Eagle 2007). Finally, heuristics and field norms may suggest how relationship strengths are incorporated into data analysis (De Choudhury et al. 2010).

Time is handled in a variety of ways in social network data, and under a range of titles (network evolution (Dorogovtsev and Mendes 2002; Kumar et al. 2010; Leskovec et al. 2007); temporal networks (Holme 2015)). A common practice is to consider a series of network snapshots, each of which may be constructed over a range of time. When these snapshots are taken over overlapping time windows, that is, with a sliding time window, this can reveal how quickly network structure varies over time (Kossinets and Watts 2006). Alternatively, sequences of non-overlapping snapshots then yield a sequence of networks for a set of discrete time steps. This yields panel data, or equivalently a tensor, which admits cross-sectional analysis.

Social network theories, such as structural holes (Burt 1992), are typically evaluated using cross-sectional or panel data, but implicitly, these theories describe processes that are dynamic, including information flow, brokerage, and access. The observation of these relationships only occurs at varying rates—potentially quite quickly in communication networks (Kossinets and Watts 2006)—and so the structure captured by a temporal network will vary, depending on the construction of the network and granularity of timescales observed (Clauset and Eagle 2007; De Choudhury et al. 2010). Even after balancing the noise and structure implied by these dynamics, there is still the problem of sampling over time: empirical data of temporal processes is still usually left-censored (i.e., we usually do not observe the system from its beginning), and we may not have enough data to 'wait' long enough to observe all relationships.

Furthermore, we are left with an impression of stability from analyzing networks derived from cross-sectional data. Even looking across a network over time, one typically observes that even if global statistics are stable, but individual (local) statics vary rapidly (Kossinets and Watts 2006). This is holds over both individual identities and the distribution over their connections. However, this invites potentially misleading conclusions drawn from panel data. where individual variability

can be interpreted as a potentially meaningful signal (Burt and Merluzzi 2016). (Alternatively, Quintane and Carnabuci (2016) instead carefully make explicit the temporal process of brokerage across structural holes, and find, roughly, that brokers are inferred to be brokers because they serve as brokers.) Extending to the setting of population-level network analysis, if we observe structure in networks to have meaningful patterns, e.g., be associated with outcomes (Chapter 4) or offline context (Chapter 3), then we would hope that these patterns can be observed across multiple network examples ($N > 1$) and multiple network construction definitions.

Across these different settings, we are still asking one question: does theory stand up to uncertainty in the instrument? Or, as we reveal later, when the settings of this instrument correspond to well-theorized social phenomena, when and how can we use theory to inform our choices of network construction? Conversely, can we use variation in network construction to better understand how these phenomena are exhibited empirically? By being explicit about these settings reveals where the literature has previously conflated temporal processes in networks and artifacts of observation. Here we attend to variables that are often treated as a pre-processing step, if treated explicitly at all, to derive social networks from communication and interaction data. These variables, once made explicit, touch on some of the most foundational ideas in the study of social networks. Explicitly considering our instrument of observation unites our perspectives on cross-sectional data, network evolution, stability, and robustness and reveals novel empirical questions in this setting.

Our contributions

We bring together three variables of network construction, often treated as a preprocessing step, which determine how networks are inferred from a set of interaction data. These variables can be implicitly or explicitly chosen, and potentially beyond control by a researcher. We describe how previous theoretical and empirical work has been drawn along these dimensions, and how this perspective can highlight new research questions. To illustrate the effectiveness of this framework, we apply this perspective empirically to a population of large email networks. First, considering network tie strength, we find that empirical structure varies in expected ways for weak and strong ties (Granovetter 1973), but that very weak ties are qualitatively different. Second, we observe that

individual properties vary rapidly over time, echoing past findings on local instability in networks (Kossinets and Watts 2006). Third, we exploit the population of differently-sized networks to empirically test the concept of network densification (Leskovec et al. 2005b, 2007). We find that, naively, our data support the densification hypothesis, but we also uncover a confounding variable: the activity level of the underlying system. Together, this network construction perspective unites past theoretical work with the challenges of making empirical claims, and this suggests a set of tools to explore the robustness of sociological results and a novel empirical perspective with which to explore the interactions between these social phenomena.

5.2 Network construction

Building social networks from online communication or interaction data is a nontrivial exercise, but is often treated as a preprocessing step. Observed interactions are assumed to be a correlate of relationship strength (Gupte and Eliassi-Rad 2012). We take the inverse approach of Kahanda and Neville (2009) and characterize how network structure varies with relationship strength.[3] We aim to construct network snapshots, i.e., collections of pairwise relationships from non-overlapping periods of time. This results in a sequence of networks, which can also be interpreted as panel data, where interactions might be represented as (i, j, t) for some pair of individuals i and j interacting at time t. For cases as we explore here, we may have attributes such as tie strength in this tuple as well. Analogously, in multi-modal data, this might also include an action type, such as emailing or instant messaging. Together this construct defines a tensor, which admits flexible modeling techniques (Schein et al. 2016).

Different choices of network construction from email communication data will lead to differences in network measures. Wuchty and Uzzi (2011) found that reciprocal email-based ties are a useful proxy for social ties. The degree to which email data encodes meaningful social structure has been examined in email data specifically (Grippa et al. 2006; Quintane and Kleinbaum 2011

[3] The study of social networks necessarily suffers many endogeneities. One relevant aspect is that increased interaction through communication or a social media platform can also increase perceived relationship strength (Burke and Kraut 2014). We ignore this endogeneity here.

and see also Chapter 4.5). De Choudhury et al. (2010) found that networks defined by 5–10 reciprocated emails per year were most predictive of future interaction; for our six month data set, this would correspond to half as many, but it may be that in the full organizational network setting, other levels of relationship strengths—weaker, stronger, or both—may better represent channels of information flow and communication structure relevant to organizational function. Different network differences could result in different inferences drawn from our data (De Choudhury et al. 2010; Hofman et al. 2017), and so comparing across thresholds can serve as a robustness check. (We employed this strategy in Chapter 4.)

Furthermore, selecting an algorithm can yield misleading conclusions. Sampling algorithms can lead to robust but degenerate discoveries of structural patterns, regardless of their true presence (Lee et al. 2006). For example, sampling algorithms or heuristic measures can misleadingly imply the presence of skewed degree distributions (Achlioptas et al. 2009) or community structure (Good et al. 2010) in networks. One natural setting where this occurs is when online networks sample from some hidden, pre-existing offline network—for example, when creating online Facebook relationships between offline friends—where this process has been shown to induce nontrivial structural patterns, including so-called network densification (Pedarsani et al. 2008; Schoenebeck 2013). Specifically with respect to time, the level of temporal resolution used in network construction can reveal wide differences in network structure. Clauset and Eagle (2007) show that deriving snapshots from proximity networks can produce wildly different structure at high resolution.

Here we emphasize the construction of network snapshots from communication or interaction data. The degree to which many of these issues apply to organizational email data is unknown, and it is an empirical question as to how construction will shed light on measurement of network processes. We first set up the notation necessary to speak precisely to the consequences of these choices. Separating these variables will be necessary to meaningfully characterize social networks drawn from interaction data and compare across networks of similar construction.

Tunable knobs: time, observation window, and relationship strength First, we clarify the variables of interest to construct network snapshots from interaction or communication

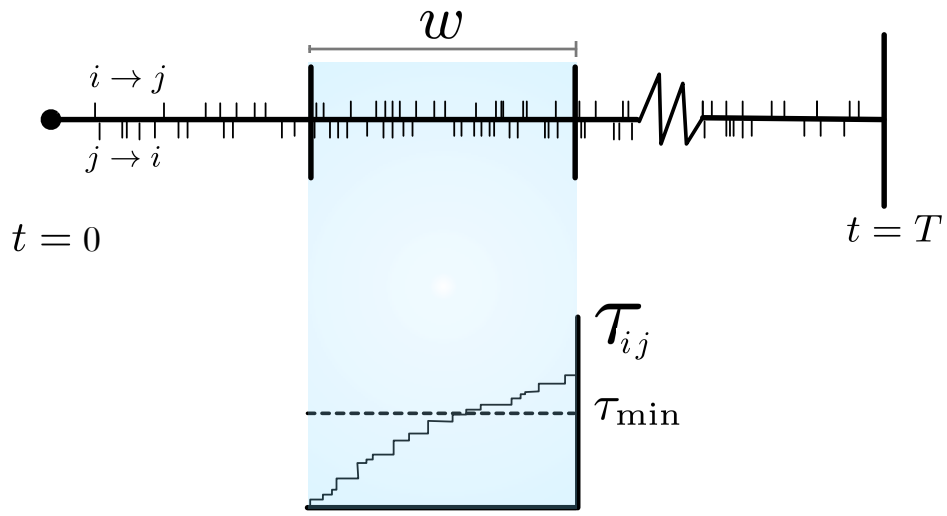

Figure 5.1: **Terms for network construction.** Top, observations of interactions between i and j are observed during time t, $t_0 \leq t \leq T$. Bottom, a network snapshot is constructed from a window of size w. Reciprocated pairwise interactions determine the value of reciprocal interaction strength τ_{ij}, and an edge between i and j is created for that time window if $\tau_{ij} \geq \tau_{\min}$.

data. These are often encoded during preprocessing or implicitly chosen or assigned. Making these variables explicit will allow us to empirically explore the consequences of network construction. Figure 5.1 combines how a set of interactions between a pair, i and j, taken over time $[0, T]$ might be aggregated over some window of length w. The strength of a relationship between i and j here is measured by τ_{ij}, and we may be interested in binarizing this network, where we define edges between i and j during some time window if their relationship strength is observed to be above $\tau_{\min}$, with no edge between them otherwise. This task is a simple network inference question, where we are inferring the presence or absence of a relationship based on some (possibly weighted) reciprocal count data.

We begin by characterizing these terms:

- **Relationship strength or reciprocal tie strength (τ)** As defined in Chapter 2.3 and 5.3.1, this represents the reciprocal interaction strength between two individuals. Specifically, we use the geometric mean of the messages exchanged between two individuals, down-weighted by the number of co-recipients of each exchange. This measure instruments the strength of the pair's relationship. The interpretation of τ is highly dependent on the size of the observation window w—four emails exchanged in one hour, one week, or one year indicate very different relationships (De Choudhury et al. 2010)—as well as the total sampling time and window size ($w \leq T$), rate of observation (for low τ compared to window size), and rate of change in the underlying system.

- **Observation window (w)** The width of the observation window w signals what timescales of human behavior we will capture, by choosing the sensitivity of the data to external influence. For example, w at the level of minutes would provide a sparse and noisy view of any social relationship, with inferred relationships flickering with observations (Clauset and Eagle 2007). This induces sampling effects that are misaligned with longer term social processes of interest: in a university email data set, for example, it is unlikely that strong ties 'forget' their relationship over Thanksgiving break (Kossinets and Watts 2006).

- **Time ($t \leq T$)** We are interested in the role of *time* to mark how a network evolves. While time t primarily serves as an index, where we iterate through the time series of interactions, aggregated over an observation window w. Given some model of underlying dynamics or sampling rate, it will also be a question of when or whether the total amount of time T will be enough to observe the patterns of interest (problems of left-censoring and resolution of social process). In single network ($N = 1$) settings, furthermore, it is difficult to tease apart the role of network size S from time T, and this has created ambiguities in the literature.

Note: snapshots and sliding windows. Let δ be the distance between observed windows. That is, for some window size w and window $w_\alpha = [t_\alpha, t_\alpha + w)$, the next window would start at $t = t_\alpha + \delta$. Here we restrict our analysis to the case where $\delta = w$, i.e., snapshots are non-overlapping, but it is also common to analyze sliding windows of social networks. We leave discussion and empirical exploration of these methods to future work.

Some subtleties First, and trivially, window sizes are limited by the total time of observation: $w \leq T$. The scale of these terms compared to the social system will determine whether one has sufficient time to observe the dynamics in a system is related to the timescale of the social processes of interest (cf. Clauset and Eagle (2007)) the problem of left-censoring (if missing historical data would have revealed pre-existing relationships), or sufficient data to reveal meaningful long-term evolution.

Second, the relationship strength of a pair τ_{ij} and the window size within which the relationship is observed w are closely related. The inferred strength of a relationship will relate to how much time has been made available to observe it. Contacts that exchange birthday cards may have an infrequent, meaningful exchange, whereas a support team/customer relationship may include many exchanges in a short period of time, followed by no future contact. These examples, though caricatures, hint at the subtleties of these measures. Simply converting relationship into a scale-free measure, e.g., by measuring τ_{ij}/w (for example: "5 emails per week; 2 phone calls per year") as a relationship strength implies a model of the relationship between pairwise interaction and time,

even though people communicate on a range of timescales. We emphasize that this is instead an *empirical* question, relevant but beyond the scope of this chapter, as to how these measures would vary together and how to convert relationship strength into a meaningful rate.

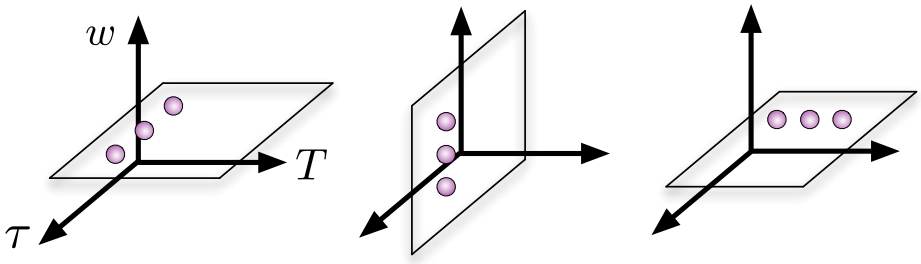

Figure 5.2: **Any point in this parameter space defines a unique network from a set of interaction data.** From left, the first panel represents different networks constructed from different minimum interaction strengths τ. This would reveal networks that vary by tie strength, as we explore in Section 5.4.1; this could also be used to verify robustness of an empirical result (Chapter 4). The second panel represents networks constructed from different observation window sizes, which could reveal differences in stability of network structures, as we explore in Section 5.4.2. The third panel represents networks sampled over different total time spans, which would reveal differences in how a network aggregates, as we explore in Section 5.4.3.

Constructing a network from ($\tau_{\mathbf{min}}$, w, T) We can then imagine this as a three-dimensional space. Then, given some source of interaction data, the location of a point in this space—the selection of these three variables—will uniquely define a network. To yield a *population* of comparable networks, we can imagine fixing a location in this space, and using this to define a network from different sources of interaction data, as we did in Chapter 4.

Alternatively, by sampling different points in this space—for example, fixed w and T but different minimum relationship strengths—we can compare how the choice of network construction yields different network structures. We visualize this idea in Figure 5.2. (As we have noted, these variables are not truly orthogonal. Some of the relationships between these variables are straightforward ($w \leq T$), but as previously discussed, other relationships in this space must be explored empirically (τ vs. w).) In this chapter, we will explore multiple points *along* each dimension, considering one variable at a time. We leave to future empirical work to explore the relationships

between each dimension. While relationships between these dimensions would be quite complex, these single dimensions are still nontrivial to explore. Our approach yields a surprisingly rich area of exploration, where each variable relates to some idea from the study of social networks, particularly where the $N \gg 1$ population setting reveals novel empirical results. Having a single framework with which to reason about these ideas suggests a promising perspective.

Final construction detail: Size from ($\tau_{\min}$, w, T) and time Subtleties abound in network construction. We note one additional detail that appears as a function of the network construction and the observation in time. For each time window observed, we will only observe the senders who were active and sufficiently active during this time. Then the number of active participants observed S may fluctuate across time during observations of an online system. For example, any online platform operator knows this intimately well: the number of user accounts is a meaningfully different statistic than the number of daily or monthly active users, and the number of daily or monthly users may vary with regularity (weekends or seasons) and due to noise. Senders who are not counted during a given time window are still, in some sense, 'meaningful' zeros, in that senders not observed may still reappear in a future network observation, but also do not represent meaningful contribution to network structure during that time. However, we expect from random graph theory and the previous chapters that empirical network structure may vary as a function of network size, and so we must take this into account during measurement.

Potential heterogeneity in observed size and structure then suggests we explicitly highlight this additional detail of interest before proceeding:

- **Size (S)**. The number of active participants observed in a network, i.e., the number of nodes or $|V|$, the size of the vertex set. Network structure is expected to vary with the number of nodes, all other variables held constant. Differences in size can be the result of different instantiations of a network: e.g., two draws from a single network model (*scaling*) or two empirically different but comparable networks; or, the result of a given graph adding (*growing* or *evolving*) or deleting (*evolving*) nodes.

5.3 Data and methods

5.3.1 Data

We take advantage of a unique data set of high-resolution organizational email communication as multiple comparable examples of graph evolution (Chapter 2.2.2). By restricting to within-organization communication, each network is disjoint with clear membership boundaries. For each network, we observe the time series of sending patterns between anonymized senders and receivers, and we analyze these patterns in aggregate, using data from a large commercial enterprise email system.

This yields a population of $N = 65$ unique networks. Considering the fully aggregated networks over the full time period ($w = T$), these networks range in size from approximately 4,500 to 220,000 unique senders, representing almost two orders of magnitude. In each of these sections, we explore different settings including $w = 1$ week and $w = 1$ month, and make our choices explicit as they change in Chapters 5.4.2 and 5.4.3. Smaller time windows capture future senders, and so we will also observe heterogeneity in the number of senders observed, and we explore and exploit this heterogeneity in Chapter 5.4.3.

We define τ following De Choudhury et al. (2010). Tie strength is defined using the geometric mean of the number of messages exchanged between each pair, weighted by the number of recipients on each message. Specifically, for each pair of individuals (i, j), and messages they exchange during some time window $w_\alpha := [t_\alpha, t_\alpha + w]$, we define $I_{ij,\alpha} = \{\iota_1, \iota_2, \ldots, \iota_{m_{ij}}\}$, aggregated over a given time window w_α, we define:

- $m_{ij,\alpha} = |I_{ij,\alpha}|$ = total messages sent from i to j during w_α
- $m_{ji,\alpha} = |I_{ji,\alpha}|$ = total messages sent from j to i during w_α
- Reciprocity $\tau_{ij,\alpha} = \tau_{ji,\alpha} = \sqrt{\omega_{ij} * \omega_{ji}}$, where

$$\omega_{ij} = \sum_{\iota \in I_{ij}} \frac{1}{\text{number of recipients}(\iota)}$$

and similarly for ω_{ji}. Note that $\tau_{ij} = 0$ when the link is unreciprocated.

We note that the reciprocity assumption for inclusion, that i and j must have each received at least one email from each other in a given time window, is actually an assumption of $\tau_{ij} > 0$, as each message could have arbitrarily many recipients.

5.4 Results

We explore the role of each variable on social network structure and, using this perspective, explore three related social phenomena: τ and weak ties in social networks (Granovetter 1973), w and network stability (Kossinets and Watts 2006), and T and densification (Leskovec et al. 2007).

5.4.1 The role of τ and the phenomenon of weak ties

The reciprocity strength τ_{ij} represents the strength of mutual engagement between two senders, by construction.[4] We follow from De Choudhury et al. (2010) and find that network structure varies systematically with relationship strength. While this may now be intuitive—one has more acquaintances and strong ties than strong ties alone—the way that network structure varies, and the degree to which this matches expectations from sociological theory, are non-obvious.

Variation about τ is meaningful in that this variable operationalizes the concept of relationship strength (equivalently: tie strength), varying from "weak ties" to "strong ties." Kilduff and Brass (2010) defines tie strength from the literature:

> "A 'combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie' (Granovetter, 1973, p. 1361). Strong ties are frequent, long-lasting, and affect-laden (Krackhardt, 1992, pp. 218–219), whereas weak ties are 'infrequent and distant' (Hansen, 1999, p. 84)."

The literature on weak ties, a concept made central by Granovetter (1973), emphasizes the utility of access to new information. These arguments are often structural: the local network structure of

[4] A detail about this discussion: here we will refer to variation in network structures determined by threshold $\tau_{\min}$ as variation by τ. That is, we always refer to the graph where all relationships $(i, j) \in E \iff \tau_{ij} \geq \tau_{\min}$. This intentional sloppiness allows us to characterize how network structure varies with tie strength, where ties are only encoded from network data conditional on having sufficiently strong relationship strength.

strong ties (many shared ties) vs. of weak ties (potentially bridges to other parts of the network, with fewer shared ties). This is the basis of Granovetter's formulation: "the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another" (Granovetter 1973). The constraints on this, and the degree to which this is a function of relationship strength or network structure, are pursued in the recent literature (cf. Aral 2016; Aral and Van Alstyne 2011; Bruggeman 2016; Quintane and Carnabuci 2016).

We find that the measure τ interpolates between very weak, weak, and strong ties, and that there is a qualitative shift in the network structure among very weak vs. weak ties, and a *different* shift, aligned with pre-existing theory, between weak and strong ties. These expected results describe how network structure, specifically degree and clustering coefficient, ought to vary between weak and strong ties, but we find non-obvious results for clustering in networks using very weak ties. Finally, we find that ties with overlapping local networks (more *embedded* ties, with more neighbors in common) are stronger than those with non-overlapping networks (i.e., those that serve as *bridges* between communities). However, we also find that weak ties are both infrequent and local *and* infrequent and distant, which suggests a departure from the original definitions put forward by Granovetter, but is supported by recent empirical results from case studies (Quintane and Carnabuci 2016).

Most ties in communication networks are weak ties First, descriptively, it is worthwhile to note that most network ties have relatively low tie strength. The distribution of all tie strengths across each organizational network is illustrated in the density plot in Figure 5.3. Each gray curve represents the distribution over each organization, and the navy blue curve represents the distribution over all observed edges in all organizations ($w = T = 6$ months; $\tau_{\min} = 0$). We find that edge strengths are approximately lognormally distributed, with an additional nontrivial peak at $\tau_{ij} = 1$, where the modal exchange has the same weight as having sent and received exactly one email with a single recipient, and smaller peaks at small-integer combinations of messages exchanged and numbers of recipients. Recall also that we are strictly considering reciprocated email relationships, therefore unanswered emails will not appear in this data set. Nonetheless, we find

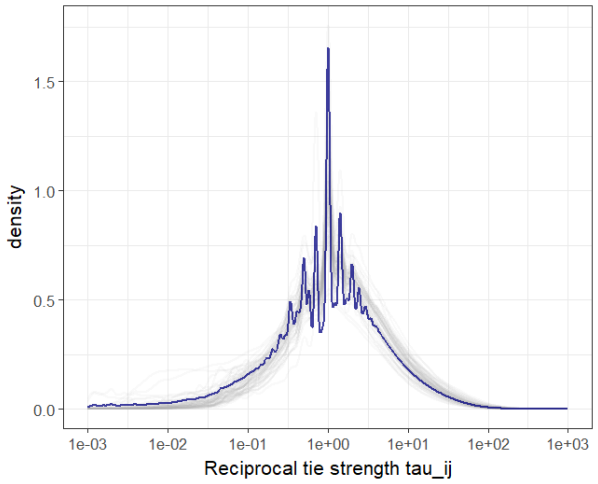

Figure 5.3: **Across organizations, relationships are approximately lognormally distributed with a significant peak at $\tau_{ij} = 1$.** Each gray curve represents the distribution of relationship strengths within a single organization, and the navy curve represents the distribution across all organizations. Across all 65 networks, $w = 1$ month, we note that most ties are quite weak—the median tie strength is 1— and there are additional small peaks for small-integer combinations of sending and receiving.

that most reciprocated pairwise strengths are weak.

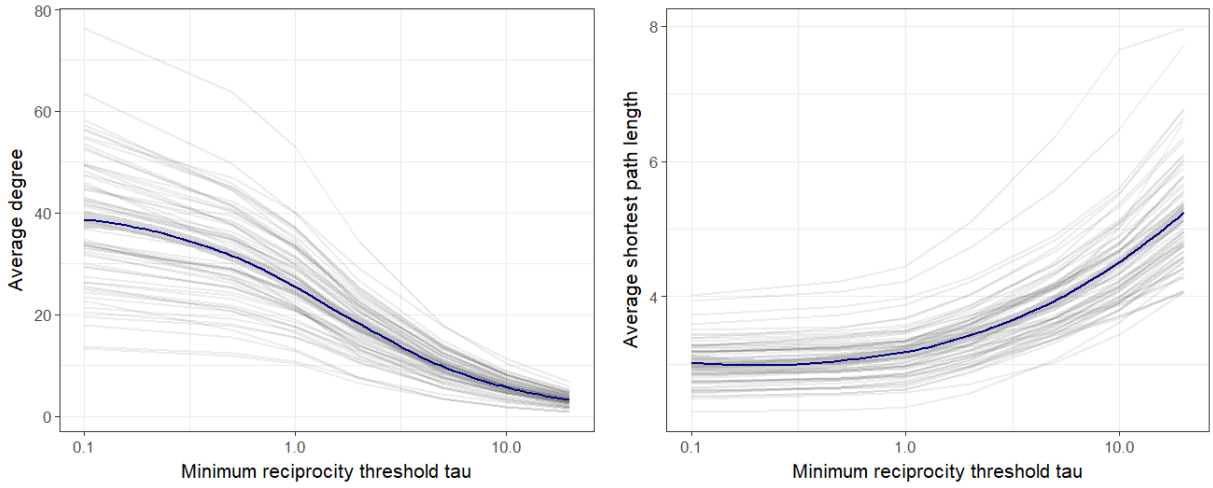

Figure 5.4: **Average degree and average shortest path lenth across all networks as a function of minimum reciprocity threshold τ.** Average degree decreases as $\langle k \rangle \sim 1/\sqrt{\tau}$. Average shortest path length, varying relationship strength τ, increases as $L \sim \sqrt{\tau}$, which matches our expectation that $L \sim \log S/\langle k \rangle$.

τ interpolates from ultra-weak/nearly bipartite ties to weak ties to strong ties; strong ties (and ultra-weak ties) are more clustered First, we recover the reasonable result that average degree decreases as the threshold to be considered a tie increases. This and, relatedly, average shortest path length (which generally changes with degree $L \sim 1/\langle k \rangle$) should increase as ties get deleted. Specifically, we fit a range of functional forms and choose the best fitting model by AIC. (Where not specified, this is the technique applied in all empirical settings in this chapter.) We find that average degree decreases as $\langle k \rangle \propto 1/\sqrt(\tau)$ as average shortest path length increases as $L \propto \sqrt(\tau)$ (Figure 5.4). Note that this matches the expectation from random graph theory that $L \sim \log S/\langle k \rangle$ (Newman (2010); and recall that $L \sim \log S$: Chapter 4).

The empirical distribution of the clustering coefficient, however, suggests a more subtle relationship. For $\tau \in [1, 20]$, clustering coefficient increases with threshold $C \propto \log(\tau)$ (Figure 5.5). This matches what is expected from theory: Granovetter would tell us that clustering should increase with relationship strength, from weak to strong ties (Granovetter 1973). But, compellingly, we see clustering coefficient start high for very weak ties ($\tau \ll 1$) and *decrease* with increasing tie

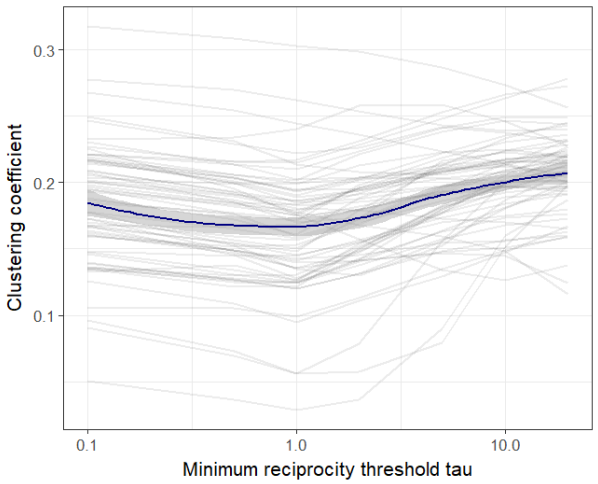

Figure 5.5: **Clustering coefficient over varying minimum relationship strength** τ. Clustering coefficient increases as $C \sim \log \tau$ for $\tau \geq 1$, but clustering coefficient *decreases* from $\tau \ll 1$ to $\tau = 1$. All networks are taken over the full six month time window, $w = T = 6$ months.

strength.[5] Why could this happen? For very weak ties ($\tau \ll 1$), relationships effectively represent co-occurrence or co-membership in groups of recipients, whether formal groups or informal. For τ_{ij} very small, these pairs are effectively recipients of each other's broadcast emails—as anyone who has been caught in a loop of disastrous "Reply All" situations can affirm, these might be unlikely to be meaningful relationships—and this creates dense graphs with high clustering.

We then claim that the reciprocity strength τ interpolates between very weak, weak, and strong ties. We make this distinction because we find that empirically there is a qualitative difference between very weak ties ($\tau \ll 1$), weak ties ($\tau \approx 1$), and strong ties ($\tau \gg 1$). While it could seem obvious that very weak ties by name alone are not evidence of a meaningful social tie, there is no known empirical threshold for interaction or communication data at which this would be necessarily true. Low τ values represent relationships between pairs that do not communicate directly, but pairs that communicate to groups that include the other. In organizations and broader social systems, it is not obvious *a priori* that these relationships are not meaningful. That is, these may be pairs that are aware of each other's existence, are connected in the offline world, or are likely to be connected to each other (through the process of cyclic or triadic closure—see Kossinets and Watts (2006) and next). Regardless, if there is a debate of the value of these very weak ties derived from communication data, then that alone suggests this task is worthwhile and non-obvious. That is, the suggestion that there would exist a relationship strength value in this space that is too weak to count as a meaningful social tie immediately affirms the relevance of this exploration.

Neighborhood overlap is low and constant for very weak ties, but increases for increasingly strong ties Here we ask how *embedded* are relationships of varying tie strength? That is, how similar are ties, conditional on their observed tie strength? Recall the hypothesis from Granovetter: "the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another" (Granovetter 1973). Again, we find that between weak and strong ties, our empirical results match both theory and past empirical results that the degree of

[5] The distributions of clustering coefficient for $\tau = 0.1$ vs. 1 and $\tau = 1$ vs. 5 are significantly different, by two-sided Kolmogorov-Smirnov test, $p = 0.013$, $p = 0.002$, respectively.

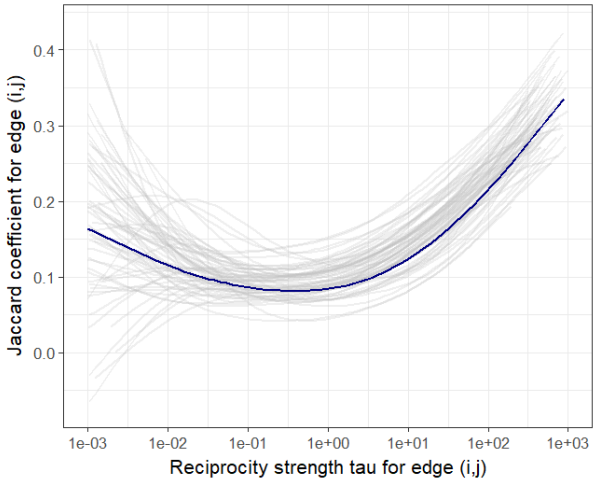

Figure 5.6: Distribution of tie strengths τ_{ij} compared to the Jaccard similiarity of the neighbors of i and j, for all neighbors. Curves shown for $w = T = 1$ month, $\tau_{\min} = 0$. Each gray line shows a smoothing spline fit to each organization's network; aggregating over all organizations, the blue line represents the spline fit to a 10% subsample of all edges.

neighborhood overlap increases with tie strength. Here, we measure this overlap by the Jaccard coefficient of shared neighbors between connected pairs: $J_{ij} = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$ where $N_i := \{j \in V : (i,j) \in E\}$ is the set of neighbors of i.

Figure 5.6 plots the Jaccard similarity over neighbors to the tie strength. We observe that edges with relationship strength $\tau_{ij} \leq 1$ have low similarity (Jaccard similarity about 0.1) but that does not vary. For $\tau_{ij} \geq 1$, similarity of neighborhoods increases with increasing tie strength, as expected. This suggests additional evidence that very weak ties ($\tau \ll 1$) behave qualitatively differently than weak ties ($\tau > 1$).

5.4.2 The role of w and the phenomenon of stability

The window size w defines the timescale of observation of social processes and network structure. Here we focus on *network stability*, where we focus on the distribution of individual properties and how they might vary with time.

Here, we are interested in the construction of network snapshots from which we compute network measures, and these snapshots are then converted into panel data. Network measures,

however, often implicitly capture a temporal process: the social strategy employed by an individual; the flow of information across ties, estimated using these observed structures; or the access or reach of individuals themselves. Given that these measures assume an implicit temporal process, we must then hope that the rate at which that process shifts is slower than the rate of observation.

Consider a practical example: if Kim Kardashian loses a fraction of her total Twitter followers between one month and the next, then her degree centrality will decrease. However, if she is still among the top accounts on Twitter, then this measure (degree centrality) may be useful in characterizing her role in the system. This example is relevant to many online systems, where, for example, there may be many users on Twitter with a few hundred followers, few will be at the comparable scale of Kim Kardashian (over 50 million followers in July 2017). If the underlying network dynamics are fast compared to the window of observation, then major shifts among the top Twitter users would be frequent. In this example, while the nature of celebrity is ephemeral, the suggestion is that the rate of observation is still faster than the mixing of that social system. In this case, cross-sectional analyses would then be overfitting to whatever arrangement happened to be observed at the time. (Burt and Merluzzi 2016 is potentially one example, where the distribution of centrality measures changes rapidly, there is insufficient data to rule out over-fitting, and the authors find that rapid change in the distribution of centrality measures is related to outcomes.) On the other hand, if the dynamics are slower than what is captured by measures taken over the window of observation, then observation over multiple windows could reveal meaningful individual changes.

Stability of central roles over time Cross-sectional analyses that involve the centrality of individuals in social networks use a series of snapshots that use individual positions to predict individual outcomes. This has been applied across social science, from Burt (1992, 2004) to Quintane and Carnabuci (2016) and Uzzi et al. (2016). We ask how stable the distributions of roles are across weeks and months, and the degree to which these properties persist at the individual level. We find that the distribution of network properties varies quickly. The most central members of these organizational networks rapidly lose their position, suggesting that the position itself is

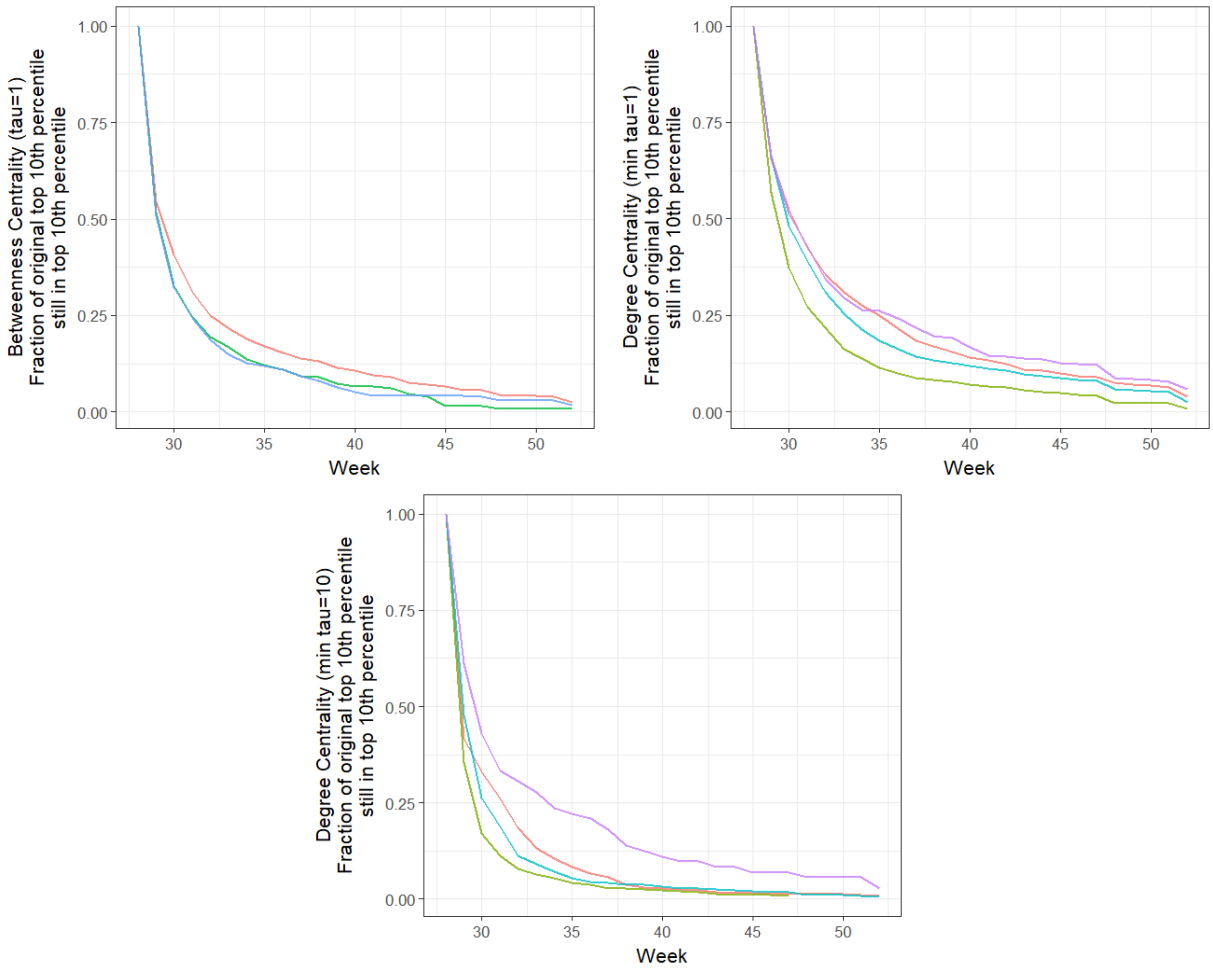

Figure 5.7: Stability of highest-betweenness and highest-degree individuals over time, per week. Each line represents a different organization (only four are shown—preliminary analysis), connecting observations taken for each week ($w = 1$ week, $\delta = 1$ week). For each network snapshot, we compare the top 10% of individuals by betweenness ($\tau = 1$), degree ($\tau = 1$) and degree ($\tau = 10$) to those who were in the top 10% in the first snapshot. The y-axis plots the percentage that have *remained* in the top 10% since the initial observation.

not a persistent property. Given a fixed starting position, the half-life of remaining in that status is quite small. Degree centrality is more persistent than betweenness centrality, and structure at the window size of one week varies much more rapidly and widely than at the window size of one month. We demonstrate this persistence, or lack thereof, of individual roles using a few organizations ($N = 3$ and $N = 4$), varying in size from a few thousand to over one hundred thousand active senders; future work will expand on this population to more accurately characterize the empirical heterogeneity across organizations.

Figure 5.7 plots the fraction of most central nodes (by degree or betweenness) that maintain these roles over time. That is, these figures show what percentage of nodes remain in the top 10% by centrality, conditional on having initially been in it. Then, we can understand Figure 5.7 to reveal the "half-life" of the most central nodes: for the weekly snapshots, half of the original top senders are no longer in the top after one or two weeks. This holds for all three centrality measures, degree centrality for $\tau_{\min} = 1$ and 10 networks and betweenness centrality for $\tau_{\min} = 1$. In Appendix C.1.2, we also show that the half-life is short for $w = 1$ month (Figure C.1.3).

Central nodes have a short half-life; monthly networks are more stable. Figure C.1.3 shows that within two month snapshots, typically half of the most central nodes (by betweenness and strong ties) are no longer the most central, and only about a third of the most central (by betweenness and strong ties) remain in the top after six months. For all centrality measures, this drop-off is precipitous when taken at the weekly snapshot level (Figure 5.7): only about 25% remain in the top after the first three weeks, and most measures drop to almost zero.

About half of most central nodes are central at a given time later; again, monthly networks are more stable. We also show a related, but qualitatively different point to the "half-life" of the most central nodes suggested by Figure 5.7. Figures C.1.4 and C.1.5 instead show how many of the initially central nodes (again, the top 10%) are *currently* in the observed snapshot by month and by week, respectively. At the weekly level, we find wide variation across organizations, and find that only about half of most central users are central at any future time point. At the monthly level, about 2/3 of the original central nodes return at any given future snapshot. This suggests whether,

and how often, the most central nodes are observed to be most central again. Figure C.1.4 and Figure C.1.5 suggest more stability, and they also show greater stability at the month-level than the week level. Monthly (Figure C.1.4), about 2/3 of the original most central nodes will be most central during any given later snapshot. Weekly, only about half of those most central nodes will be most central during that given time.

Together, these results support past work that suggests that while global properties of a network are stable, individual properties may vary rapidly (Kossinets and Watts 2006). Furthermore, this lack of individual stability could yield misleading inferences. In settings that use cross-sectional data, it is often an unstated assumption that the dynamics of the process of interest are relevant to the timescales observed. Centrality measures in particular implicitly encode a temporal process over a static network, and these processes may vary more or less rapidly than the rate of observation (operationalized by w). Robustness checks across various values of w may reveal weaknesses in cross sectional analyses, and these results otherwise suggest caution in the use of cross-sectional analysis. Future work using populations of networks should more thoroughly explore the empirical heterogeneity across networks, the stability of centrality measures and other local and global network features across varying network definitions. A compelling empirical and modeling task could explore the robustness of inferences made using these highly variable network measures, and characterize the conditions under which meaningful measurements can be made.

5.4.3 The role of T and the phenomenon of densification

Variation about T is meaningful in that this allows us to operationalize the total time window of observation. In studying network evolution, this variable is often fixed and handled during preprocessing. Networks that grow or vary over time may also be defined to aggregate their histories (by increasing w until $w = T$, or, equivalently, fixing $w = T$ and increasing T) or may carry over historical structure from past time windows (snapshots drawn from the same social system over time). Then to precisely define dynamic networks, and furthermore make *comparisons* across networks of different ages (and potentially different sizes), we must be explicit about the role of T

when considering a population of networks. While this may seem like a trivial point, this sheds light on empirical network evolution research which focuses on the dynamics of networks that vary over time, and may vary in size, but are only observed in single $N = 1$ settings. Teasing apart size and time is then at the heart of understanding empirical network scaling and network evolution.

In network evolution settings, network size S often varies with the amount of time observed. Consider, for example, the number of users on Facebook in 2005 vs. 2017: about six million users at the end of 2005 (Chapter 3) to over two billion users in June 2017[6] . Then, the role of size immediately becomes intertwined with the role of time. In $N = 1$ network evolution settings, it is not obvious how to tease apart these competing processes.

Network structure is well understood to vary with network size, yet ambiguity remains about the ways these properties vary with size (Newman 2003), growth (Callaway et al. 2001) and evolution (Leskovec et al. 2005b) in random and empirical graphs. One canonical property of random graphs, of many types, is the average shortest path length (mean geodesic) and diameter varying as $O(\log S)$ or $O(\log \log S)$ with the size of the graph S, with constant average degree, by assumption (Newman 2010). Another widely accepted and well-decorated finding is that empirical networks appear to densify: that is, average degree increases, i.e., become more dense.[7] Teasing apart these issues, we find robust empirical evidence of *constant* average degree *across* a population of organizational communication networks. And yet, we find that average degree increases *within* a population of organizational communication networks. Within each organization, the network structure varies with the size of the network observed.

Dissonance between these results and the densification literature is in part due to the attribution of structural properties to network growth (aggregation and evolution) as opposed to variation strictly due to size. However, we note that this distinction is difficult if not impossible to highlight when $N = 1$, i.e., in the absence of a comparative setting. We shed light on this problem using a

[6] `https://newsroom.fb.com/news/2017/06/two-billion-people-coming-together-on-facebook/`

[7] For the moment, we omit an additional argument from Leskovec et al. (2007) about the diameter or distance between nodes shrinking over time. We found *increasing* $O(\log S)$ scaling of shortest path lengths across organizational networks both for firms and in the early Facebook data, but leave analysis within organizations for future work.

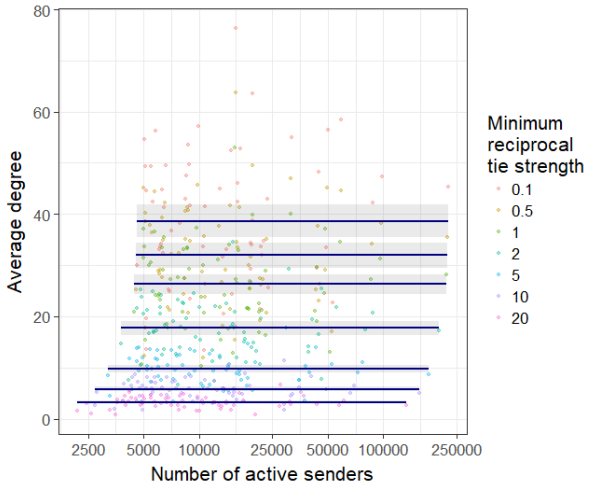

Figure 5.8: Average degree (number of contacts) within an organization, for different definitions of the network (varying $\tau_{\min}$). The regression lines show no relationship between network size S and average degree $\langle k \rangle$ across all network definitions.

rich data set of a population of dynamic communication networks of that are of comparable origin but of different size. We begin by establishing some basic relationships between size and time across this population of networks.

Average degree does not vary with size *across* networks, corresponding to different social systems. However, as demonstrated in Chapter 4, recall that we found that degree did not vary with the number of nodes. (We also found this in Chapter 3, on a different type of social setting.) Here, shown over a range of values of $\tau_{\min}$, Figure 5.8 shows how $\langle k \rangle$ does not vary with size over fixed $w = T$ over a range of values of $\tau_{\min}$. Then in settings where the size of a network increases with time, it would not be feasible to tease apart size from time.

Average degree increases with a greater window of observation (timespan). Trivially, we also note that average degree increases when we aggregate over time, rather than considering non-overlapping snapshots. It is a straightforward observation that during time $t = [0, 2w)$, we observe at least as much as we observed during time $t = [0, w)$. We show this in Figure 5.9 and observe that $\langle k \rangle$ increases with time, where $w = T$ is increasing. A range of the empirical settings considered in Leskovec et al. 2007 consider graphs that aggregated connections over time or ex-

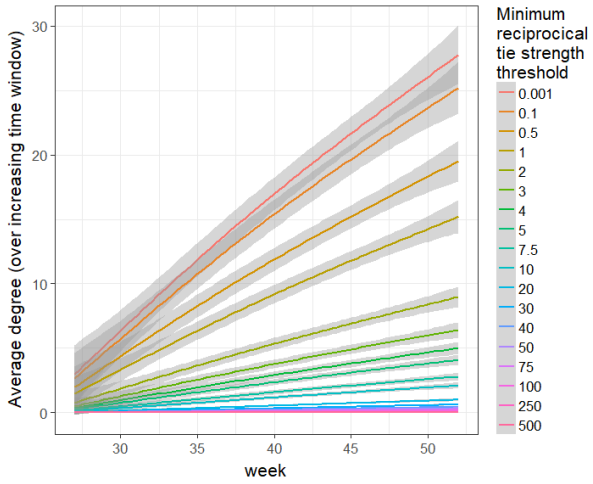

Figure 5.9: Average degree (number of contacts) within a single organization by increasing time window, two standard errors about the mean shown as gray bars. Here, the graph starts at $[0, w = 1 \text{ week}]$, and then increases w by intervals of one week, so the graph is aggregating edges over time. The colors (as shown in the legend) correspond the the minimum reciprocity value τ_{ij} that each edge must have had by the time the window size reached w. (For example, the green and blue lines can be considered the average number of strong ties, which are increasing as the time window increases.) This shows, but does not fully differentiate, that new edges are still being observed as time goes on ($\tau > 0.001$) but also that edges are being activated over time, that is, they reach high enough reciprocity levels as time continues.

hibited increases in the underlying population, such that size would necessarily be increasing with time. However, we note that we must differentiate network size from the timespan of observation, that is, $S \neq T$. The key argument of Leskovec et al. (2007), emphasis ours, states that: " The networks are becoming denser over *time*, with the average degree increasing (and hence with the number of edges growing super-linearly in the *number of nodes*)." Without access to a population, it is not obvious how or even feasible to separate these dimensions of time and number of nodes.

Average degree does not vary meaningfully over time. Having already distinguished the effect of aggregation, we return to the setting of network snapshots of fixed window size w, $w < T$. We note that both average degree and observed network size does not vary meaningfully over time. Whereas Leskovec et al. (2007) considered examples with aggregation or systems where the system size increased with time (and, as a result, made mixed claims about the roles of size and time), we are able to distinguish network scaling and network evolution over our observed windows.

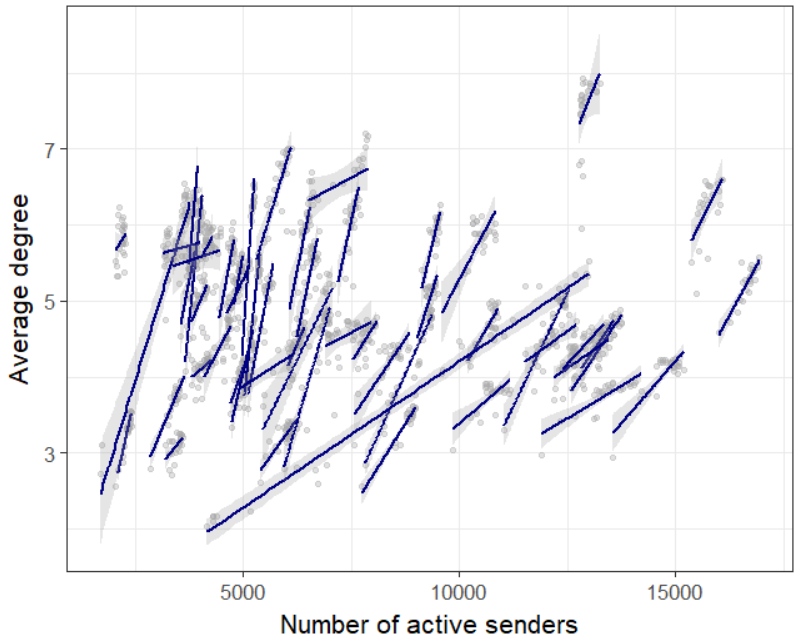

Figure 5.10: **Average degree (number of contacts) within an organization generally increases with the number of observed senders.** Observations are taken across $w = 1$ week periods over $T = 6$ months, minimum $\tau = 1$. We use the 24 week periods for which we have complete coverage. Linear regressions and standard error about the mean are taken across observations for each organization. For visual clarity, we only show organizations with no more than 20,000 weekly active senders; all organizations are shown in Figure C.1.7.

So, where does this leave us?

Is densification real? In order to address this question, we must be precise about the concept of network scaling vs. network evolution. As we saw in Chapter 4, and as we see here in Figure 5.8, average degree does not vary with size *across* a population of networks, even if they are of comparable origin. Within the context of network evolution, however, we do find evidence that *within* a system, average degree does increase with size. That is, within each observed social system, we find evidence of network densification, where average degree increases with size. First we expand on the notion of densification, review the present evidence, and then present a potential confounder and explanation for the densification process observed in this type of empirical online social data.

Densification: some definitions We first define a weak and strong form.

- Weak form: networks *densify* with increasing density, i.e., increasing average degree with the number of nodes. Equivalently, the number of edges increases superlinearly in the number of nodes.

- Strong form: Densification power law. Leskovec et al. (2007) argue for both densification in the general sense and for the "denisfication power law" where $|E| \sim S^{a+1},\ 0 < a < 1$, therefore $\langle k \rangle = |E|/S \sim S^a,\ 0 < a < 1$.[8]

Here we only assert to demonstrate the weak form. We introduce statistical evidence that for some $|E| \sim S^a$, where $a \neq 0$. That is, we show that the number of edges is increasing superlinearly with the size of the network. We demonstrate this by providing statistical evidence that $k = 2|E|/S$ is increasing with S. If $a = 0$ then we would have $|E| \sim S$, and k would be constant with respect to S. This has not previously been systematically demonstrated on a population of comparable networks. However, since we only have 24 network snapshots drawn from each of the 65 different generating distributions, we have insufficient data to suggest a more precise functional form of a scaling relationship.[9]

Average degree increases with size of a network *within* a social system Figure 5.10 is suggestive that average degree increases with size *within* a network. To fully test this hypothesis across all 65 instances, we construct a model to ask whether or not size and degree are positively related. We combine the weekly snapshots data into a single hierarchical linear model, and seek to test whether we can reject the hypothesis that degree does not vary with size.

We frame this by asking if there is a population-level effect of size on degree using a random effects model. This strategy accounts for heterogeneity across firms and mitigating overestimating

[8] In the setting of probabilistic generative models for network structure, there is currently a debate and open technical problem about the "sparsity" of network models (Jacobs and Clauset 2014). In the machine learning literature, Caron and Fox (2014) and Veitch and Roy (2015) and the extant literature consider "sparse" graphs to be precisely those defined by the densification power law setting. This is in contrast to so-called "dense" graphs for which $a = 2$ (Orbanz and Roy 2014). Reconciling these notions of sparsity with generative models with the scaling properties of networks is a meaningful technical challenge (see, e.g., Fosdick et al. (2016)).

[9] Demonstrating this over a set of sliding window snapshots ($\delta < w < T$) would provide more observations that were strongly dependent, based on overlapping windows of the original interaction data. Leskovec et al. (2007) suggest that approaches such as this that only consider active nodes may increase the densification process. We develop a related idea that the distribution of sender activity may suggest an explanation (and confounder) for the observed densification.

	Model: $\langle k \rangle$ under $\tau_{\min} = 1$ $w = 1$ week, $T = 24$ full weeks	Model: $\langle k \rangle$ under $\tau_{\min} = 5$ $w = 1$ week, $T = 24$ full weeks
(Intercept)	11.92^{***}	2.66^{***}
	(1.36)	(0.22)
S_{rescaled}	21.27^{***}	3.24^{***}
	(2.60)	(0.32)
AIC	820.72	-3390.72
BIC	852.01	-3359.43
Log Likelihood	-404.36	1701.36
Num. obs.	1360	1360
Num. groups: Symbol	65	65
Var: Symbol (Intercept)	108.16	3.10
Var: Symbol S_{rescaled}	366.53	6.35
Cov: Symbol (Intercept) S_{rescaled}	190.40	4.33
Var: Residual	0.06	0.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.1: Multilevel model relating average degree to observed network size over 24 weekly snapshots across 65 firms.

effects by multiple comparisons (Gelman et al. 2012). We allow for random intercepts and random slopes, allowing for variation in firm-level effects of size, and ask whether there is a population-level effect of size to predict degree. We find that we can reject the hypothesis that degree does not vary with size, $p < 0.001$ (Table 5.1).[10] In addition, we find that the slope is qualitatively smaller for the $\langle k \rangle$ in the $\tau_{\min} = 5$, which lends support to our hypothesis that densification occurs as a product of shifts in activity.

Densification: the present evidence Why does this happen? How does average degree vary within systems, with network size, and yet not across systems? We review what evidence we have for how networks densify through network evolution, scaling, and aggregation, and present a potential explanation. We suggest that variation in distributions of activity are the primary variable being captured, which increases both observed network size and average degree.

We have found evidence for the following series of observations:

- $\langle k \rangle$ is unrelated to S across networks, fixing comparable time window snapshots (Figure 5.8)
- $\langle k \rangle$ is trivially related to time under aggregation ($w = T$, w increasing) (Figure 5.9)
- $\langle k \rangle$ is not meaningfully to the passage of time across snapshots ($w < T$, fixed w) (Figure C.1.8)
- $\langle k \rangle$ is related to S within networks, across comparable time windows (Figure 5.10)

We suggest that this apparent contradiction is related to the distribution of online activity. Recall that we are in the empirical setting of observing online activity across many offline organizations. Using two proxies for online activity, we first show that these systems vary in activity level during different time windows (intuitively: imagine a worker's output on Friday afternoon, or the week of Thanksgiving, vs. a more active time period). We first show that we observe more

[10] For numerical stability, we rescale S to $S_{\text{rescaled}} = S - \text{mean}(S)/\text{sd}(S)$, and find comparable results to . Then an intercept of 21.27 for the rescaled model corresponds to an intercept of 0.0011 (taking the intercept and dividing by $\text{sd}(S)$). Similarly, for the $\tau_{\min} = 5$ networks, this corresponds to a much smaller slope, 0.0003. These "unscaled" intercepts are similar to the coefficient found under the original model of 0.0011 and 0.0003, respectively. These are also statistically significant, but are unreliable estimates due to poor convergence properties of our algorithm.

active senders during more active time periods. Then high user counts are then snapshots of high activity time periods, and measures that are correlated with higher *activity* will then appear to be a function of larger *size* of observed active users. We attempt to tease apart the relationships between activity and size. We will then show the non-obvious finding that this activity level is positively correlated with degree in the next section.

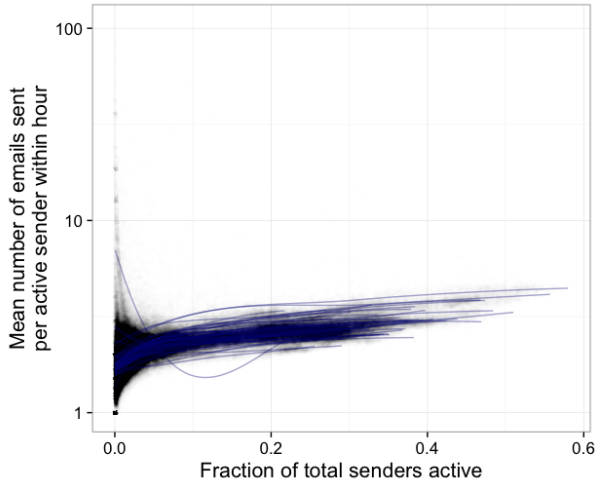

Figure 5.11: **At the hour level, senders send more messages when more other people are active.** Conditional on a sender being active, the mean number of messages sent in an hour period *per active hour user* increases with the fraction of active senders within an organization. Each point represents an observation of the average number of messages sent for a given hour ($w = 1$ hour) across all active senders within that hour. The fraction of active senders is given by N_{observed} divided by the total number of unique active senders ever observed ($T = 6$ months).

Densification: an artifact of activity level? Activity level vs. number of senders observed. First, we consider one measure of activity, number of messages sent, by the number of senders active. There are a number of relationships we could observe: if the distribution of user activity was *unrelated* to the number of active senders, that would suggest that we are taking samples of differently-sized user populations. Alternatively, if the number of senders observed was a function of observing more low activity users, then it is plausible that the average and median numbers of messages would *decrease* as a function of the number of senders observed. On the other hand, if we are instead observing a system where more senders are only active during high activity

periods, and senders were more active during those times, then we would observe the average and median number of messages *increase* as a function of active senders.

We find that the number of senders observed and the activity level are positively correlated. Figure 5.11 shows that the average number of messages sent per active sender increases with the number of active senders in organization observed. The total number of messages and the median per active user increases as well: see Figure C.1.9. Here, we show size as the fraction of all senders in an organization observed, such that size is comparable across networks. Then, *conditional on being active* within a time frame, users are more productive (as a function of total messages sent) during time periods when more other senders are active (as a function of total reciprocated active senders observed).

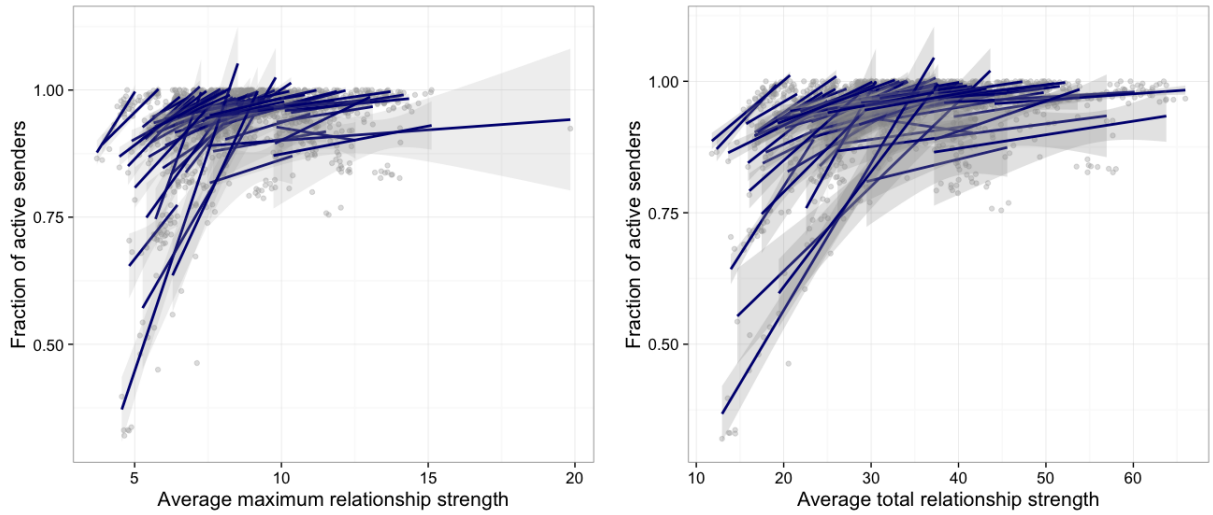

Figure 5.12: **Left, individuals' strongest relationship is observed to be *stronger* when more senders are active. Right, the total weight of relationship strength exchanged is higher when more active senders are observed.** Taken over $w = 1$ week, $T = 24$ fully observed weeks.

Then, we show that the number of senders observed in a given time period increases as a function of activity level.

We first consider the *maximum* relationship strength of each user, i.e., each individual's strongest relationship, and take the average of this maximum across all users in each organization.

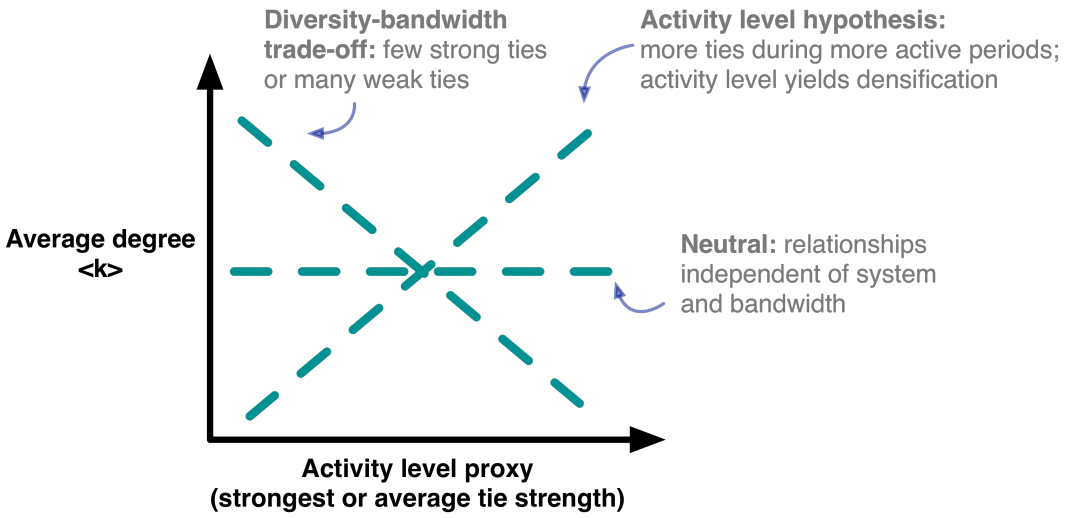

Figure 5.13: **Possible hypotheses for the relationship between activity level and degree.** First, we might find support for the diversity-bandwidth trade-off if we observe low activity (low bandwidth) relationships with higher degree and highly active relationships (high bandwidth) with lower degree. Alternatively, we would observe the opposite if senders exchange message with more contacts during more active periods. If neither or both are true, and in aggregate, individuals' behavior (by sender degree) is independent of activity level in the system, then we should observe no relationship between bandwidth and degree.

The strongest relationship of an individual can vary in multiple ways as a function of the larger social system. First, during times when more active senders, the strength of the strongest relationship may *decrease*, due to constraints on total bandwidth of an individual. (An additional argument for observing weaker relationships with more observed senders would be if the total number of senders observed is only increasing by observing more low activity users.) Second, the strength of this strongest relationship may be *unrelated* to the larger social system: an individual's email relationship with their strongest tie—potentially a collaborator, their boss, or their friend—may be independent of the total number of active users. Third, and what we find here (Figure 5.12, left), is that the number of active observed users *increases* with the strongest relationship strength. This suggests evidence for a positive relationship between activity level in a time period—even of one's strongest relationship—and the total number of users observed. Analogous arguments hold for the *total* relationship strengths per user, and we find these results hold for the average across all users of their total relationship strengths, a proxy for their total activity per user per time window (Figure 5.12, right). As this measure is taken across all users, this suggests that the total number of active senders observed is not simply increasing as a result of observing more low activity users, but that more users are observed during higher activity time periods.

Densification: an artifact of activity level? Activity level vs. degree. Next, we compare sender degree to activity as a function of relationship strengths. Figure 5.13 suggests a range of hypotheses we could support by comparing our proxies for activity level to sender degree.

Degree operationalizes the *diversity* of users' networks—i.e., their degree, or number of contacts—to users' *bandwidth*—i.e., users' total information and communication expenditure, taken here using their relationship strengths $\{\tau_{ij}\}_{j \in N(i)}$. We again consider individuals' strongest relationship using the maximum over all of their neighbors, averaged over all individuals ($\text{mean}_{i \in V} \max_j \tau_{ij}$), and the total of all individuals' reciprocated relationships, averaged over all individuals ($\text{mean}_{i \in V} \sum_j \tau_{ij}$). Following the argument that there should be a trade-off between an individual's diversity of ties to their total bandwidth of all communication (Aral and Van Alstyne 2011), there should be some limit to an individual's capacity for communication across many recipients. Then with higher user

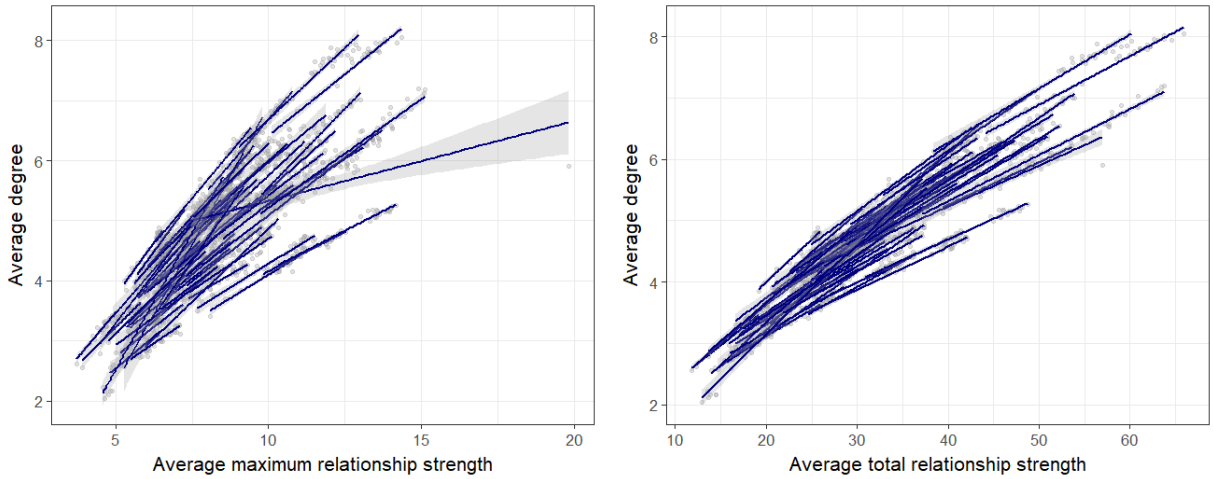

Figure 5.14: **Left, average degree is higher when an individuals' strongest relationship is observed to be *stronger*. Right, average degree increases with total reciprocated relationship strengths.** Together, this suggests that rather than observing a trade-off between bandwidth (total relationship strength, $\sum_j \tau_{ij}$) and diversity (degree k), we are instead observing higher average degree during higher activity time periods. Networks taken over $w = 1$ week, $T = 24$ fully observed weeks, $\tau_{\min} = 1$.

degree, the total bandwidth (total relationship strength) expended should remain constant, and the maximum bandwidth expended on their strongest ties ought to remain constant or decrease. Instead, we observe that, on average, the average degree of the system is positively related to maximum tie strength and average degree is positively related to total bandwidth of the system (Figure 5.14). This means that we observe senders exchanging messages with more people when they are also exchanging more messages with each other, in total and to their strongest tie. This suggests evidence for densification appearing as a consequence of activity level.

As the densification hypothesis is about the relationship between size and degree, a lack of relationship between activity and size and activity and degree would be important to establish this connection. Instead, we find that activity and size and activity and degree are positively related, and so our observations of networks of varying size are also observations of systems of varying activity level. This suggests an alternative explanation of how we might observe densification in online systems: as we tend to observe more active users during more active times, we observe differences in network structure due to differences in activity level, not network size *per se*.

5.5 Discussion

Human social processes represent multiscale temporal processes, and our ability to observe social phenomena depends on the resolution at which we observe these processes. In the context of online social data, we can also observe relationships at a resolution that may be less informative. These temporal processes and measurement questions reveal concrete concerns. For example: if we speak on Monday and Wednesday, are we really no longer socially connected on Tuesday? Am I really less popular because I had one fewer contact that day? Do my acquaintances on this bulk email list broaden my network? Had I stayed home sick, would my network position be lost forever? These questions would be operationalized using relationship strength τ and window size w, about the strength of ties and network stability. For network densification, we can imagine the organizational setting: I might have high incentives to respond frequently to emails from my boss during peak hours when many others are active (high activity, high relationship strength, many

active senders observed), but on average, less during the weekend. For degree related to activity level, I might have incentives to be more available on email when the rest of my team is active (high degree following from high activity).

While these are stylized, the more general versions of these questions demand subtle answers, beyond a simple link prediction or interaction modeling task. Persistent social interactions are meaningful; they can serve, for example, to access to diverse information within an organization (e.g., Aral and Van Alstyne 2011). Then there should be a meaningful way to assess how to describe the best settings to observe phenomena of interest (as in De Choudhury et al. (2010) for optimal τ); whether our observed results are robust (as in Chapter 4). Together this allows us to understand how past work about weak ties, stability, and densification fit together.

Deriving networks from interaction and communication data forces us to confront the instrument we use to observe social networks. These instrument settings determine what we social processes we observe. By being precise about these instrument settings, we can understand past exploration of networks with explicit or implicit temporal processes. This sheds light on the robustness on past results. For example, varying τ, w or T can serve as a robustness check on network results. Alternatively, conditioning on τ, w and T across a *population* of networks can reveal meaningful instability or sources of variation across network structure, where case studies would fall short. Precise language allows us to characterize the space of problems we can now address within the framework of population-level analysis.

By emphasizing the construction of these networks, we find a common framework with which to characterize a qualitative shift between very weak, weak and strong ties; the instability of individual network positions over time; and densification in a population of evolving networks.

We find that the organizational networks show no relationship between average degree and size *across* organizations, but we find evidence of densification (increasing average degree with size) within organizations. This can be thought of as an example of Simpson's paradox, where we find no relationship between these quantities when we fail to account for organization identities, but a strong relationship when we do. However, we also uncovered a confound for the observed pattern of

densification, where networks are observed to be larger and denser during periods of high activity. Furthermore, this activity level is unrelated to network size and evolution in time. Together, this suggests that we do not have evidence for densification. In addition, this observation that senders are observed sending more to more people is in contrast to the "diversity-bandwidth trade-off," where we would expect senders to trade off relationship strengths and quantity.

Limitations & future work The empirical results here are meant to be illustrative, not exhaustive, examples of the questions afforded by the perspective introduced here. Furthermore, a number of these analyses can and should be performed more thoroughly across a range of settings to explore the densification result, the implications of very weak ties, and the stability of individual network positions.

This work directly prods the relationship between network construction and network structure in dynamic communication data, and the patterns that emerge over different time and size scales. Having explored sampling and the strength of relationships, an open question is how to infer 'meaningful' edges in a graph given a time series of observations.

We further note that here we focus on a simple and deterministic model that maps reciprocated communication patterns into weighted relationships. This mapping could be easily extended to a model for prediction. De Choudhury et al. (2010) provides a relevant template for evaluating such a model on network data. With heldout data, within or even *across* networks, this could be treated as a link prediction problem for which we could do more sophisticated probabilistic inference. Our simple model, and related model extensions, could also be combined with simulation data (potentially drawn from the empirical distributions given here) to determine robustness and sensitivity of our measures. We leave these tasks for future work. That is: given that we know that some local changes in structure are meaningful, and some are simply a function of sampling, how can we tell when dynamics in local social network structure are 'real'? We consider this challenge of dynamic network inference in future work. Allowing networks to 'forget' noisy or very weak edges may be the most effective way to balance changes in structure over time; while the sliding window construction of the network allows forgetting of edges with respect to time, it may be relationship

strength that is the useful indicator here. Combining network measures with models of processes on networks (such as information flow) or on edges in networks (such as dyadic communication patterns) may provide a richer problem space, and empirical settings in which we have outcomes at the individual level would provide a meaningful benchmark for activity and network prediction. We consider future work on inferring social networks from dynamic communication data in Chapter 6.2.1.

Typically, and in the setting observed here, empirical network construction also must handle the problem of left-censoring, where relationships existed prior to any data observed. This is related to network stability, where with every fixed window snapshot, we observe relationships as if new. Future work should involve developing the empirical tools, ideally across a population of networks with a longer timespan T, to determine when, if ever, we have fully observed a network. A related question asks when global properties stabilize, if they do. These remain open empirical challenges.

Here we focused on reciprocated relationships. Directed exchanges are meaningful and reflect power and status (Ball and Newman 2013; Guo et al. 2015); meaningful timing in the rate and mode of reciprocation. This is a shortcoming because both requires undirectedness and removes those pairwise dynamics, which is an interesting phenomenon to be studied in future work. Furthermore, relationship strength τ and observation window size w are closely related, but it is an empirical question how they interact. Beyond the network construction questions available here, a meaningful test of the impact of these choices on models of individual outcomes. Relatedly, with this data we lack outcomes for individuals. a meaningful test of the impact of these choices on models of individual outcomes.

A wealth of questions about empirical network densification remains. As a first step, this includes considering how network diameter and average shortest path length vary as a function of size, time, and activity level, but we leave this to future work. It would also be meaningful to deeply explore how network sampling relates to ideas of network densification: we are necessarily exploring the interaction between sampling frequency, local dynamics, and observation windows. Furthermore, past work has shown that densification can emerge as a function of encoding a latent

offline network in an online system (Pedarsani et al. 2008; Schoenebeck 2013). This notion of sampling from an offline network may be a useful model for the setting here, where offline relationships in an organization may exist and be 'sampled' by an email communication process. Aligning a deeper exploration of the role of activity levels with the literature on sampling may be a fruitful future direction.

We largely leave the dynamics of networks, and the relationship between dynamics and network scaling, to future work. It is clear from the previous section that differentiating structural patterns induced by changes in network size from social processes and evolution is a nontrivial task. Using a population of networks—particularly evolving, high-resolution, comparable networks—with simulation suggests a novel opportunity to clarify these terms and tease apart these processes in a structured manner. The framework put forward here should illuminate areas to test past hypotheses and explore novel questions about online social systems.

Conclusions Social network data sets are constructed using a range of choices about the boundary of observations. We unite these settings into a single representation from which we can describe the space of observable networks from a set of interaction data. We demonstrate the effectiveness of adopting this representation through three explorations, in which we confirm, dispute, and offer novel empirical results for previously distinct, foundational areas in social networks research. Crucially, we show how previous work on network densification could have conflated the roles of size and time, given an (understandable) $N = 1$ network perspective. However, this may have further hidden an underlying explanation that differences in user activity may have produced the patterns observed by network densification. Future work in network science will be heavily built off of communication data and, hopefully, increasingly using populations of social systems. We provide the precise and explicit framework that will be necessary to effectively use tease apart the varied and interacting roles of measurement through timescales, dynamics, scaling, evolution, and local and global social processes in networks.

Acknowledgements This work was done in collaboration with Duncan Watts, and was supported by NSF Graduate Research Fellowship award no. DGE 1144083 and Microsoft Re-

search. The authors thank Ashton Anderson, Aaron Clauset, Jake Hofman, Aaron Schein, and Amit Sharma for useful discussions and feedback.

Chapter 6

Discussion and future work

6.1 Contributions

This dissertation advances the idea of using populations of comparable networks. The population-level perspective helps us uncover and understand social processes within and between social networks, dynamics of networks, and subtleties in the methods we use to understand social network structure. Paraphrasing Roosevelt, comparison is the thief of joy and the wealth of opportunity for studying social systems, including social networks, organizations, and online social platforms. Questions that would otherwise be untestable or based on single case studies become available in this context. Characterizing heterogeneity within and across social systems allows us to explore our measures, platforms, and rigorously test theories of social structure. Leveraging the external context of networks allows us to tease apart heterogeneity in social systems due to environment vs. due to network structure, and variation in organizations and platforms allows us to explore empirical scaling properties of graphs.

Using these strategies, we move forward in a number of domains. First, in early Facebook data, we discover that one can detect differences in online social strategies (social search vs. social browsing) due to physical context and differences in product adoption (Chapter 3). Then, using a new data set of organizational communication patterns, we discover the significant diversity in informal network structure of organizations; while this is a challenge to the organizational theory literature, it is also encouraging that we find that productivity is still achievable across a wide range of network structures (Chapter 4). Inspired by the surprising empirical results of Chapters 3 and 4

that average degree does not change with network size, and that diameter does scale logarithmically with network size, we investigate and clarify past literature on empirical graph evolution and stability by investigating how network snapshots are constructed from interaction data (Chapter 5). We show that network densification can be understood as a product of the levels of activity in an online system, and this will drive structure to a greater degree than a diversity-bandwidth trade-off.

6.2 Future work beyond the scope of this dissertation

We look forward to future work that leverages this comparative, population-level structural approach. As laid out in Chapter 1, a number of perspectives have emerged to begin to address existing theory in novel ways and suggest new hypotheses about the structure of social systems. Maps to this new field—as attempted in Chapter 1 and by Hill and Shaw (2017)—will help navigate this area. Future empirical work using online systems will leverage platforms that serve multiple communities (such as Wikia, Reddit, and StackExchange), and enterprise companies and technologies will serve as key contributors and beneficiaries of this type of research.

A number of open technical challenges remain in comparative work. First, the network science community has been lacking large scale comparative analyses across different network types (although see Ghasemian et al. (2017)). Second, theoretical challenges of applying generative network models are nontrivial (D'Amour and Airoldi 2016). The field of neuroscience may be a first area where this is addressed—see, e.g., Durante et al. (2016)—but the setting of online social data will introduce new challenges, including the endogeneities of community assembly, networks of varying size and composition, and differences across sampling methods.

In domains that have begun to grapple with multiple community or network instances, new research questions abound. In the ecological community, the food webs community, among others, have faced debate particularly where ambiguous roles of network structure and size interplay at the global (Dunne et al. 2013) and local level (Klaise and Johnson 2016)). For organization theory, online communities, informal work organizations, and open source communities provide a novel opportunity for computational social science. Here, access to online data has already altered the

view of organization theory on online communities (cf. Benkler 2001, Shaw and Hill 2014).

Next, I describe several concrete future projects that build off of the technical themes in dissertation but are beyond the scope of this document.

6.2.1 Inferring social networks from dynamic communication networks

In this future work expanding on Chapter 5, we are motivated by our results on the large natural heterogeneity in informal social networks in firms (Chapter 4); the significant individual dynamics comprising significant local heterogeneity but relatively stable global dynamics in communication networks (Chapter 5); and, in contrast, the emphasis in the literature on cross-sectional data, which heavily emphasizes small changes in local structure. Cross-sectional data allows for picking up on structural signals that capture a surprising amount of heterogeneity. Instead, the target here is to infer the "best" social network from communication data by varying precision and recall for individual structural characteristics, pairwise relationships, and global network structure. We can quantitatively and precisely explore how network inference relates to network construction. By using a population of comparable networks over time, that presumably evolve on comparable timescales, we can further explore how this inference varies across networks. Together with Chapter 5, this contributes to a story about the distance between social networks and communication networks, as well as the hazards of making inferences from highly dynamic systems about individual- and group-level performance.

6.2.2 Online change point detection for network data

Networks derived from real-world social or communication patterns may go through large-scale structural shifts due to internally and externally induced changes. For example, in an organization, people may get reorganized formally into different divisions or reorganize informally around different projects, and social groups may organize and change over time. In online systems, different groups of people may interact differently depending on the time of day (time zones) or time of year (academic year, holidays), or exogenous events (such as news events); the online platform

may be interested in serving and redistributing computational resources to these different groups efficiently, without inferring exactly the timing of the change point.

While this dissertation has emphasized heterogeneity across a population of networks, a significantly more common setting is to have a single network that varies through time. This section introduces a new project, in collaboration with Aaron Clauset and Hanna Wallach, that brings together several of the themes of this dissertation: specifically, how to determine if and when when meaningful shifts occur in the structure of communication networks over time. We focus on the informal social networks of organizations using the large-scale structure of communication patterns (Chapter 4). Instead of emphasizing the emergent graph properties of the graph, which may themselves be changing over time (Chapter 5), we focus on changes in the underlying generative model, which may correspond to shifts in large-scale social processes (Jacobs and Clauset 2014).

We focus on the setting of temporal networks, where for each $t \in T = \{t_1, t_2, \ldots\}$, we have network $G_t = (V_t, E_t)$ generated from some model $f_t(\theta_t)$. In the case of *nonstationarity* of interest here, the model f_t and parameters θ_t may vary over different values of t. Here we will work under a fixed model assumption, where only the latent variable θ_t is potentially varying with time. We specifically contrast the setting of *change point detection*, i.e., detecting and modeling shifts in the underlying model, to the problem of *anomaly detection*, where we would seek only to detect (and potentially ignore) rogue individual deviations in the data from a baseline model.

Using this methodology, we examine the time series of communication networks derived from email metadata for several real-world organizations. In contrast to Chapter 4, where individual characteristics and real-world events were unknown, we look to different organizational data sources where change points are known, such as as in Enron (Peel and Clauset 2015), in messages among traders (Romero et al. 2016), and in the structure and dynamics of business school students (Uzzi et al. 2016). We specifically explore the problem of online change point detection (Adams and MacKay 2007), where we incrementally observe the network over time (Figure 6.1), as opposed to fully *post hoc* analysis (Peel and Clauset 2015). Automatically characterizing organizational dynamics then creates a meaningful baseline to help understand team assembly, the impact of

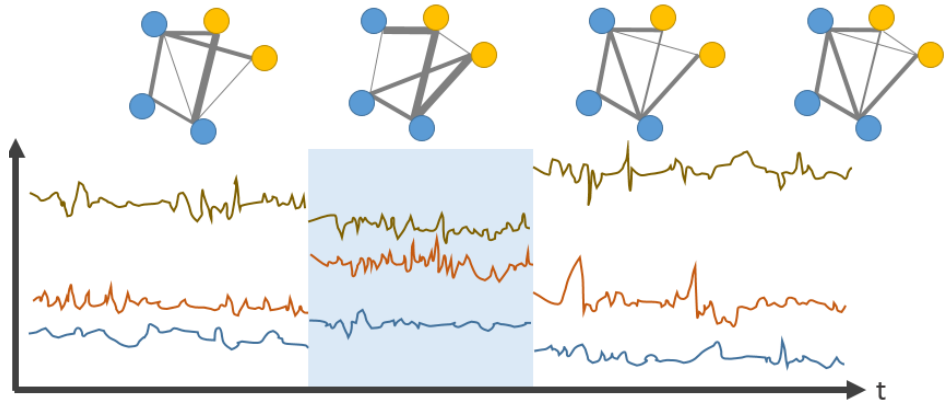

Figure 6.1: **Organizational change points.** We define a change point as an abrupt variation in the parameters of a generative model, such as a stochastic block model. These shifts may happen in conjunction with changes in other network measures, such as assortativity, reciprocity or degree distribution. Here, we show a series of network snapshots that are structurally similar, but still different, within a certain epoch. The squiggly lines may represent the time series of some network measures over all of the networks in that epoch, and one can imagine changes to the network model parameters where some network measures strongly capture this difference (the orange line) and others that have less signal (brown and blue).

management practices, and the scale of responses to outside shocks, even when change points are not known *a priori* (Romero et al. 2016). Rigorous, data-driven approaches to modeling dynamics, communication and structure in organizations will allow us to empirically evaluate previously untested ideas from organization theory in this ubiquitous social and economic setting.

We draw on two lines of work in change point detection and network models. The first is in the automatic detection of change points for networks as a function of large-scale structure, following previous work from Peel and Clauset (Peel and Clauset 2015). In this setting, Peel and Clauset use a single—but descriptively rich—generative model for network structure, the generalized hierarchical random graph model. This detection is done *post hoc*, and the authors use a Bayesian hypothesis test to infer when and whether a change point in the underlying network structure has occurred.

We additionally draw on previous work on *online change point detection*, in which change points are inferred in an online fashion. We specifically take inspiration from the Bayesian online change point detection (BOCD) framework of Adams and MacKay (2007), which yields a fully generative specification of a time series as well as the change points themselves. (This is in contrast

to Shalizi et al. (2011) on online change point detection, where we use an ensemble of models as weighted 'experts' to predict and characterize nonstationarities in time series data. That perspective is oriented towards exploratory data analysis, as opposed to the fully generative specification of Adams and MacKay (2007).) Cai and Adams later introduced variational approximations for the BOCD setting (Cai 2014; Cai and Adams 2015), as did Turner et al. for non-exponential family models (Turner et al. 2013). We also contrast the problem of online change point detection with the problem of (retrospective) topic segmentation and nonstationary time series prediction (Chib 1998; Clements and Hendry 1999; Green 1995; Stephens 1994). These methods have been extended to similarly generative settings, for example, effectively clustering in non-exchangeable settings, particularly time series (Blei and Frazier 2011), or automatically discovering shifts in conversations and language (Nguyen et al. 2014; Purver 2011). While this work similarly focuses on discovering a partition over the temporal sequence of data, we explicitly seek an online approach.

Change points, assembly, and observational data analysis in organizational networks Given organization communication network data, shocks to an organization afford the opportunity to study social processes through shifts in communication structure, content and dynamics. As an illustrative example, recent work by Romero et al. (2016)explored changes in social processes in organizational communication networks in response to exogenous shocks. Romero et al. find that traders in a hedge fund shift towards internal communication, decreased inhibition, and emotional responses to price shocks. This work, exploring network structure around change points known *a priori*, illustrates the potential of using temporal network data to understand social systems, particularly organizations. Similarly, interventions to organization structure—i.e., firm reorganization—support the empirical exploration of tie strength and dynamics before and after these changes: for example, the maintenance of relationships across different network structures and personal attributes (Kleinbaum 2017).

Teasing apart social processes on networks using shocks is similar to the approach we took in Chapter 3 to understand network assembly in early adoption of Facebook. Previously, we considered network assembly in the setting of *post hoc* observational data analysis: revisiting assembly in the

change point framework highlights the opportunities revealed by past work and points to open challenges in this space. Recalling the discussion of future opportunities for research in this space (Chapter 1.1.2), shocks, natural and designed experiments in online systems will reveal a novel opportunities to explore social structure from a methodologically rigorous approach.

6.3 Conclusions and future outlook.

Organizations can take on a diversity of structural forms. In their landmark paper, White et al. (1976) had laid the groundwork for the empirical approach of whole-network analysis, which we carry forward here. White et al. (1976) were already looking ahead to characterizing how such a diversity of organizational forms could emerge, and how this could relate to network evolution and structure:

> " A natural next step, then, is to identify how flows of information and other transactions relate to images and their change. One fundamental problem here is that many social settings may admit not just a single equilibrium outcome, but multiple alternative equilibria, with which particular equilibrium is reached depending in part on accidents of early interaction ... In turn, the interesting questions may bear on what external forces may cause a social structure to pass from one equilibrium configuration to another."

This evolutionary perspective on the diversity of organizations and network structure helps us understand both organization theory and the space of potential future social networks methods.

Social processes within and between systems, historical dependence, and system design create constraints on the diversity of structural forms we observe. Characterizing the extent and sources of empirical heterogeneity across these systems is impossible without adopting a comparative, population-level approach. Online systems provide a broad opportunity to study populations of communities (Hill and Shaw 2017), and studying unsuccessful organizations, whether online or offline, will be necessary to fully reveal the space of configurations (Hill 2013). Furthermore, understanding how our choices of methods impact what we measure is crucial to make meaningful empirical claims, whether based off of algorithms or models (Jacobs and Clauset 2014) or preprocessing (Chapter 5; see also Hofman et al. 2017). Looking forward, enterprise companies and

online platforms that support multiple organizations will provide novel opportunities to empirically explore the diversity, evolution, and assembly of social networks, organizations and online social systems.

Bibliography

Achlioptas, Dimitris, Aaron Clauset, David Kempe, and Cristopher Moore (2009). On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *Journal of the ACM (JACM)* 56.4, p. 21.

Acquisti, Alessandro and Ralph Gross (2006). Imagined communities: Awareness, information sharing, and privacy on the Facebook. *Privacy Enhancing Technologies*. Springer, pp. 36–58.

Adamic, Lada and Eytan Adar (2005). How to search a social network. *Social Networks* 27.3, pp. 187–203.

Adams, Ryan Prescott and David J. C. MacKay (2007). **Bayesian Online Changepoint Detection**. Tech. rep. University of Cambridge.

Ahuja, Gautam (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative Science Quarterly* 45.3, pp. 425–455.

Ahuja, Gautam, Giuseppe Soda, and Akbar Zaheer (2012). The genesis and dynamics of organizational networks. *Organization Science* 23.2, pp. 434–448.

Aicher, Christopher, Abigail Z. Jacobs, and Aaron Clauset (2015). Learning latent block structure in weighted networks. *Journal of Complex Networks* 3.2, pp. 221–248. eprint: `http://comnet.oxfordjournals.org/content/3/2/221.full.pdf+html`.

Alcácer, Juan and Minyuan Zhao (2012). Local R&D strategies and multilocation firms: The role of internal linkages. *Management Science* 58.4, pp. 734–753.

Aldrich, Howard E and Jeffrey Pfeffer (1976). Environments of organizations. *Annual Review of Sociology* 2.1, pp. 79–105.

Altenburger, Kristen M and Johan Ugander (2017). Bias and variance in the social structure of gender. *arXiv preprint arXiv:1705.04774*.

Aral, Sinan (2016). The Future of Weak Ties. *American Journal of Sociology* 121.6, pp. 1931–1939.

Aral, Sinan and Marshall Van Alstyne (2011). The diversity-bandwidth trade-off. *American Journal of Sociology* 117.1, pp. 90–171.

Argote, Linda and Paul Ingram (2000). Knowledge transfer: A basis for competitive advantage in firms. *Organizational behavior and human decision processes* 82.1, pp. 150–169.

Asta, Dena and Cosma Rohilla Shalizi (2014). Geometric Network Comparison. Preprint, *arXiv:1411.1350*.

Backstrom, Lars, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna (2012). Four degrees of separation. *Proc. 3rd ACM Web Science Conference*, pp. 33–42.

Balkundi, Prasad and David A Harrison (2006). Ties, leaders, and time in teams: Strong inference about network structures effects on team viability and performance. *Academy of Management Journal* 49.1, pp. 49–68.

Ball, Brian and Mark E J Newman (2013). Friendship networks and social status. *Network Science* 1.01, pp. 16–30. arXiv: 1205.6822.

Banerjee, Abhijt, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson (2013). The diffusion of microfinance. *Science* 341.6144, p. 1236498.

Barbosa, Samuel, Dan Cosley, Amit Sharma, and Roberto M Cesar Jr (2016). Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior. *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 829–841.

Barrat, Alain, Marc Barthelemy, and Alessandro Vespignani (2008). *Dynamical processes on complex networks*. Vol. 1. Cambridge University Press.

Bascompte, Jordi and Daniel B Stouffer (2009). The assembly and disassembly of ecological networks. *Phil. Trans. Roy. Soc. B* 364.1524, pp. 1781–1787.

Baum, JA and Andrew V Shipilov (2006). Ecological Approaches to Organizations, p. 55.

Baum, Joel AC and Jitendra V Singh, eds. (1994a). *Evolutionary dynamics of organizations*. Oxford University Press.

— (1994b). Organizational niches and the dynamics of organizational founding. *Organization Science* 5.4, pp. 483–501.

— (1994c). Organizational niches and the dynamics of organizational mortality. *American Journal of Sociology*, pp. 346–380.

Baum, Joel AC, Andrew V Shipilov, and Tim J Rowley (2003). Where do small worlds come from? *Industrial and Corporate Change* 12.4, pp. 697–725.

ben-Aaron, James, Matthew Denny, Bruce Desmarais, and Hanna Wallach (2017). Transparency by Conformity: A Field Experiment Evaluating Openness in Local Governments. *Public Administration Review* 77.1, pp. 68–77.

Benkler, Yochai (2001). Coases penguin, or Linux and the nature of the firm. *arXiv preprint cs.CY/0109077* 8.

Blau, Peter M (1965). The comparative study of organizations. *ILR Review* 18.3, pp. 323–338.

— (1970). A formal theory of differentiation in organizations. *American Sociological Review*, pp. 201–218.

Blau, Peter Michael and Richard A Schoenherr (1971). *The structure of organisations*. Basic Books New York.

Blei, David M and Peter I Frazier (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research* 12.Aug, pp. 2461–2488.

Bloom, Nicholas and John Van Reenen (2010). Why do management practices differ across firms and countries? *The Journal of Economic Perspectives* 24.1, pp. 203–224.

boyd, danah (2006). None of this is real: Identity and participation in Friendster. *Structures of Participation in Digital Culture*. Ed. by Joe Karaganis. SSRC.

— (2013). "White flight in networked publics? How Race and Class Shaped American Teen Engagement with Myspace and Facebook". *Race after the Internet*. Routledge, pp. 203–22.

boyd, danah m and Nicole B Ellison (2007). Social network sites: Definition, history, and scholarship. *J. Computer- Mediated Communication* 13.1, pp. 210–230.

Bruggeman, Jeroen (2016). The Strength of Varying Tie Strength: Comment on Aral and Van Alstyne. *American Journal of Sociology* 121.6, pp. 1919–1930.

Burke, Moira and Robert E Kraut (2014). Growing closer on Facebook: changes in tie strength through social network site use. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, pp. 4187–4196.

Burt, Ronald S (1992). *Structural holes: The social structure of competition*. Cambridge: Harvard.

— (1997). The contingent value of social capital. *Administrative Science Quarterly*, pp. 339–365.

— (2000). The network structure of social capital. *Research in Organizational Behavior* 22, pp. 345–423.

— (2002). Bridge decay. *Social Networks* 24.4, pp. 333–363.

— (2004). Structural holes and good ideas. *American Journal of Sociology* 110.2, pp. 349–399.

— (2010). *Neighbor networks: Competitive advantage local and personal*. Oxford University Press.

Burt, Ronald S and Jennifer Merluzzi (2016). Network Oscillation. *Academy of Management Discoveries* 2.4, pp. 368–391.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14.5, pp. 365–376.

Cai, Diana (2014). "Scalable Methods for Bayesian Online Changepoint Detection". Undergraduate Thesis. Harvard University.

Cai, Diana and Ryan P Adams (2015). "Efficient variational approximations for flexible online changepoint detection". Unpublished tech report.

Callaway, Duncan S, John E Hopcroft, Jon M Kleinberg, Mark E J Newman, and Steven H Strogatz (2001). Are randomly grown graphs really random? *Physical Review E* 64.4, p. 041902.

Campbell, Karen E and Barrett A Lee (1991). Name generators in surveys of personal networks. *Social Networks* 13.3, pp. 203–221.

Caron, Francois and Emily B Fox (2014). Bayesian nonparametric models of sparse and exchangeable random graphs. *NIPS Workshop on Frontiers in Network Analysis*. arXiv: 1401.1137.

Carroll, Glenn R (1984). Organizational ecology. *Annual review of Sociology*, pp. 71–93.

— (1985). Concentration and specialization: Dynamics of niche width in populations of organizations. *American journal of sociology* 90.6, pp. 1262–1283.

Carroll, Glenn R and Michael T Hannan (2000). *The demography of corporations and industries*. Princeton University Press.

Centola, Damon (2010). The spread of behavior in an online social network experiment. *Science* 329.5996, pp. 1194–1197.

Centola, Damon and Andrea Baronchelli (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences* 112.7, pp. 1989–1994.

Chib, Siddhartha (1998). Estimation and comparison of multiple change-point models. *Journal of econometrics* 86.2, pp. 221–241.

Child, John (1973). Predicting and understanding organization structure. *Administrative Science Quarterly*, pp. 168–185.

Clauset, Aaron and Nathan Eagle (2007). Persistence and periodicity in a dynamic proximity network. *DIMACS/DyDAn Workshop on Computational Methods for Dynamic Interaction Networks*.

Clauset, Aaron, Cristopher Moore, and M. E. J. Newman (2008). Hierarchical structure and the prediction of missing links in networks. *Nature* 453.7181, pp. 98–101. arXiv: 0811.0484.

Clements, Michael P and David F Hendry (1999). *Forecasting non-stationary economic time series*. MIT Press.

Contractor, Noshir (2013). Some assembly required: leveraging Web science to understand and enable team assembly. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 371.1987, p. 20120385.

Cross, Rob, Stephen P Borgatti, and Andrew Parker (2002). Making invisible work visible: Using social network analysis to support strategic collaboration. *California Management Review* 44.2, pp. 25–46.

Dalton, Dan R, William D Todor, Michael J Spendolini, Gordon J Fielding, and Lyman W Porter (1980). Organization structure and performance: A critical review. *Academy of Management Review* 5.1, pp. 49–64.

D'Amour, Alexander and Edoardo Airoldi (2016). "Misspecification, Sparsity, and Superpopulation Inference for Sparse Social Networks".

Davis, Gerald F (2010). Do theories of organizations progress? *Organizational Research Methods*.

— (2015a). Celebrating Organization Theory: The After-Party. *Journal of Management Studies* 52.2, pp. 309–319.

— (2015b). What Is Organizational Research For? *Administrative Science Quarterly* 60.2, pp. 179–188.

Davis, Gerald F, Mina Yoo, and Wayne E Baker (2003). The small world of the American corporate elite, 1982-2001. *Strategic organization* 1.3, pp. 301–326.

Davis, James A (1970). Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices. *American Sociological Review*, pp. 843–851.

Davis, James A and Samuel Leinhardt (1972). "The structure of positive interpersonal relations in small groups." *Sociological Theory in Progress*. Houghton-Mifflin.

De Choudhury, Munmun, Winter A Mason, Jake M Hofman, and Duncan J Watts (2010). Inferring relevant social networks from interpersonal communication. *Proceedings of the 19th international conference on World Wide Web*. ACM, pp. 301–310.

Diesner, Jana, Terrill L Frantz, and Kathleen M Carley (2005). Communication networks from the Enron email corpus It's always about the people. Enron is no different. *Computational & Mathematical Organization Theory* 11.3, pp. 201–228.

DiMaggio, Paul and Walter W Powell (1983). The iron cage revisited: Collective rationality and institutional isomorphism in organizational fields. *American Sociological Review* 48.2, pp. 147–160.

Dobrev, Stanislav D and Glenn R Carroll (2003). Size (and competition) among organizations: modeling scale-based selection among automobile producers in four major countries, 1885–1981. *Strategic Management Journal* 24.6, pp. 541–558.

Dorogovtsev, Sergey N and Jose FF Mendes (2002). Evolution of networks. *Advances in physics* 51.4, pp. 1079–1187.

Ducheneaut, Nicolas and Victoria Bellotti (2001). E-mail as habitat: an exploration of embedded personal information management. *interactions* 8.5, pp. 30–38.

Dunne, Jennifer A, Richard J Williams, and Neo D Martinez (2002). Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences* 99.20, pp. 12917–12922.

Dunne, Jennifer A, Kevin D Lafferty, Andrew P Dobson, Ryan F Hechinger, Armand M Kuris, Neo D Martinez, John P McLaughlin, Kim N Mouritsen, Robert Poulin, Karsten Reise, et al. (2013). Parasites affect food web structure primarily through increased diversity and complexity. *PLoS Biol* 11.6, e1001579.

Durante, Daniele, David B Dunson, and Joshua T Vogelstein (2016). Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association*.

Eagle, Nathan and Alex Pentland (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10.4, pp. 255–268.

Eagle, Nathan, Michael Macy, and Rob Claxton (2010). Network diversity and economic development. *Science* 328.5981, pp. 1029–1031.

Eckles, Dean, René F Kizilcec, and Eytan Bakshy (2016). Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences* 113.27, pp. 7316–7322.

Ellison, Nicole B, Charles Steinfield, and Cliff Lampe (2007). The benefits of Facebook "friends": Social capital and college students use of online social network sites. *J. Computer-Mediated Comm.* 12.4, pp. 1143–1168.

Faust, Katherine and John Skvoretz (2002). Comparing Networks across Space and Time, Size and Species. *Sociological Methodology* 32.1, pp. 267–299.

Fire, Michael and Carlos Guestrin (2016). Analyzing Complex Network User Arrival Patterns and Their Effect on Network Topologies. *arXiv preprint arXiv:1603.07445*.

Fisher, Danyel (2004). "Social and Temporal Structures in Everyday Collaboration (dissertation)". PhD thesis.

Fisher, Danyel, Marc Smith, and Howard T Welser (2006). You are who you talk to: Detecting roles in usenet newsgroups. *HICSS*. IEEE.

Flap, Henk, Bert Bulder, and Völker Beate (1998). Intra-organizational networks and performance: A review. *Computational & Mathematical Organization Theory* 4.2, pp. 109–147.

Fosdick, Bailey K, Tyler H McCormick, Thomas Brendan Murphy, Tin Lok James Ng, and Ted Westling (2016). Multiresolution network models. *arXiv preprint arXiv:1608.07618*.

Foster, Lucia, John Haltiwanger, and Chad Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review* 98.1, pp. 394–425.

Garcia, David, Pavlin Mavrodiev, and Frank Schweitzer (2013). Social resilience in online communities: The autopsy of Friendster. *COSN*, pp. 39–50.

Gee, Laura K, Jason J Jones, Christopher J Fariss, Moira Burke, and James H Fowler (2017). The paradox of weak ties in 55 countries. *Journal of Economic Behavior & Organization* 133, pp. 362–372.

Gelman, Andrew and Eric Loken (2014). The Statistical Crisis in Science Data-dependent analysisa garden of forking pathsexplains why many statistically significant comparisons don't hold up. *American Scientist* 102.6, p. 460.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5.2, pp. 189–211.

George, Gerard, Martine R Haas, and Alex Pentland (2014). Big data and management. *Academy of Management Journal* 57.2, pp. 321–326.

Ghasemian, Amir, Homa Hosseinmardi, and Aaron Clauset (2017). Evaluating and comparing overfit in models of network community structure. *Preprint*.

Ghoshal, Sumantra and Christopher A Bartlett (1990). The multinational corporation as an interorganizational network. *Academy of Management Review* 15.4, pp. 603–626.

Gibson, Cristina B and Jennifer L Gibbs (2006). Unpacking the concept of virtuality: The effects of geographic dispersion, electronic dependence, dynamic structure, and national diversity on team innovation. *Administrative Science Quarterly* 51.3, pp. 451–495.

Golder, Scott A and Michael W Macy (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology* 40, pp. 129–152.

Good, Benjamin H, Yves-Alexandre de Montjoye, and Aaron Clauset (2010). Performance of modularity maximization in practical contexts. *Physical Review E* 81.4, p. 046106.

Gooding, Richard Z and John A Wagner III (1985). A meta-analytic review of the relationship between size and performance: The productivity and efficiency of organizations and their subunits. *Administrative Science Quarterly*, pp. 462–481.

Granovetter, Mark (1994). "Business Groups". *The Handbook of Economic Sociology*. Ed. by Neil J Smelser and Richard Swedberg. Princeton University Press, pp. 453–475.

— (2005). The impact of social structure on economic outcomes. *The Journal of Economic Perspectives* 19.1, pp. 33–50.

Granovetter, Mark S (1973). The strength of weak ties. *American Journal of Sociology*, pp. 1360–1380.

Green, Peter J (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82.4, pp. 711–732.

Grippa, Francesca, Antonio Zilli, Robert Laubacher, and Peter A Gloor (2006). E-mail may not reflect the social network. *Proceedings of the North American Association for Computational Social and Organizational Science Conference*, pp. 1–6.

Guare, John (1990). *Six degrees of separation: A play*. Vintage.

Gulati, Ranjay, Nitin Nohria, and Akbar Zaheer (2000). Strategic networks. *Strategic Management Journal* 21.3, p. 203.

Gulati, Ranjay, Maxim Sytch, and Adam Tatarynowicz (2012). The rise and fall of small worlds: Exploring the dynamics of social structure. *Organization Science* 23.2, pp. 449–471.

Guo, Fangjian, Charles Blundell, Hanna M Wallach, Katherine A Heller, and UCL Gatsby Unit (2015). The Bayesian Echo Chamber: Modeling Social Influence via Linguistic Accommodation. *AISTATS*.

Gupte, Mangesh and Tina Eliassi-Rad (2012). Measuring tie strength in implicit social networks. *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, pp. 109–118.

Hamilton, William L, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec (2017). Loyalty in online communities. *arXiv preprint arXiv:1703.03386*.

Hannan, Michael T (2005). Ecologies of organizations: Diversity and identity. *Journal of Economic Perspectives* 19.1, pp. 51–70.

Hannan, Michael T and John Freeman (1977). The population ecology of organizations. *American Journal of Sociology* 82.5, pp. 929–964.

— (1984). Structural inertia and organizational change. *American Sociological Review*, pp. 149–164.

— (1993). *Organizational ecology*. Harvard University Press.

Hansen, Gary S and Birger Wernerfelt (1989). Determinants of firm performance: The relative importance of economic and organizational factors. *Strategic Management Journal* 10.5, pp. 399–411.

Hansen, Morten T (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly* 44.1, pp. 82–111.

Hargittai, Eszter (2007). Whose space? Differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication* 13.1, pp. 276–297.

Heaberlin, Bradi and Simon DeDeo (2016). The evolution of Wikipedias norm network. *Future Internet* 8.2, p. 14.

Healy, Kieran (2015). The performativity of networks. *European Journal of Sociology* 56.02, pp. 175–205.

Hill, Benjamin Mako (2013). Almost Wikipedia: What Eight Early Online Collaborative Encyclopedia Projects Reveal about the Mechanisms of Collective Action. *Essays on Volunteer Mobilization in Peer Production. Ph.D. Dissertation*. Massachusetts Institute of Technology.

Hill, Benjamin Mako and Aaron Shaw (2017). Studying Populations of Online Organizations. In: *Oxford Handbook of Networked Communication*. Oxford University Press.

Hill, Russell A and Robin IM Dunbar (2003). Social network size in humans. *Human Nature* 14.1, pp. 53–72.

Hobbs, William and Margaret E Roberts (2016). How Sudden Censorship Can Increase Access to Information. *Preprint*.

Hofman, Jake M, Amit Sharma, and Duncan J Watts (2017). Prediction and explanation in social systems. *Science* 355.6324, pp. 486–488.

Holme, Petter (2015). Modern temporal network theory: a colloquium. *The European Physical Journal B* 88.9, pp. 1–30.

Holme, Petter and Beom Jun Kim (2002). Growing scale-free networks with tunable clustering. *Physical Review E* 65.2, p. 026107.

Holme, Petter and Jari Saramäki (2012). Temporal networks. *Physics Reports* 519.3, pp. 97–125.

House, Robert, Denise M Rousseau, and Melissa Thomas-Hunt (1995). The Meso Paradigm: A framework for the integration of micro and macro organizational behavior. *Research in Organizational Behavior* 17, pp. 71–114.

Huberman, Bernardo A and Lada A Adamic (2004). "Information dynamics in the networked world". *Complex networks*. Springer, pp. 371–398.

Hutchinson, G Evelyn (1959). Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist* 93.870, pp. 145–159.

Ibarra, Herminia, Martin Kilduff, and Wenpin Tsai (2005). Zooming in and out: Connecting individuals and collectivities at the frontiers of organizational network research. *Organization Science* 16.4, pp. 359–371.

Immarigeon, Russ (2014). The 'ungovernable' Ella Fitzgerald. (`https://www.prisonpublicmemory.org/blog/2014/the-ungovernable-ella-fitzgerald`). *Prison Public Memory Project*, October 29, 2014. Accessed June 15, 2017.

Jacobs, Abigail Z. and Aaron Clauset (2014). A unified view of generative models for networks: models, methods, opportunities, and challenges. Preprint, *arXiv:1411.4070; NIPS Networks Workshop 2014*.

Jacobs, Abigail Z., Samuel F. Way, Johan Ugander, and Aaron Clauset (2015a). Assembling thefacebook: Using Heterogeneity to Understand Online Social Network Assembly. *Proc. ACM WebSci*.

Jacobs, Abigail Z., Jennifer A. Dunne, Cristopher Moore, and Aaron Clauset (2015b). Untangling the roles of parasites in food webs with generative network models. Preprint, *arXiv:1505.04741*.

Jansen, Justin JP, Frans AJ Van Den Bosch, and Henk W Volberda (2006). Exploratory innovation, exploitative innovation, and performance: Effects of organizational antecedents and environmental moderators. *Management Science* 52.11, pp. 1661–1674.

Kahanda, Indika and Jennifer Neville (2009). Using Transactional Information to Predict Link Strength in Online Social Networks. *ICWSM* 9, pp. 74–81.

Karinthy, Frigyes (1929). Láncszemek (Chains). *Minden másképpen van. Atheneum, Budapest, Hungary*.

Kilduff, Martin and Daniel J Brass (2010). Organizational social network research: Core ideas and key debates. *The Academy of Management Annals* 4.1, pp. 317–357.

Kilduff, Martin and Wenpin Tsai (2003). *Social Networks and Organizations*. Sage.

Kimberly, John R (1976). Organizational size and the structuralist perspective: A review, critique, and proposal. *Administrative Science Quarterly*, pp. 571–597.

Klaise, Janis and Samuel Johnson (2016). The Origin of Motif Families in Food Webs. *arXiv preprint arXiv:1609.04318*.

Kleinbaum, Adam M (2017). Reorganization and tie decay choices. *Management Science*.

Kleinbaum, Adam M and Toby Stuart (2014). Network Responsiveness: The Social Structural Microfoundations of Dynamic Capabilities. *Academy of Management Perspectives*.

Kleinbaum, Adam M and Michael L Tushman (2007). Building bridges: The social structure of interdependent innovation. *Strategic Entrepreneurship Journal* 1.1-2, pp. 103–122.

Kleinbaum, Adam M, Toby E Stuart, and Michael L Tushman (2013). Discretion within constraint: Homophily and structure in a formal organization. *Organization Science* 24.5, pp. 1316–1336.

Kleinbaum, Adam Michael (2008). "The Social Structure of Organization: Coordination in a Large, Multi-Business Firm". PhD thesis. Harvard University.

Kleineberg, Kaj-Kolja and Marián Boguñá (2015). Digital ecology: Coexistence and domination among interacting networks. *Scientific Reports* 5.

— (2016). Competition between global and local online social networks. *Scientific Reports* 6.

Klemm, Konstantin and Victor M Eguiluz (2002a). Growing scale-free networks with small-world behavior. *Physical Review E* 65.5, p. 057102.

— (2002b). Highly clustered scale-free networks. *Physical Review E* 65.3, p. 036123.

Kogut, Bruce (2000). The network as knowledge: Generative rules and the emergence of structure. *Strategic Management Journal* 21.3, pp. 405–425.

Kogut, Bruce and Gordon Walker (2001). The small world of Germany and the durability of national networks. *American Sociological Review*, pp. 317–335.

Kooti, Farshad, Haeryun Yang, Meeyoung Cha, Krishna P Gummadi, and Winter A Mason (2012). The emergence of conventions in online social networks. *Sixth International AAAI Conference on Weblogs and Social Media*.

Kooti, Farshad, Nathan O Hodas, and Kristina Lerman (2014). Network Weirdness: Exploring the Origins of Network Paradoxes. *ICWSM*.

Kossinets, Gueorgi and Duncan J. Watts (2006). Empirical Analysis of an Evolving Social Network. *Science* 311.88, pp. 88–90.

Kossinets, Gueorgi and Duncan J Watts (2009). Origins of homophily in an evolving social network1. *American Journal of Sociology* 115.2, pp. 405–450.

Krackhardt, David (1994a). Constraints on the interactive organization as an ideal type. *The post-bureaucratic organization: New perspectives on organizational change*, pp. 211–222.

— (1994b). "Graph Theoretical Dimensions of Informal Organizations". *Computational Organizational Theory*. Ed. by Kathleen Carley and Michael Prietula. Lawrence Erlbaum Associates, Inc., pp. 89–111.

Krackhardt, David and Jeffrey R Hanson (1993). Informal networks. *Harvard Business Review* 71.4, pp. 104–111.

Kreiss, Daniel, Megan Finn, and Fred Turner (2011). The limits of peer production: Some reminders from Max Weber for the network society. *New Media & Society* 13.2, pp. 243–259.

Kumar, Ravi, Jasmine Novak, and Andrew Tomkins (2010). "Structure and evolution of online social networks". *Link Mining: Models, Algorithms, and Applications*. Springer, pp. 337–357.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*. ACM, pp. 591–600.

Lahiri, Nandini (2010). Geographic distribution of R&D activity: How does it affect innovation quality? *Academy of Management Journal* 53.5, pp. 1194–1209.

Lampe, Cliff, Nicole Ellison, and Charles Steinfield (2006). A Face(book) in the crowd: Social searching vs. social browsing. *CSCW*, pp. 167–170.

Larremore, Daniel B., Aaron Clauset, and Abigail Z. Jacobs (2014). Efficiently inferring community structure in bipartite networks. *Physical Review E* 90.1, p. 012805. arXiv: 1403.2933.

Laumann, Edward O, Peter V Marsden, and David Prensky (1989). The boundary specification problem in network analysis. *Research methods in social network analysis* 61, p. 87.

Lazega, Emmanuel and Philippa E Pattison (1999). Multiplexity, generalized exchange and cooperation in organizations: a case study. *Social Networks* 21.1, pp. 67–90.

Leana, Carrie R and Harry J Van Buren (1999). Organizational social capital and employment practices. *Academy of Management Review* 24.3, pp. 538–555.

Lee, Sang Hoon, Pan-Jun Kim, and Hawoong Jeong (2006). Statistical properties of sampled networks. *Physical Review E* 73.1, p. 016102.

Leskovec, Jure and Eric Horvitz (2008). Planetary-scale views on a large instant-messaging network. *Proceedings of the 17th International Conference on the World Wide Web*. ACM, pp. 915–924.

Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos (2005a). Graphs over time: densification laws, shrinking diameters and possible explanations. *KDD*, pp. 177–187.

— (2005b). Graphs over time: densification laws, shrinking diameters and possible explanations. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, pp. 177–187.

— (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 497–506.

Lewin, Arie Y and John W Minton (1986). Determining organizational effectiveness: Another look, and an agenda for research. *Management Science* 32.5, pp. 514–538.

Li, Xiao and Karl Rohe (2015). Central limit theorems for network driven sampling. *Preprint, arXiv:1509.04704*.

Liu, Christopher C, Sameer B Srivastava, and Toby E Stuart (2015). An intraorganizational ecology of individual attainment. *Organization Science* 27.1, pp. 90–105.

Lungeanu, Alina, Yun Huang, and Noshir S Contractor (2014). Understanding the assembly of interdisciplinary teams and its impact on performance. *Journal of Informetrics* 8.1, pp. 59–70.

MacArthur, Robert H and Edward O Wilson (1963). An equilibrium theory of insular zoogeography. *Evolution*, pp. 373–387.

Malik, Momin M and Jürgen Pfeffer (2016). Identifying Platform Effects in Social Media Data. *Tenth International AAAI Conference on Web and Social Media*.

Mao, Andrew, Winter Mason, Siddharth Suri, and Duncan J Watts (2016). An experimental study of team size and performance on a complex task. *PloS one* 11.4, e0153048.

March, James G and Herbert Alexander Simon (1958). *Organizations*. Wiley.

March, James G and Robert I Sutton (1997). Crossroads—organizational performance as a dependent variable. *Organization Science* 8.6, pp. 698–706.

Marlow, Cameron (2004). Audience, structure and authority in the weblog community. *ICAC*. Vol. 27.

— (2009). Maintained Relationships on Facebook.

Martin, Travis, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J Watts (2016). Exploring limits to prediction in complex social systems. *Proceedings of the 25th International Conference on World Wide Web*, pp. 683–694.

Matias, J Nathan (2016). Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 1138–1151.

May, Robert M (2009). Food-web assembly and collapse: mathematical models and implications for conservation. *Philosophical Transactions of the Royal Society B* 364.1524, pp. 1643–1646.

Mayer, Adalbert and Steven L Puller (2008). The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics* 92.1, pp. 329–347.

McEvily, Bill, Giuseppe Soda, and Marco Tortoriello (2014). More formally: Rediscovering the missing link between formal organization and informal social structure. *The Academy of Management Annals* 8.1, pp. 299–345.

Michels, Robert (1915). *Political parties: A sociological study of the oligarchical tendencies of modern democracy*. Hearst's International Library Company.

Milgram, Stanley (1967). The small world problem. *Psychology Today* 2.1, pp. 60–67.

Miner, John B (1984). The validity and usefulness of theories in an emerging organizational science. *Academy of Management Review* 9.2, pp. 296–306.

Mitzenmacher, Michael (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1.2, pp. 226–251.

Monge, Peter R and Noshir S Contractor (2001). Emergence of communication networks. *The new handbook of organizational communication: Advances in theory, research, and methods*, pp. 440–502.

— (2003). *Theories of communication networks*. Oxford University Press.

Moreno, Jacob Levy (1934). *Who shall survive?* Beacon House Inc.

— (1953). Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama.

National Center for Education Statistics (2014). Institute of Education Sciences, U.S. Department of Education. Integrated Postsecondary Education Data System (IPEDS), http://nces.ed.gov/ipeds/.

Nelson, Reed E (2001). On the shape of verbal networks in organizations. *Organization Studies* 22.5, pp. 797–823.

Newell, Edward, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths (2016). User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. *ICWSM*, pp. 279–288.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford, UK: Oxford University Press.

Newman, Mark E J (2003). The structure and function of complex networks. *SIAM Review* 45.2, pp. 167–256.

Newman, Mark EJ and Juyong Park (2003). Why social networks are different from other types of networks. *Physical Review E* 68.3, p. 036122.

Nguyen, Viet-An, Jordan Boyd-Graber, Philip Resnik, Deborah Cai, Jennifer Midberry, and Yuanxin Wang (2014). Modeling Topic Control to Detect Influence in Conversations using Nonparametric Topic Models. *Machine Learning* 95, pp. 381–421.

Nohria, Notin and Ranjay Gulati (1994). "Firms and Their Environments". *The Handbook of Economic Sociology*. Princeton University Press, pp. 529–555.

Oktay, Hüseyin, Brian J Taylor, and David D Jensen (2010). Causal discovery in social media using quasi-experimental designs. *Proceedings of the First Workshop on Social Media Analytics*. ACM, pp. 1–9.

Onnela, Jukka-Pekka, Daniel J Fenn, Stephen Reid, Mason A Porter, Peter J Mucha, Mark D Fricker, and Nick S Jones (2012). Taxonomies of networks from community structure. *Physical Review E* 86.3, p. 036104.

Orbanz, Peter and Daniel M Roy (2014). Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–25. arXiv: `arXiv:1312.7857v1`.

Pedarsani, Pedram, Daniel R Figueiredo, and Matthias Grossglauser (2008). Densification arising from sampling fixed graphs. *ACM SIGMETRICS Performance Evaluation Review*. Vol. 36. 1. ACM, pp. 205–216.

Peel, Leto and Aaron Clauset (2015). Detecting change points in the large-scale structure of evolving networks. *AAAI*.

Phan, Tuan Q and Edo M Airoldi (2015). A natural experiment of social network formation and dynamics. *PNAS* 112.21, pp. 6595–6600.

Powell, Walter W, Kenneth W Koput, and Laurel Smith-Doerr (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, pp. 116–145.

Provan, Keith G, Amy Fish, and Joerg Sydow (2007). Interorganizational networks at the network level: A review of the empirical literature on whole networks. *Journal of management* 33.3, pp. 479–516.

Pugh, Derek S, David J Hickson, Christopher R Hinings, and Christopher Turner (1968). Dimensions of organization structure. *Administrative Science Quarterly*, pp. 65–105.

— (1969). The context of organization structures. *Administrative Science Quarterly*, pp. 91–114.

Purver, Matthew (2011). Topic segmentation. *Spoken language understanding: systems for extracting semantic information from speech*, pp. 291–317.

Quintane, Eric and Gianluca Carnabuci (2016). How Do Brokers Broker? Tertius Gaudens, Tertius Iungens, and the Temporality of Structural Holes. *Organization Science* 27.6, pp. 1343–1360.

Quintane, Eric and Adam M Kleinbaum (2011). Matter over mind? E-mail data and the measurement of social networks. *Connections* 31.1, pp. 22–46.

Rapoport, Anatol (1953). Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *Bulletin of Mathematical Biology* 15.4, pp. 523–533.

Ravasz, Erzsébet and Albert-László Barabási (2003). Hierarchical organization in complex networks. *Physical Review E* 67.2, p. 026112.

Reagans, Ray and Bill McEvily (2003). Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly* 48.2, pp. 240–267.

Reagans, Ray and Ezra W Zuckerman (2001). Networks, diversity, and productivity: The social capital of corporate R&D teams. *Organization Science* 12.4, pp. 502–517.

Resnick, Michael D, Peter S Bearman, Robert Wm Blum, Karl E Bauman, Kathleen M Harris, Jo Jones, Joyce Tabor, Trish Beuhring, Renee E Sieving, Marcia Shew, et al. (1997). Protecting adolescents from harm: findings from the National Longitudinal Study on Adolescent Health. *JAMA* 278.10, pp. 823–832.

Ribeiro, Bruno (2014). Modeling and predicting the growth and death of membership-based websites. *WWW*.

Rice, Ronald E (1994). Relating electronic mail use and network structure to R&D work networks and performance. *Journal of Management Information Systems* 11.1, pp. 9–29.

Roethlisberger, Fritz J and William J Dickson (1939). *Management and the Worker*. Harvard University Press.

Rohe, Karl (2015). Network driven sampling; a critical threshold for design effects. *arXiv preprint arXiv:1505.05461*.

Romero, Daniel M., Brian Uzzi, and Jon Kleinberg (2016). Social Networks Under Stress. *WWW*.

Rowley, Timothy J (1997). Moving beyond dyadic ties: A network theory of stakeholder influences. *Academy of Management Review* 22.4, pp. 887–910.

Saavedra, Serguei, Felix Reed-Tsochas, and Brian Uzzi (2008). Asymmetric disassembly and robustness in declining networks. *Proc. Natl. Acad. Sci. USA* 105, pp. 16466–16471.

Salganik, Matthew J (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Schein, Aaron, Mingyuan Zhou, David M Blei, and Hanna Wallach (2016). Bayesian poisson tucker decomposition for learning the structure of international relations. *Proceedings of the 33rd International Conference on Machine Learning*.

Schoenebeck, Grant (2013). Potential networks, contagious communities, and understanding social network structure. *WWW*.

Schuessler, Jennifer (2017). Six Degrees Forevermore. How Six Degrees Became a Forever Meme (`https://www.nytimes.com/2017/04/19/theater/six-degrees-of-separation-meme.html`). *The New York Times*, April 23, 2017: pg. AR1. Published online April 19, 2017. Accessed June 12, 2017.

Schwarz, Gavin M, Stewart Clegg, Thomas G Cummings, Lex Donaldson, and John B Miner (2007). We see dead people? The state of organization science. *Journal of Management Inquiry* 16.4, pp. 300–317.

Shalizi, Cosma Rohilla and Alessandro Rinaldo (2013). Consistency under sampling of exponential random graph models. EN. *Annals of Statistics* 41.2, pp. 508–535.

Shalizi, Cosma Rohilla, Abigail Z. Jacobs, Kristina Lisa Klinkner, and Aaron Clauset (2011). Adapting to non-stationarity with growing expert ensembles. Preprint, *arXiv:1103.0949*.

Sharma, Amit, Jake M Hofman, and Duncan J Watts (2015). Estimating the causal impact of recommendation systems from observational data. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. ACM, pp. 453–470.

Shaw, Aaron and Benjamin M Hill (2014). Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication* 64.2, pp. 215–238.

Shipilov, Andrew V and Stan Xiao Li (2008). Can you have your cake and eat it too? Structural holes' influence on status accumulation and market performance in collaborative networks. *Administrative Science Quarterly* 53.1, pp. 73–108.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22.11, pp. 1359–1366.

Simon, Herbert A (1962). The architecture of complexity. *Proceedings of the American Philosophical Society* 106.6, pp. 467–482.

Smith, D Brent, Benjamin Schneider, and Marcus W Dickson (2006). "Meso Organizational Behaviour: Comments on the Third Paradigm". *The Sage Handbook of Organization Studies*. Ed. by Stewart R. Clegg, Cynthia Hardy, Thomas B. Lawrence, and Walter R. Nord. Sage. Chap. 5, pp. 149–164.

Soda, Giuseppe and Akbar Zaheer (2012). A network perspective on organizational architecture: performance effects of the interplay of formal and informal organization. *Strategic Management Journal* 33.6, pp. 751–771.

Srivastava, Sameer B (2015). Intraorganizational network dynamics in times of ambiguity. *Organization Science* 26.5, pp. 1365–1380.

Staw, Barry M, Lance E Sandelands, and Jane E Dutton (1981). Threat rigidity effects in organizational behavior: A multilevel analysis. *Administrative Science Quarterly*, pp. 501–524.

Stephens, DA (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, pp. 159–178.

Su, Jessica, Aneesh Sharma, and Sharad Goel (2016). The Effect of Recommendations on Network Structure. *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1157–1167.

Tan, Chenhao and Lillian Lee (2015). All who wander: On the prevalence and characteristics of multi-community engagement. *Proceedings of the 24th International Conference on World Wide Web*. ACM, pp. 1056–1066.

TeBlunthuis, Nathan, Aaron Shaw, and Benjamin Mako Hill (2017). Density Dependence Without Resource Partitioning: Population Ecology on Change.org. *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, pp. 323–326.

Traud, Amanda L, Eric D Kelsic, Peter J Mucha, and Mason A Porter (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Review* 53.3, pp. 526–543.

Traud, Amanda L, Peter J Mucha, and Mason A Porter (2012). Social structure of Facebook networks. *Physica A* 391.16, pp. 4165–4180.

Tufekci, Zeynep (2008). Grooming, Gossip, Facebook, and MySpace. *Information, Communication and Society* 11.4, pp. 544–564.

Turner, Ryan D, Steven Bottone, and Clay J Stanek (2013). Online variational approximations to non-exponential family change point models: with application to radar tracking. *Advances in Neural Information Processing Systems*, pp. 306–314.

Tyler, Joshua R, Dennis M Wilkinson, and Bernardo A Huberman (2005). E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society* 21.2, pp. 143–153.

Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow (2011). The anatomy of the Facebook social graph. Preprint, *arXiv:1111.4503*.

Ugander, Johan, Lars Backstrom, and Jon Kleinberg (2013). Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. *Proceedings of the 22nd Interna-*

tional Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 1307–1318.

Uzzi, Brian (1996). The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American Sociological Review*, pp. 674–698.

— (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, pp. 35–67.

Uzzi, Brian and Jarrett Spiro (2005). Collaboration and creativity: The small world problem. *American Journal of Sociology* 111.2, pp. 447–504.

Uzzi, Brian, Luis AN Amaral, and Felix Reed-Tsochas (2007). Small-world networks and management science research: A review. *European Management Review* 4.2, pp. 77–91.

Uzzi, Brian, Yang Yang, and Kevin Gaughan (2016). The Formation and Imprinting of Network Effects Among the Business Elite. *arXiv preprint arXiv:1606.02283*.

van de Rijt, Arnout, Soong Moon Kang, Michael Restivo, and Akshay Patil (2014). Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences* 111.19, pp. 6934–6939.

Veitch, Victor and Daniel M Roy (2015). The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*.

Walsh, Toby (1999). Search in a small world. *IJCAI*. Vol. 99, pp. 1172–1177.

Wang, Xiaoqing, Brian S Butler, and Yuqing Ren (2013). The impact of membership overlap on growth: An ecological competition view of online groups. *Organization Science* 24.2, pp. 414–431.

Warren, Ben H, Daniel Simberloff, Robert E Ricklefs, Robin Aguilée, Fabien L Condamine, Dominique Gravel, Hélène Morlon, Nicolas Mouquet, James Rosindell, Juliane Casquet, et al. (2015). Islands as model systems in ecology and evolution: prospects fifty years after MacArthur-Wilson. *Ecology Letters* 18.2, pp. 200–217.

Wasserman, Stanley and Katherine Faust (1994). *Social network analysis: Methods and applications*. Vol. 8. Cambridge University Press.

Watts, Duncan J (2004). *Six degrees: The science of a connected age*. WW Norton & Company.

— (2017). Should social science be more solution-oriented? *Nature Human Behaviour* 1, p. 0015.

Watts, Duncan J and Steven H Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature* 393.6684, pp. 440–442.

Weber, Matthew S, Janet Fulk, and Peter Monge (2016). The emergence and evolution of social networking sites as an organizational form. *Management Communication Quarterly* 30.3, pp. 305–332.

Weber, Max (1947). *The theory of economic and social organization*. Ed. by Talcott Parsons. Trans. by Alexander Morell Henderson and Talcott Parsons. Oxford University Press.

Wellman, Barry and Stephen D Berkowitz (1988). *Social structures: A network approach*. Vol. 2. Cambridge University Press.

Wellman, Barry and Caroline Haythornthwaite (2008). *The Internet in everyday life*. John Wiley & Sons.

White, Harrison C, Scott A Boorman, and Ronald L Breiger (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology* 81.4, pp. 730–780.

Williams, Richard J and Drew W Purves (2011). The probabilistic niche model reveals substantial variation in the niche structure of empirical food webs. *Ecology* 92.9, pp. 1849–57.

Williamson, Oliver E (1994). "Transaction Cost Economics and Organization Theory". *The Handbook of Economic Sociology*. Princeton University Press, pp. 77–107.

Wuchty, Stefan and Brian Uzzi (2011). Human communication dynamics in digital footsteps: A study of the agreement between self-reported ties and email networks. *PloS One* 6.11, e26972.

Zachary, Wayne W (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research* 33.4, pp. 452–473.

Zaheer, Akbar, Remzi Gözübüyük, and Hana Milanov (2010). It's the connections: The network perspective in interorganizational research. *The Academy of Management Perspectives* 24.1, pp. 62–77.

Zaheer, Srilata, Stuart Albert, and Akbar Zaheer (1999). Time scales and organizational theory. *Academy of Management Review* 24.4, pp. 725–741.

Zhu, Haiyi, Jilin Chen, Tara Matthews, Aditya Pal, Hernan Badenes, and Robert E Kraut (2014). Selecting an effective niche: an ecological view of the success of online communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 301–310.

Zignani, Matteo, Sabrina Gaito, Gian Paolo Rossi, Xiaohan Zhao, Haitao Zheng, and Ben Y Zhao (2014b). Link and Triadic Closure Delay: Temporal Metrics for Social Network Dynamics. *ICWSM*.

— (2014a). Link and Triadic Closure Delay: Temporal Metrics for Social Network Dynamics. *ICWSM* 14, pp. 564–573.

Appendix A

Appendix: Natural experiments in online social network assembly

A.1 Appendix: Facebook100 temporal data

Tables A.1.1 and A.1.2 lists the calendar date that thefacebook become available to students at each of the first 100 colleges. The principle sources for this data are (i) Thefacebook LLC's "Spring 2004 Media Kit," which lists the dates for the first 20 colleges, and (ii) snapshots of the landing page for thefacebook.com as recorded by the Internet Archive (`archive.org`). Exact dates were discernible for 30 schools. When exact dates were not discernible, upper bounds (the latest possible date) were used. Our sources identified 84 of the 100 schools to within a window of at most 3 days. The two schools (Rochester and Bucknell) with the least certain dates are known to fall within a window of 9 days, so may be up to 9 days earlier than listed here.

Table A.1.3 and!A.1.4 lists the start of 2005 freshman orientation for the schools in the facebook100 dataset. Dates were amassed from individual academic calendars, and reflect the start of freshman orientation for non-international students. If such a date could not be found, dates reflect the day dormitories opened. Failing that, the date was set at 1 week before the start of classes. Summer pre-orientations were not considered. Calendars from 2005 were found for 71 of the 100 schools. For the remaining schools a judgement was performed based on more recent calendars and the relative position of orientation/dorms opening to Labor Day on the oldest available calendar vs. Labor Day in 2005. All 100 colleges are located in the United States.

For additional sources and complete methodology, see `http://azjacobs.com/fb100`.

FB100 Index	Name	Date Joined	FB100 Index	Name	Date Joined
1	Harvard	2/4/2004	51	USF	8/21/2004
2	Columbia	2/25/2004	52	UCF	8/21/2004
3	Stanford	2/26/2004	53	FSU	8/21/2004
4	Yale	2/29/2004	54	GWU	8/21/2004
5	Cornell	3/7/2004	55	Johns	8/21/2004
6	Dartmouth	3/7/2004	56	Syracuse	8/22/2004
7	UPenn	3/14/2004	57	Notre Dame	8/22/2004
8	MIT	3/14/2004	58	Maryland	8/22/2004
9	NYU	3/21/2004	59	Maine	9/7/2004
10	BU	3/21/2004	60	Smith	9/7/2004
11	Brown	4/4/2004	61	UC	9/7/2004
12	Princeton	4/4/2004	62	Villanova	9/7/2004
13	Berkeley	4/4/2004	63	Virginia	9/7/2004
14	Duke	4/11/2004	64	UC	9/7/2004
15	Georgetown	4/11/2004	65	Cal	9/7/2004
16	UVA	4/11/2004	66	Mississippi	9/7/2004
17	BC	4/19/2004	67	Mich	9/7/2004
18	Tufts	4/19/2004	68	UCSC	9/7/2004
19	Northeastern	4/19/2004	69	Indiana	9/7/2004
20	UIllinois	4/19/2004	70	Vermont	9/7/2004
21	UF	4/25/2004	71	Auburn	9/7/2004
22	Wellesley	4/25/2004	72	USFCA	9/7/2004
23	Michigan	4/25/2004	73	Wake	9/7/2004
24	MSU	4/25/2004	74	Santa	9/7/2004
25	Northwestern	4/25/2004	75	American	9/7/2004
26	UCLA	4/27/2004	76	Haverford	9/7/2004
27	Emory	4/30/2004	77	William	9/7/2004
28	UNC	4/30/2004	78	MU	9/7/2004
29	Tulane	4/30/2004	79	JMU	9/7/2004
30	UChicago	4/30/2004	80	Texas	9/7/2004

Table A.1.1: Calendar date thefacebook arrived on campus to Facebook100 schools, 1 of 2.

FB100 Index	Name	Date Joined	FB100 Index	Name	Date Joined
31	Rice	4/30/2004	81	Simmons	9/7/2004
32	WashU	5/2/2004	82	Binghamton	9/7/2004
33	UC	5/20/2004	83	Temple	9/7/2004
34	UCSD	5/20/2004	84	Texas	9/7/2004
35	USC	6/23/2004	85	Vassar	9/7/2004
36	Caltech	6/25/2004	86	Pepperdine	9/7/2004
37	UCSB	6/25/2004	87	Wisconsin	9/7/2004
38	Rochester	8/4/2004	88	Colgate	9/7/2004
39	Bucknell	8/4/2004	89	Rutgers	9/7/2004
40	Williams	8/8/2004	90	Howard	9/7/2004
41	Amherst	8/8/2004	91	UConn	9/7/2004
42	Swarthmore	8/8/2004	92	UMass	9/7/2004
43	Wesleyan	8/8/2004	93	Baylor	9/7/2004
44	Oberlin	8/8/2004	94	Penn	9/7/2004
45	Middlebury	8/8/2004	95	Tennessee	9/7/2004
46	Hamilton	8/8/2004	96	Lehigh	9/7/2004
47	Bowdoin	8/8/2004	97	Oklahoma	9/7/2004
48	Vanderbilt	8/21/2004	98	Reed	9/7/2004
49	Carnegie	8/21/2004	99	Brandeis	9/7/2004
50	UGA	8/21/2004	100	Trinity	9/24/2004

Table A.1.2: Calendar date thefacebook arrived on campus to Facebook100 schools, 2 of 2.

FB100 Index	Name	2005 Orientation	FB100 Index	Name	2005 Orientation
1	Harvard	9/10/2005	51	USF	8/22/2005
2	Columbia	8/29/2005	52	UCF	8/17/2005
3	Stanford	9/20/2005	53	FSU	8/20/2005
4	Yale	8/26/2005	54	GWU	8/27/2005
5	Cornell	8/19/2005	55	Johns	8/24/2005
6	Dartmouth	9/14/2005	56	Syracuse	8/24/2005
7	UPenn	9/1/2005	57	Notre Dame	8/19/2005
8	MIT	8/28/2005	58	Maryland	8/24/2005
9	NYU	8/28/2005	59	Maine	9/2/2005
10	BU	8/30/2005	60	Smith	9/2/2005
11	Brown	9/3/2005	61	UC	9/19/2005
12	Princeton	9/7/2005	62	Villanova	8/20/2005
13	Berkeley	8/23/2005	63	Virginia	8/19/2005
14	Duke	8/24/2005	64	UC	9/22/2005
15	Georgetown	8/27/2005	65	Cal	9/12/2005
16	UVA	8/20/2005	66	Mississippi	8/17/2005
17	BC	8/30/2005	67	Mich	8/21/2005
18	Tufts	8/31/2005	68	UCSC	9/17/2005
19	Northeastern	9/1/2005	69	Indiana	8/24/2005
20	UIllinois	8/18/2005	70	Vermont	8/26/2005
21	UF	8/17/2005	71	Auburn	8/10/2005
22	Wellesley	8/29/2005	72	USFCA	8/22/2005
23	Michigan	8/30/2005	73	Wake	8/18/2005
24	MSU	8/25/2005	74	Santa	9/17/2005
25	Northwestern	9/13/2005	75	American	8/21/2005
26	UCLA	9/26/2005	76	Haverford	8/24/2005
27	Emory	8/24/2005	77	William	8/19/2005
28	UNC	8/27/2005	78	MU	8/17/2005
29	Tulane	8/26/2005	79	JMU	8/24/2005
30	UChicago	9/17/2005	80	Texas	8/26/2005

Table A.1.3: Start of 2005 freshman orientation for Facebook100 schools, 1 of 2.

FB100 Index	Name	2005 Orientation	FB100 Index	Name	2005 Orientation
31	Rice	8/14/2005	81	Simmons	9/3/2005
32	WashU	8/11/2005	82	Binghamton	8/25/2005
33	UC	9/26/2005	83	Temple	8/22/2005
34	UCSD	9/15/2005	84	Texas	8/22/2005
35	USC	8/15/2005	85	Vassar	8/30/2005
36	Caltech	9/18/2005	86	Pepperdine	8/23/2005
37	UCSB	9/17/2005	87	Wisconsin	8/25/2005
38	Rochester	8/24/2005	88	Colgate	8/20/2005
39	Bucknell	8/17/2005	89	Rutgers	8/25/2005
40	Williams	8/31/2005	90	Howard	8/20/2005
41	Amherst	8/28/2005	91	UConn	8/26/2005
42	Swarthmore	8/23/2005	92	UMass	8/29/2005
43	Wesleyan	8/31/2005	93	Baylor	8/18/2005
44	Oberlin	8/30/2005	94	Penn	8/25/2005
45	Middlebury	9/7/2005	95	Tennessee	8/13/2005
46	Hamilton	8/20/2005	96	Lehigh	8/25/2005
47	Bowdoin	8/27/2005	97	Oklahoma	8/18/2005
48	Vanderbilt	8/20/2005	98	Reed	8/30/2005
49	Carnegie	8/22/2005	99	Brandeis	8/28/2005
50	UGA	8/15/2005	100	Trinity	9/1/2005

Table A.1.4: Start of 2005 freshman orientation for Facebook100 schools, 2 of 2.

Appendix B

Appendix: A comparative study of informal social networks in firms

B.1 Appendix: Network properties, scaling, and organizational context: additional results

B.1.1 Additional scaling results: network properties and size

Median degree does not vary with network size.

Just as average degree does not vary with network size (Figure 4.3), neither does median degree (Figure B.1.1; $R^2 = 0$). This result is also nontrivial in the organizations literature and network theory literature. We also find insufficient evidence that industry varies meaningfully with median degree: we fail to reject the intercept model ($p = 0.12$).

Scaling in the degree distribution The Gini coefficient of the degree distribution, measuring how unevenly or how skewed the degree distribution is, does not vary with the size of the organization (Figure B.1.2). This is reassuring for the data cleaning process; behavior from real, active email users (not noise induced by our data source) should not become more extremal in large organizations. If we were to, say, drop the threshold and reciprocity requirement, then this may no longer apply: for example, a C.E.O. emailing all of her employees will reach more recipients at a larger firm. However, the reciprocity requirement still implies a constraint on the number of people that can be engaged with directly, so this may be less extreme.

Scaling in the diameter of the networks We find that the diameter of the networks, defined as the longest shortest path between any two senders in the organization, scales like $O(\log S)$, i.e., logarithmically with the size of the organization (Figure B.1.3). This quantity is similar to the

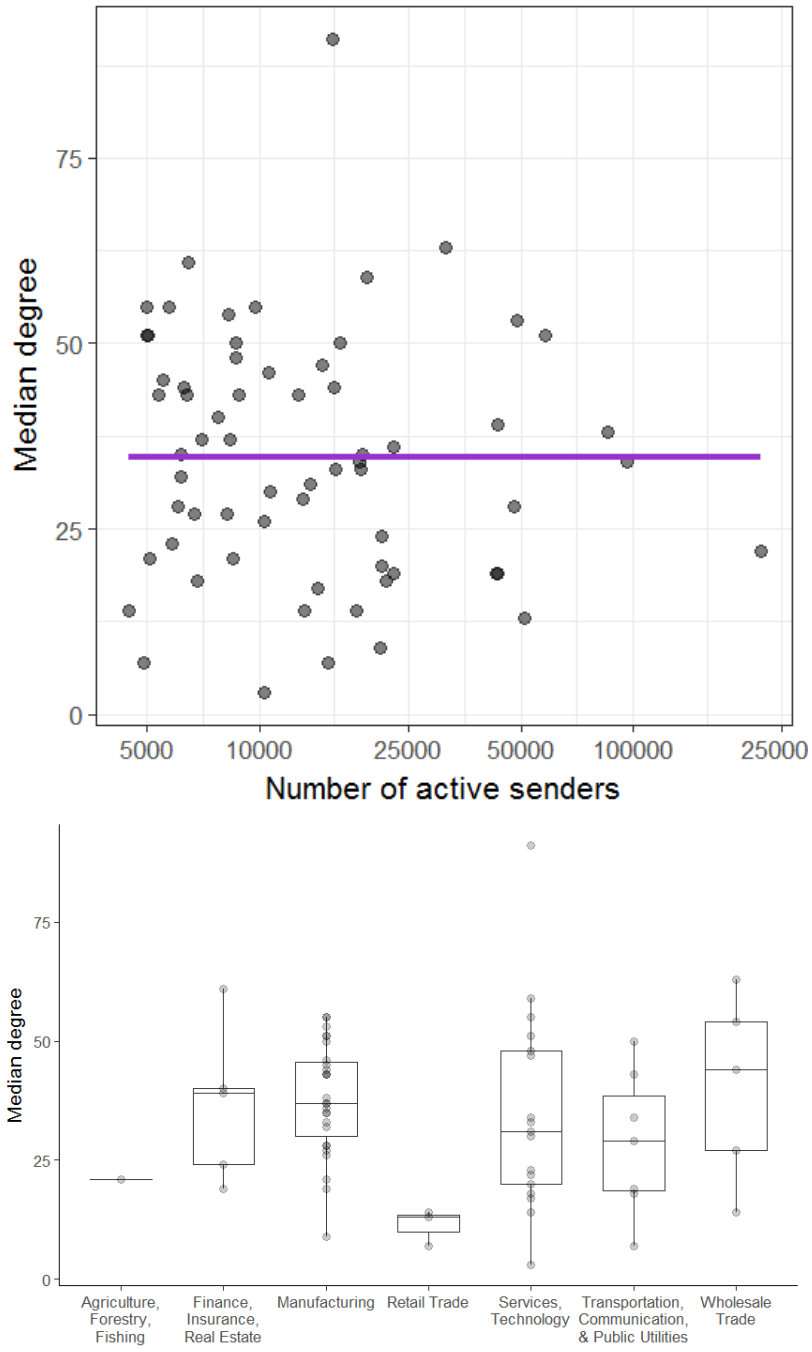

Figure B.1.1: **Median degree does not vary meaningfully with size or industry.** The median degree, i.e., the median number of contacts that a sender has in an organization, does not vary with the size of the organization.

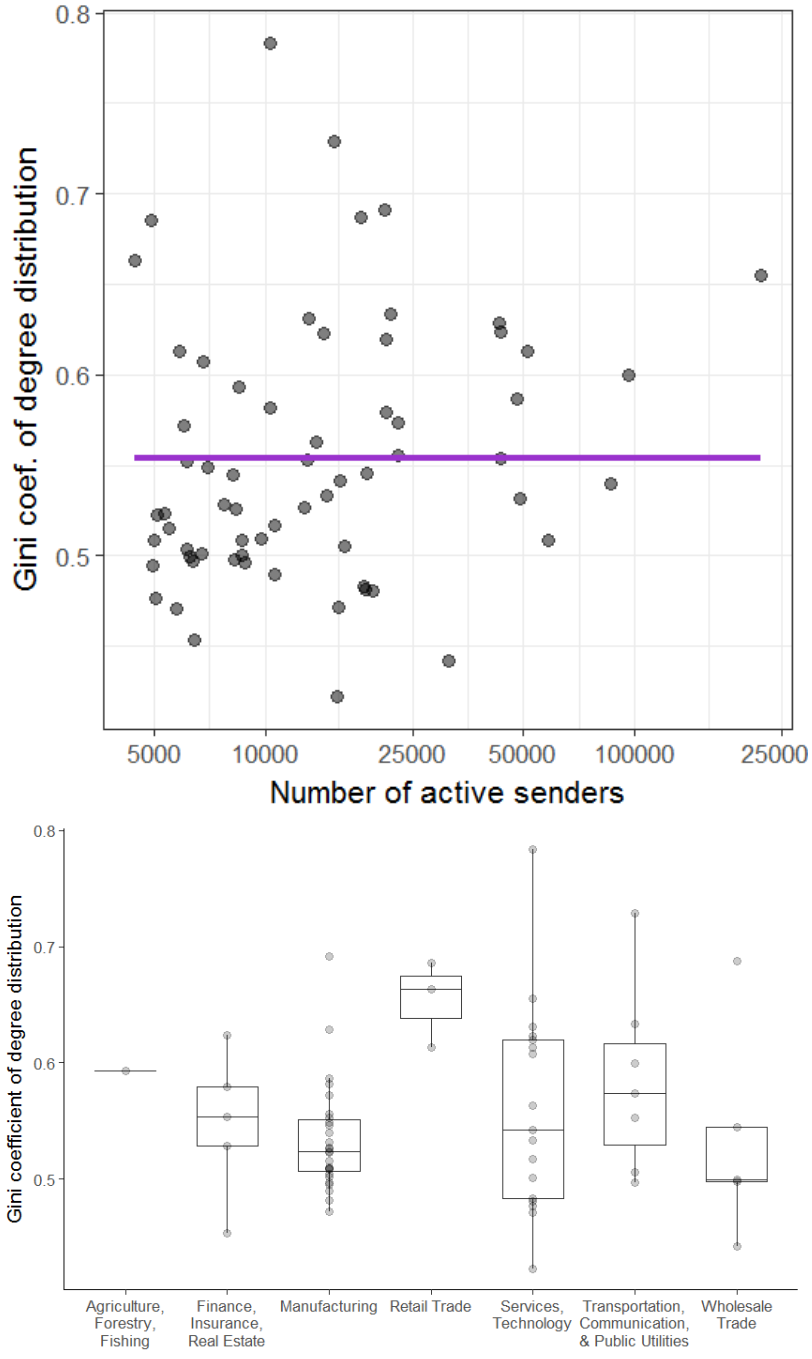

Figure B.1.2: **Additional informal social network features.** The Gini coefficient, a measure of inequality, of the degree distribution. A high Gini coefficient would suggest a very skewed degree distribution, with fewer senders contacting most of the recipients. A low Gini coefficient suggests more evenly distributed numbers of contacts. We find no relationship to size and how (un)evenly distributed contacts are.

average shortest path length (Section 4.4.1), in that it is expected in many random graph models, and is observed in empirical data, to scale logarithmically (Newman (2010)). Given that firms typically have an underlying formal hierarchy, and that informal networks are related to the the formal network structure, it is reasonable that the network can be traversed in $O(\log S)$ hops.

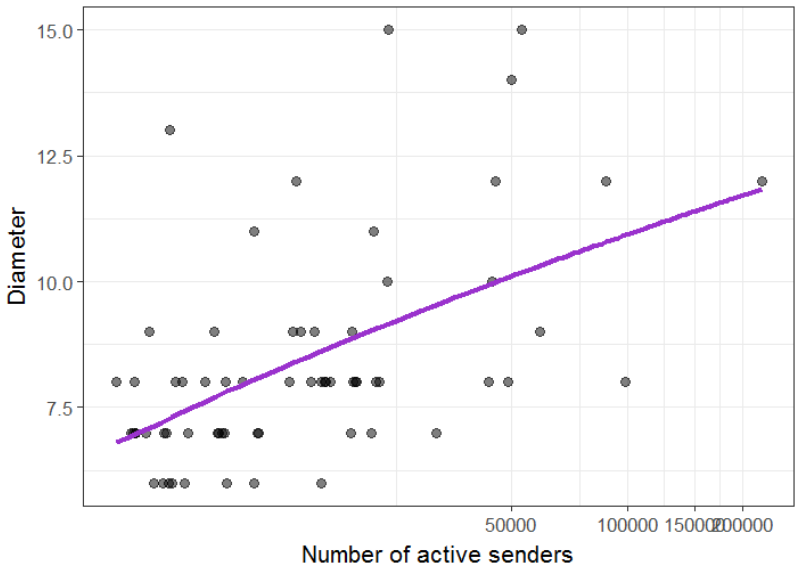

Figure B.1.3: **The diameter of the network increases logarithmically with the size of the network.**

Scaling in the clustering coefficient and small world quotient Clustering coefficient scales too slowly to distinguish similar models, so we fail to reject several similar models. For example, the function $\log(S)/S$ over only two orders of magnitude is difficult to distinguish from a noisy $1/S$ scaling function. Trivially, these networks have much higher clustering than an equivalently-sized network with constant degree and no structure, and we expect that any meaningful social network from a firm, or otherwise, will have higher clustering than $1/S$ in the limit, and even if it becomes small, it will not become arbitrarily close to zero. We find that $1/S$, $1/(S \log S)$, and $\log S/S$ are effectively indistinguishable with the amount (65 data points) and scale of data available (about 2 orders of magnitude). However, by exploring the small world quotient, we can investigate how modified clustering coefficient C/C_R varies, and through that find evidence that

$C = O(\log S/S)$.

Figure B.1.4 finds that C/S increases with $\log S$ of the network ($R^2 = 0.544$). This suggests that we can differentiate models for clustering coefficient, and we should interpret $C = O(\frac{\log S}{S})$.

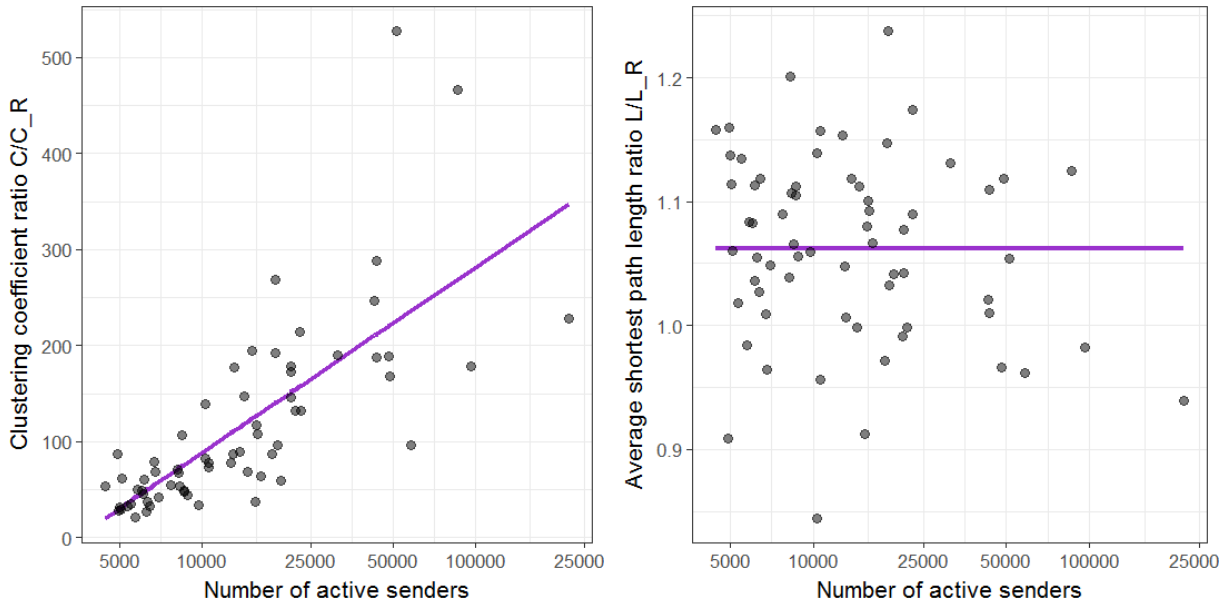

Figure B.1.4: **Clustering coefficient deviates from random as** $\log S$**; deviations of average shortest path length do not vary with size.**

The small world quotient represents the distance of the small world structure of a network from that expected by random. We find that the empirical variation in the small world quotient is explained by variation in the clustering coefficient from random (Figure B.1.5). The numerator compares the clustering coefficient to th $C/C_R = C/(\langle k \rangle/S)$, reveals the pattern remaining by C/S, as we know that the average degree $\langle k \rangle$ does not vary with S. The clustering coefficient ratio (C/C_R) represents almost all of the variation $R^2 = 0.937$ and $R^2 = 0.024$ L/L_R for Q

The denominator reflects that the average shortest path length L varies with log the size of the network (Figure 4.3), and this reflects that the variation about that relationship.

Finally, while we we do not expect the average degree $\langle k \rangle$ to vary with size, and can treat it as a constant in the small world quotient $Q = \frac{C/C_R}{L/L_R} = \frac{C \times S/\langle k \rangle}{L/\log S}$, we can be even more confident that this is not responsible for variation in Q. For graphs of given network size S, the shortest path

length will vary about that as a function of average degree, because higher degree simply creates more ways to shorten any given path (we see this correlation as well in Table B.2.1).

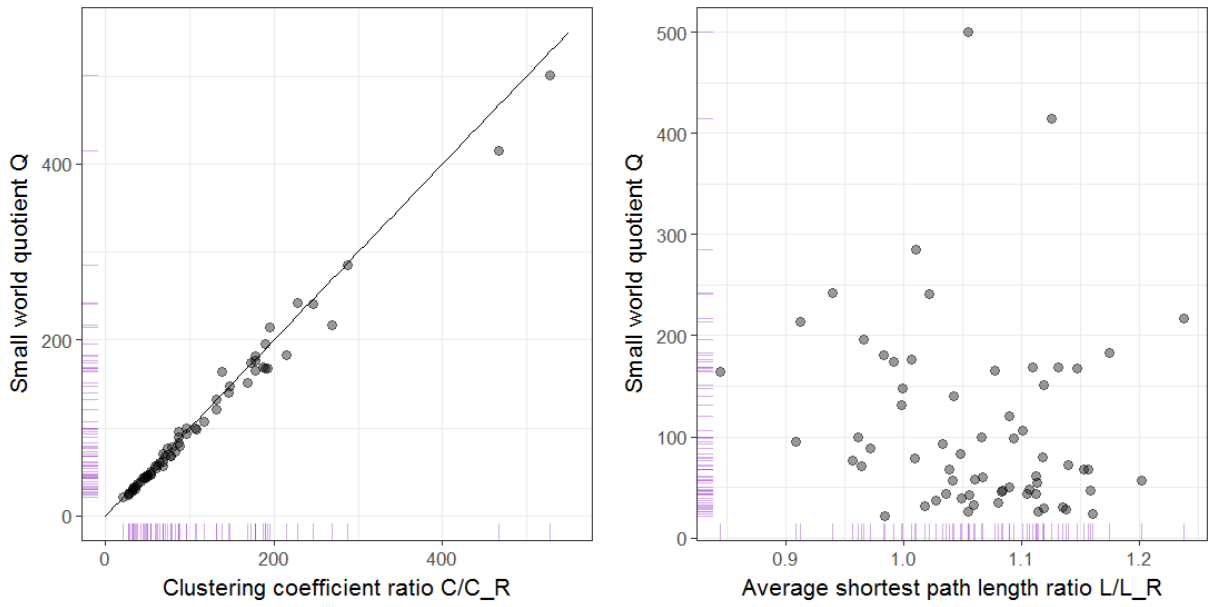

Figure B.1.5: **Variation in the small world quotient is due to the variation in the clustering coefficient.** The small world quotient is defined as $Q = (C/C_R)/(L/L_R)$. Top panel, C/C_R is compared to Q. The identity function is shown for reference. Bottom panel, L/L_R by Q.

B.1.2 Firm age, dispersion & size

Size and dispersion are positively related, but unrelated to age. Geographic dispersion represents the number of physical entities across which these organizations are distributed. Greater dispersion could be due to differentiation of manufacturing plants; dependence on local or regional decentralized processes; distribution of teams over buildings; or a byproduct of acquisitions. First, we find that firm age is unrelated to geographic dispersion (Figure B.1.6). This is non-obvious: by diversifying industries, moving and expanding domain, or through acquisitions, it seems plausible that firms could have become more dispersed with age. We do find that geographic dispersion is related to network size (Figure B.1.7). Specifically, we find that dispersion scales logarithmically with the size of the firm ($R^2 = 0.09$, p value for the model $p = 0.015$). While this is

intuitive—it is unlikely that a firm of 3,000 employees has 1,000 unique locations associated with it—this aligns with our expectation from Blau (1970) that differentiation should increase at a declining rate with the size of the firm. Dispersion in this setting is our best proxy for differentiation across the underlying formal network.

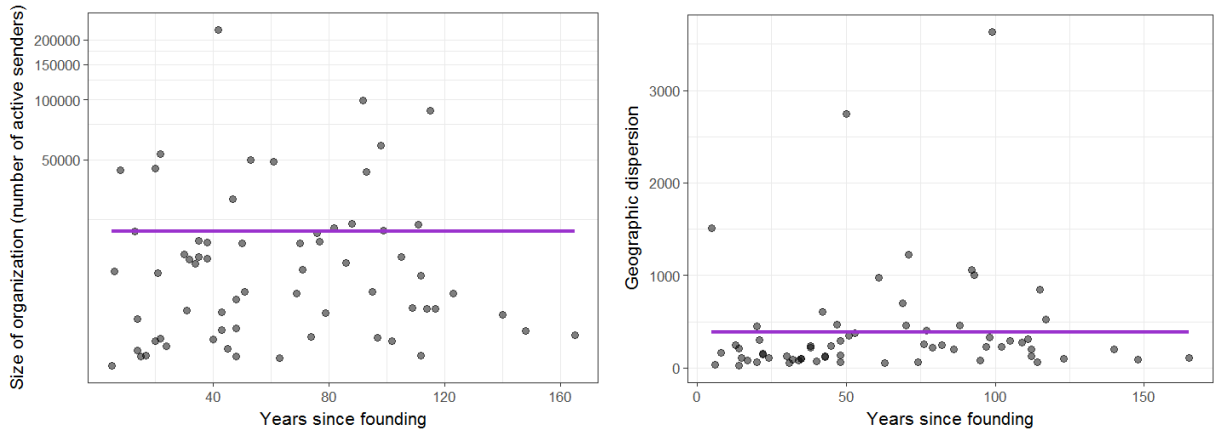

Figure B.1.6: **Firm age is unrelated to firm size or dispersion.** While firms tend to go through mergers, acquistions, and potentially diversify over time, they do not increase their dispersion over time.

We compare firm age to firm size in Figure B.1.6. An implied question—of how age and size are related, or implicitly, how long firms of certain sizes survive—is a demographic question, beyond the scope of this study (Carroll and Hannan (2000)). Here we have a biased subset of all publicly traded firms, limited to those who use a similar resource, Microsoft Exchange, in a comparable manner. Instead, we test for whether the existence of a relationship between firm age and size would carry over to other parts of our analysis. We find no relationship between age and firm size.

Finally, the *rate* of dispersion helps validate which network properties were varying as a function of firm network size—L, C and Q—and show that centralization varies with dispersion, not size (Figure B.1.8).

Considering the effective rate of dispersion, the scaling patterns found for average shortest path length L, clustering coefficient C, and small world quotient Q in Figure 4.6 appear to reflect

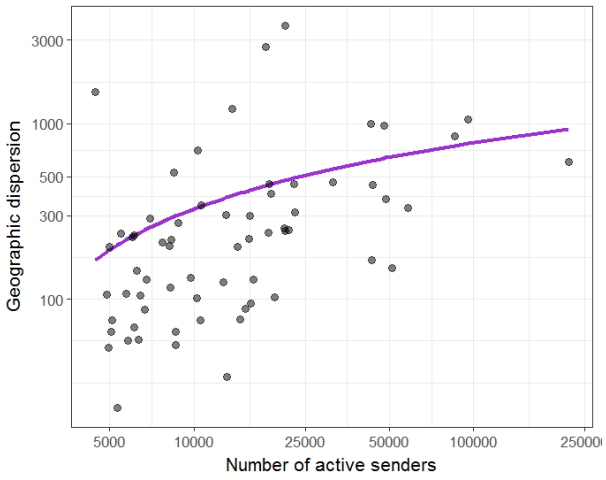

Figure B.1.7: **Geographic dispersion increases at a declining rate with the size of the firm.**

the total size of the firms. This is suggestive that the scaling relationships against centralization are more likely artifacts of size. Average degree remains unrelated to the size, total dispersion, and rate of dispersion. Finally, centralization *does* vary with respect to dipsersion.

Specifically, we find that centralization varies as $\log(d)$ for dispersion d, and varies as $\log(d/S)/(d/S)$ with d/S.

B.1.3 Industry and network structure

Across the network statistics, we find that at least 87–94% of variance *remains* after conditioning on industry: Table 4.2 shows the coefficient of determination R^2 for network measures explained by industry. The lack of significance of our models, combined with the number of parameters used to fit them, suggests that this may be an underestimate of our uncertainty (Table B.1.1).

We regress network statistics over the set of industry categories, introducing dummy variables for each nontrivial inudstry category. We use 64 of the 65 firms, where we exclude the single firm from the Agriculture, Forestry, and Fishing industry category and variance is undefined. (See Table 2.2 for the distribution of industry categories in our data.) We report the results from these

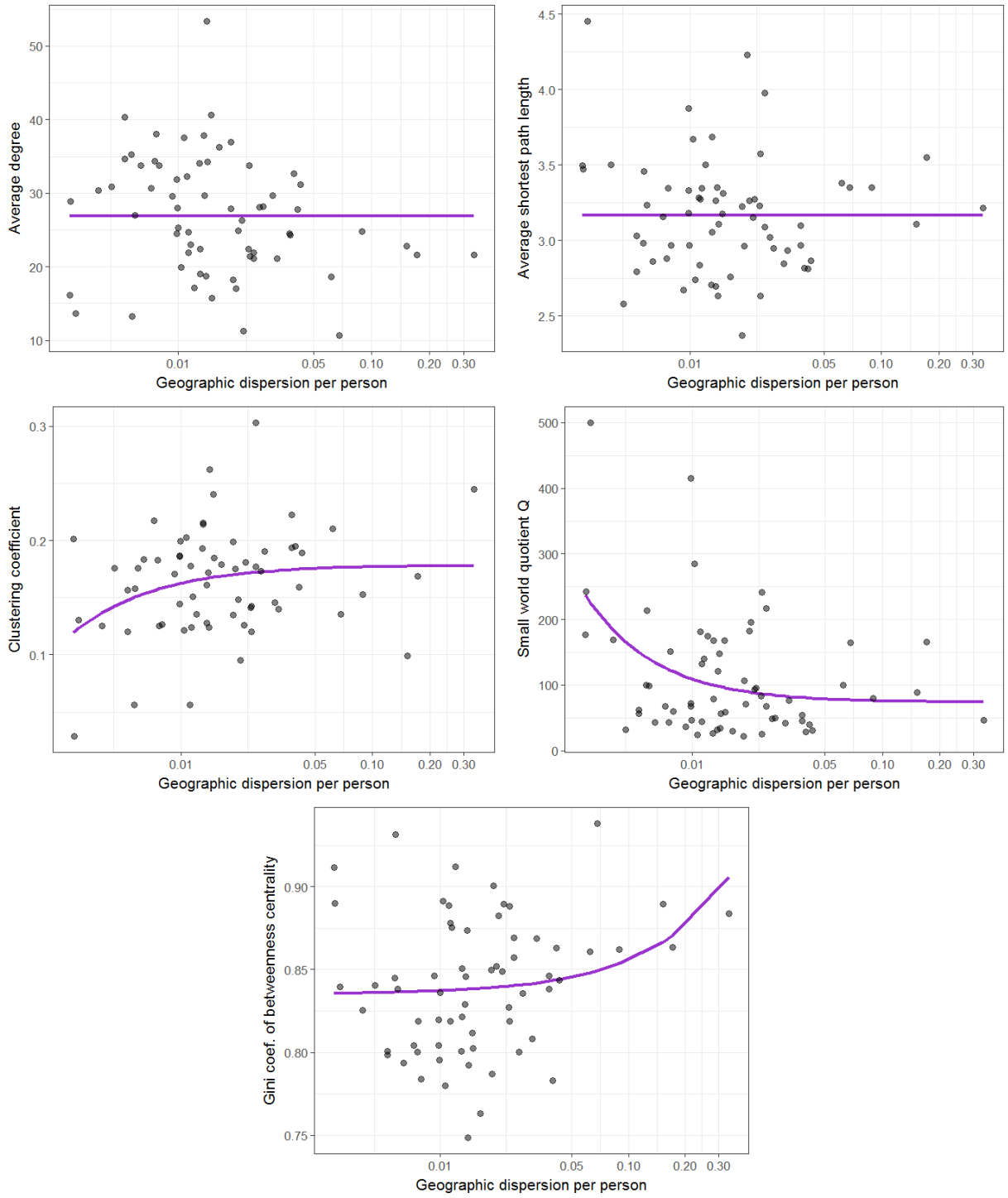

Figure B.1.8: **The rate of dispersion per person helps tease apart the effects of size from dispersion.** $\langle k \rangle$ does not vary with rate of dispersion, but the scaling patterns of L, C, and Q are more likely reflective of network size than dispersion (Figure 4.6), whereas centralization *is* better explained by dispersion.

regressions in Table B.1.1 (this R^2 is also reported in Table 4.2; *N.B.* the adjusted R^2 values suggest a more pessimistic view). We note that this R^2 value can be equivalently derived by measuring variance across groups. Following the discussion in Martin et al. (2016), let F be the fraction of variance remaining to explain network feature Z after conditioning on industry I:

$$F = \frac{\mathbb{E}[\mathrm{Var}(Z|I)]}{\mathrm{Var}(Z)} = \frac{\sum_{\text{Industry } i \in I} \sum_{\text{firm } \alpha \in i:} (\bar{z}_i - z_\alpha)^2}{\sum_\beta (\bar{z} - z_\beta)^2} \tag{B.1}$$

for $\bar{z}_i$ the average network statistic value for all firms from industry i and $\bar{z}$ the average for all firms. For an unbiased model of the network statistics (such as regression, which we use here), $\bar{z}_i = f(i)$, and substituting into Equation B.1, this yields $F = 1 - R^2$.

We find that industry fails to predict each of the network statistics (Table B.1.1). We highlight the *lack* of significant regression coefficients, which suggests that the industry categories are not meaningfully distinguishable. These results are qualitatively similar, even once we account for size (Table B.1.2).

There is one possible exception for an industry effect, for which we have weak-to-moderate evidence. Retail trade is weakly significantly associated with lower average degree ($p = 0.0396$). However, at the full model level, the model fails to reject the intercept model (p value for the F statistic is 0.13). The F statistic, high p value, and the fact that this is drawn from so few observations (3 firms in that category), we do not have meaningful evidence that this category is different. Conditioning on size in our models (Table B.1.2), and using the models of size as established in Section 4.4.1, we have weak evidence for retail trade being associated with lower average degree ($p = 0.042$) and higher average shortest path length ($p = 0.013$) and small world quotient ($p = 0.003$).

The effect disappears for average degree for other definitions of the communication networks ($\tau \geq 0.1,\ \tau \geq 5$: $p > 0.05$), however these results are robust for average shortest path length ($\tau \geq 0.1 :\ p = 0.016,\ \tau \geq 5 :\ p = 0.026$) and small world quotient ($\tau \geq 0.1 :\ p = 0.015,\ \tau \geq 5 :\ p < 0.001$).

A difference in email usage patterns, rather than differences in the informal social networks

themselves, is a plausible mechanism for these differences being limited to the retail sector. Future confirmatory research is necessary to determine if these differences are meaningful.

	$\langle k \rangle$	L	C	Q	Gini (betweenness)
(Intercept)	27.54^{***}	3.24^{***}	0.15^{***}	134.64^{**}	0.84^{***}
	(3.50)	(0.17)	(0.02)	(40.49)	(0.02)
Manufacturing	0.45	−0.13	0.02	−37.72	−0.01
	(3.81)	(0.18)	(0.02)	(44.08)	(0.02)
Retail Trade	-12.04^{*}	0.40	0.03	79.47	0.03
	(5.72)	(0.28)	(0.03)	(66.12)	(0.03)
Services, Technology	−0.02	−0.14	0.00	−35.82	0.01
	(3.99)	(0.19)	(0.02)	(46.06)	(0.02)
Transportation, Communication & Public Utilities	−3.22	0.08	−0.02	−6.85	0.01
	(4.59)	(0.22)	(0.03)	(53.01)	(0.02)
Wholesale Trade	2.99	−0.22	0.03	−53.19	0.00
	(4.95)	(0.24)	(0.03)	(57.26)	(0.03)
Num. obs.	64	64	64	64	64
R^2	0.13	0.12	0.11	0.09	0.06
Adj. R^2	0.06	0.05	0.03	0.01	-0.02
F statistic	1.77	1.60	1.39	1.18	0.77
p-value	0.13	0.17	0.24	0.33	0.58

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table B.1.1: **Regressions: industry fails to predict different network statistics**. Each column indicates the coefficients and performance of the model for each network statistic ($\langle k \rangle$, L, etc.). Note that the F statistic is not significant for all models, i.e., all models fail to reject the null intercept model (where $R^2 = 0$ and best model is the population average).

	$\langle k \rangle$	L	C	Q	Gini (betweenness)
(Intercept)	27.14^{***}	0.54	0.12^{***}	-684.96^{***}	0.83^{***}
	(3.61)	(0.46)	(0.02)	(85.96)	(0.02)
Manufacturing	0.55	-0.02	0.01	-5.56	0.00
	(3.84)	(0.15)	(0.02)	(27.06)	(0.02)
Retail Trade	-11.97^{*}	0.56^{*}	0.01	126.39^{**}	0.03
	(5.76)	(0.22)	(0.03)	(40.58)	(0.03)
Services, Technology	-0.02	-0.06	0.00	-11.10	0.01
	(4.01)	(0.15)	(0.02)	(28.19)	(0.02)
Transportation, Communication & Public Utilities	-3.27	0.05	-0.02	-15.70	0.01
	(4.62)	(0.17)	(0.02)	(32.33)	(0.02)
Wholesale Trade	3.15	-0.11	0.02	-17.85	0.00
	(5.00)	(0.19)	(0.03)	(35.09)	(0.03)
S	0.00				0.00
	(0.00)				(0.00)
$\log S$		0.63^{***}		192.40^{***}	
		(0.10)		(19.33)	
$\log S/S$			102.99^{***}		
			(24.63)		
Num. obs.	64	64	64	64	64
R^2	0.14	0.47	0.32	0.67	0.10
Adj. R^2	0.05	0.41	0.24	0.63	0.00
F statistic	1.49	8.41^{***}	4.40^{**}	19.16^{***}	1.02
p value	0.196	$\ll 0.001$	0.001	$\ll 0.001$	0.423

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table B.1.2: **Regressions for the role of industry on network structure that account for size primarily reflect the role of size.** Each column indicates the coefficients for size and industry category and the performance of the model for each network statistic ($\langle k \rangle$, L, etc.). The models include meaningful size terms, as previously modeled in Section 4.4.1. We find some support for retail trade being predictive of a difference in average shortest path length L and small world quotient Q; the model for average degree is not significant.

B.1.4 Case study: intratypical comparison within manufacturing and technology

Despite our restrictive rules for inclusion in our dataset—U.S.-based firms that publicly NASDAQ or NYSE traded, of at least several thousand employees, and that were active and consistent users of the platform— our dataset consists of a wide array of firm types. Presumably intertypical analysis should be able to uncover patterns that are robust across settings; as Kimberly (1976) explains, "intertypical sampling ... is justified on the grounds that a general theory of organizations ought to enable one to derive hypotheses which can be tested-and, presumably, supported on a heterogeneous sample of organizations." On the other hand, intratypical sampling mitigates still empirically unknown variation in structural characteristics due to organization type.

In the spirit of intratypical analysis, then, we consider a disparate but internally homogeneous collections of firms. We compare two collections of firms with the same four-digit SIC codes: seven firms from a technology subsector and four from the same manufacturing 4-SIC device specialization. We additionally compare another set of four equally similar manufacturing firms to the other set with the same top two SIC levels. Firms within sector may be competing within the same niche, and be pushed to diversify over time (Hannan (2005)); on the other hand, through potentially similar distributions of roles, similar market pressures and contingencies, and legitimation through mimetic and normative processes, firms within the same sector may become more similar over time (DiMaggio and Powell (1983)).

Across different network measures, we find that each collection of firms are each highly heterogeneous in their own right, even once we account for size. Compellingly, these firms are widely distributed within subsector: Figure B.1.9 highlights not only the diversity within each highly specialized type, but also that these distinct specialized groups are difficult to distinguish. While there may be fewer competitive pressures for firms to differentiate across sectors, this suggests evidence for a wide diversity of firm types, both within and across sector.

Management and productivity has been shown to vary widely in firms (Bloom and Van Reenen (2010); Foster et al. (2008), e.g, compared across producers of homogeneous products, such

as corrugated and solid fiber boxes or mixed concrete). While the comparison here is statistically underpowered, there is a rich tradition of small-scale comparative analysis in the organization theory literature (Ahuja et al. (2012), Blau (1965), and Kilduff and Brass (2010)), and the role of organization context has remained under-specified (Hansen and Wernerfelt (1989) and Pugh et al. (1969)) and the degree of organization heterogeneity is both unknown and under debate (Hannan (2005)).

B.2 Appendix: Predicting performance

B.2.1 Regression using informal network structure and organizational productivity

One prediction task we have performed has been using regression to predict organizational productivity in our firms. Here we treat network structural measures as features and attempt to predict different economic outcomes. We fit a nested model with different network measures as features, as described in Section B.2.1.1. We try to generalize outside of this training set to look at other versions of the firm networks. That is, we fit the models using the network structure derived from the reciprocity strength $\tau \geq 1$ networks, and then fit the best model from that exercise to the network features derived from the network of other relationship strengths. For example, if the best model included only industry, average degree, and clustering coefficient based on a given performance outcome Y and the networks $\{G\}_{\tau \geq 1}$, we would fit the model including only industry, average degree, and clustering coefficient to fit Y for the measures derived from $\{G\}_{\tau \geq T}$ for different values of T. If there was systematic variation in the performance across different versions of the network, that would provide insight into what types of informal relationships are most relevant to firm performance.

Data We have 65 networks for publicly traded firms as described in Section 4.3.4, of which we use 62 in the regressions. We exclude two firms for nonrepresentative financial information for 2015 due to mergers and acquisitions. In addition, we exclude the single firm in the Mining/Agriculture/Construction sector, as we are using industry category as a predictor.

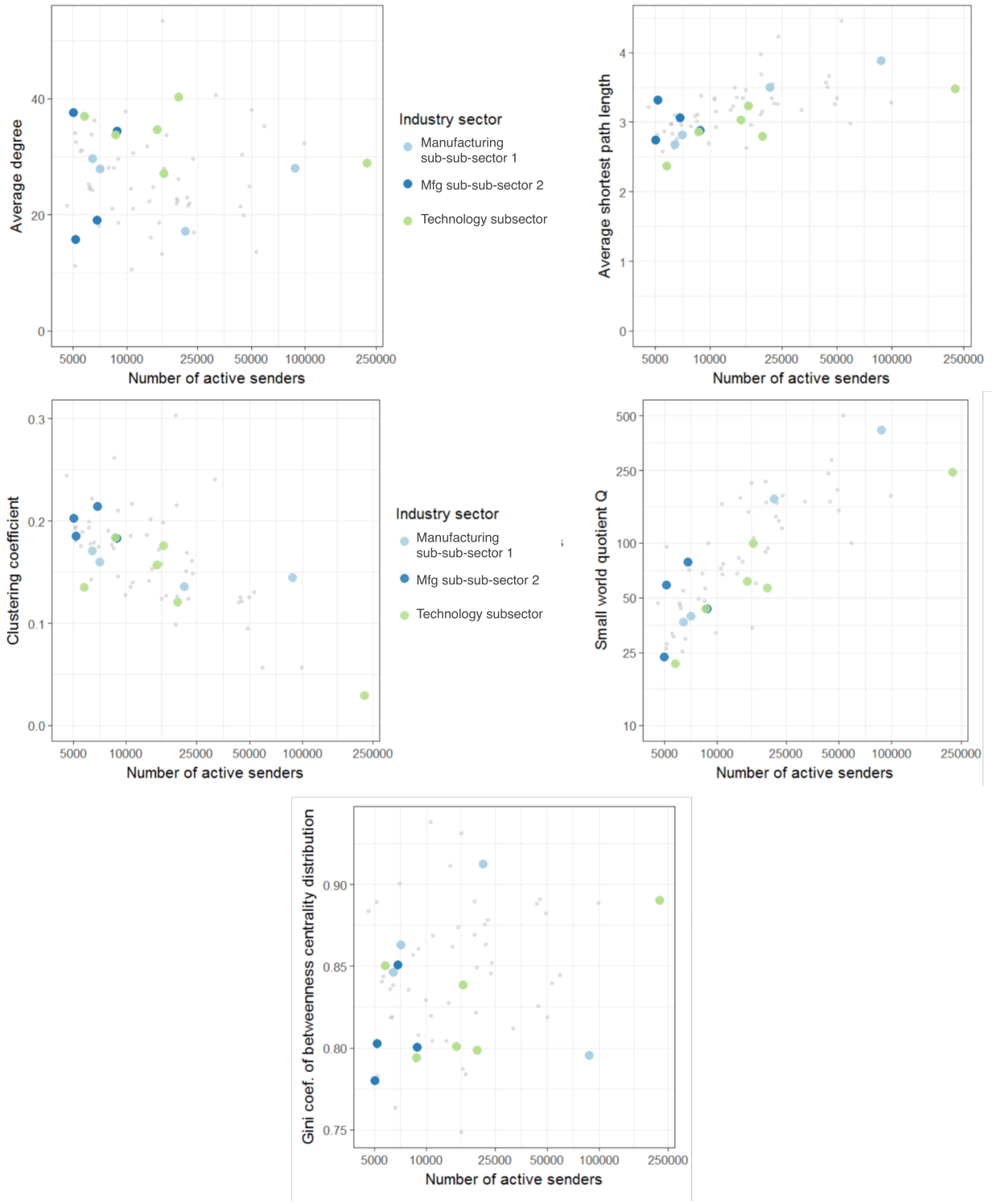

Figure B.1.9: **Comparison of network measures on three sets of firms from similar within-industry domains.** Blue points represent manufacturing firms, with light and dark representing two different sectors: manufacturing of surgical and medical devices and manufacturing of transportation equipment, respectively. Technology firms are represented in green. All other firms are represented as small gray points. Despite the similarity of function of these firms, we find heterogeneity within-type well exceeds heterogeneity across organization types.

B.2.1.1 Models of performance

For each outcome variable (performance measures), we apply models 1–6. All performance measures are derived from MSN Money.

Outcome variables (performance measures):

- Log of sales per employee (Income.Employee, by MSN Money). *N.B.* based on 'Net Income', not Revenue
- Sales (Revenue) Q/Q (last year, growth rate)
- Return on Assets (5 year average)
- Return on Equity (5 year average)
- Average performance rank: rank order of organizations by these measures, then average rank position across measures

Note about constructing the data: we don't have complete data for the regression. We are currently missing 7 values for sales per employee; 2 for Sales Q/Q; 2 for Return on Assets. This is noted on the tables for each set of regressions. We compute the combined performance ranking by:

- For each measure (e.g., Return on Assets), compute rankings. Rank is descending and ties take the average of the ranks they would take: for example, for some four performance scores of [5, 25, 100, 25], assign rankings [4, 2.5, 1, 2.5], here averaging second and third place.
- Given the rankings, for each firm, we take the average over the ranks available. This means some are an average of three rankings, not four. (We do not penalize for missing data in this measure.)

Models:

- **1** Industry + log of size

Industry is given by first level of primary SIC code classification; categorical/binary variables. We exclude the only agriculture/mining company as it is the only firm in that category.

(Size given by total employees, Hoover's)

- **1a** M1 + senders/size (ratio of total senders S / total employees)

  (N is active senders, $\tau = 1$)

- **2** M1a + Average degree ($\langle k \rangle$)

- **3** M1a + Average degree + Clustering (C)

- **4** M1a + Average degree + Clustering + Avg Geodesic (L)

- **5** M1a + Average degree + Clustering + Avg Geodesic + S/W index (Q)

  (Walsh small world index)

- **6** M1a + Average degree + Clustering + Avg Geodesic + S/W index + Centralization (G)

  (Centralization is Gini coefficient of betweenness centrality scores)

B.2.1.2 Results

First, we consider multicollinearity (Table B.2.1). Between the relatedness of our network features and the number of organizations in each industry, our regression results are unlikely to be particularly robust, which we could explore further if we were to find any strong signal in the data.

We apply each model to each of the performance measures.

Notation. We use the standard significance codes: p value is labeled *** 0.001 ,** 0.01, * 0.05, . 0.1 , 1. Adjusted R^2 penalizes R^2 for having a large number of explanatory variables to available data:

$$\text{Adjusted } R^2 = \bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p} = R^2 - (1 - R^2)\frac{p-1}{n-p} \tag{B.2}$$

where p is the number of explanatory variables (not including the constant term), and n is the sample size. We also use AIC, the Akaike Information Criterion, for model selection.

Multicollinearity See Table B.2.1.

We note that log(Size) and average shortest path length are strongly correlated: this makes sense, recalling the scaling results ($L \sim \log(S)$). Controlling for size, average shortest path length is strongly negatively correlated with average degree; this is also intuitive, as more links should make it more likely that shortest paths are shorter.

Average degree $\langle k \rangle$ and centralization G are also strongly (negatively) correlated. As degree increases, the distribution of betweenness centrality scores decreases. This could potentially be understood as greater degree creating more potential short paths throughout the network. Notably the degree distribution also becomes less skewed with increasing average *and median* degree.

As a side note, as the minimum tie strength τ increases, the remaining networks become less centralized: stronger ties are more equally distributed across the network.

As we discovered previously, the small world quotient ($Q = \frac{C/C_R}{L/L_R}$) varies with $\log(S)$, so these variables are also naturally correlated, although this effect goes down with increasing tie strength. Controlling for size, this does then covary with clustering coefficient, but much more weakly ($\langle \rho \rangle = -.233$).

Finally, we note that the relationship between the average shortest path length and size is primarily getting *stronger* with stronger restrictions on the definition of a tie; on the other hand, the relationship between degree and centralization are getting *weaker* with stronger tie definitions.

Regression results We show the results from applying models 1–6 to each of the five performance measures in Tables B.2.2, B.2.3, B.2.4, B.2.5, including the combined performance ranking in Table B.2.6. Almost all models do not have any significant coefficients, but Model 2 (Industry, Employees, Senders/Employee, and Mean Degree) does have some signal towards predicting Income per Employee, as Model 4 (Model 2, plus Clustering Coefficient C and Average Shortest Path Length L) for predicting Return on Equity. We apply these models across different versions of the network (given by minimum reciprocity to define a relationship, τ) in Tables B.2.7

τ	$\rho(\log(S), L)$	$\rho(\langle k \rangle, G)$	$\rho(\log(S), Q)$	$\rho(\langle k \rangle, L/\log(S))$
0.1	0.659	-0.740	0.804	-0.740
1	0.635	-0.722	0.734	-0.670
2	0.662	-0.730	0.705	-0.608
5	0.691	-0.665	0.653	-0.490
10	0.702	-0.569	0.522	-0.400
20	0.745	-0.424	0.073	-0.318

Table B.2.1: **Correlation between network variables across different definitions of the network.** The first three combinations (log size ($\log(S)$) and average shortest path length L; average degree $\langle k \rangle$ and centralization of betweenness centrality G; log size $\log(S)$ and the small world quotient Q) are explicitly included in the model. The latter ($\langle k \rangle$ vs. $L/\log S$) confirms that variation about average shortest path length is negatively related to average degree, and this relationship is weaker in stronger-tie networks.

and B.2.8. Any meaningful effects seem to disappear in this setting as well.

Table B.2.2: **Income per employee.** Log of income per employee. N_{orgs} evaluated: 55 (7 missing values). Model 2 is best by AIC and adjusted R^2.

	β_0	Industry SIC code						Emplys.	Senders/ Emplys.	$\langle k \rangle$	C	L	Q	G	RSS	R^2	Adj. R^2	AIC
		3	4	5	6	7	8											
1	4.018***	-0.003	0.226	-0.011	-0.358	-0.403	-.475	0.116							0.607	0.142	0.014	110.5
1a	3.125**	-0.055	0.052	-0.147	-0.608	-0.666*	-0.557	0.228	0.511*						0.587	0.214	0.077	107.7
2	2.794**	-0.069	0.092	-0.087	-0.571	-0.627*	-0.551	0.222	0.409	0.017 .					0.575	0.262	0.114	106.2
3	1.690	-0.083	0.138	-0.129	-0.580	-0.630*	-0.541	0.373	0.524 .	0.017 .	2.137				0.575	0.278	0.114	107.0
4	1.814	-0.077	0.148	-0.127	-0.579	-0.617*	-0.535	0.253	0.466	0.021	1.630	0.134			0.582	0.279	0.094	108.9
5	2.750	-0.101	0.147	-0.163	-0.597	-0.631*	-0.565	0.195	0.452	0.020	2.078	-0.143	0.001		0.586	0.286	0.082	110.4
6	5.958	-0.134	0.098	-0.152	-0.622	-0.633*	-0.549	0.600	0.624	-0.005	2.842	-0.586	0.002	-3.723	0.586	0.304	0.083	111.0

Table B.2.3: **Revenue growth rate.** Sales (Revenue) Q/Q (last year, growth rate). N_{orgs} evaluated: 60 (2 missing values)

	β_0	Industry SIC code						Emplys.	Senders/ Emplys.	$\langle k \rangle$	C	L	Q	G	RSS	R^2	Adj. R^2	AIC
		3	4	5	6	7	8											
1	37.262 .	2.796	4.955	8.423	-0.190	1.153	5.741	-9.279 .							13.5	0.101	-0.020	492.0
1a	36.183	2.786	4.733	8.348	-0.430	0.909	5.675	-9.150 .	0.606						13.63	0.101	-0.040	494.0
2	36.079	2.782	4.750	8.359	-0.416	0.919	5.677	-9.147 .	0.577	0.005					13.77	0.101	-0.061	496.0
3	64.573 .	2.641	3.350	8.689	-0.301	0.785	5.074	-12.934 *	-2.566	0.001	-56.111				13.74	0.123	-0.056	496.6
4	59.113	2.398	2.927	8.759	-0.527	0.404	4.840	-7.634	-0.103	-0.150	-33.746	-5.892			13.86	0.126	-0.075	498.4
5	93.517 .	1.553	3.018	7.271	-1.545	0.155	3.774	-9.949	-1.275	-0.164	-20.497	-15.452	0.049		13.87	0.142	-0.077	499.2
6	-56.891	3.529	4.806	7.403	-0.926	-0.094	2.894	-28.109	-8.001	0.959	-48.915	4.962	0.039	171.29	13.42	0.215	-0.007	495.9

Table B.2.4: **Return on Assets.** Return on assets (5 year average). N_{orgs} evaluated: 62

	β_0	Industry SIC code						Emplys.	Senders/ Emplys.	$\langle k \rangle$	C	L	Q	G	RSS	R^2	Adj. R^2	AIC
		3	4	5	6	7	8											
1	-12.054	0.747	-2.272	1.581	-2.042	0.721	-0.431	4.403 .							6.929	0.085	-0.034	425.4
1a	-10.308	0.766	-1.939	1.770	-1.624	1.080	-0.332	4.174 .	-0.895						6.988	0.087	-0.051	427.3
2	-10.772	0.750	-1.862	1.848	-1.575	1.121	-0.324	4.173 .	-1.019	0.022					7.052	0.087	-0.071	429.3
3	1.032	0.702	-2.483	2.123	-1.575	0.994	-0.582	2.563	-2.127	0.017	-22.806				7.067	0.101	-0.075	430.3
4	7.956	1.017	-1.947	2.104	-1.450	1.453	-0.284	-4.144	-5.153	0.206	-50.495	7.415			7.065	0.120	-0.075	431.1
5	21.954	0.665	-1.946	1.564	-1.728	1.283	-0.740	-5.069	-5.441	0.199	-44.112	3.388	0.021		7.091	0.131	-0.082	432.2
6	-26.555	1.285	-1.296	1.331	-1.390	1.321	-1.055	-11.381	-8.187 .	0.584	-55.910	10.156	0.019	57.107	7.031	0.163	-0.064	431.9

Table B.2.5: **Return on Equity.** Return on equity (5 year average). N_{orgs} evaluated: 60 (2 missing values). Model 4 is best by AIC, although we have insufficient sample size to have high confidence in these values.

	β_0	Industry SIC code						Emplys.	Senders/ Emplys.	$\langle k \rangle$	C	L	Q	G	RSS	R^2	Adj. R^2	AIC
		3	4	5	6	7	8											
1	-21.493	-1.419	-6.452	-3.835	-10.556	-1.693	-2.521	9.415							17.07	0.076	-0.048	520.2
1a	-5.137	-1.038	-3.458	-2.161	-7.421	1.497	-1.634	7.181	-7.973						17.02	0.010	-0.042	520.6
2	-9.608	-1.303	-2.874	-1.530	-6.989	1.790	-1.585	7.277	-8.876	0.184					17.12	0.107	-0.054	522.2
3	6.924	-1.319	-3.712	-1.149	-7.008	1.631	-1.934	5.018	-10.443	0.176	-31.426				17.25	0.111	-0.070	523.9
4	43.383	0.694	-1.070	-1.232	-5.659	3.954	-0.458	-28.285	-25.610*	1.085*	-169.474 .	36.470*			16.71	0.183	-0.004	520.8
5	76.162	-0.427	-1.113	-2.557	-6.924	3.467	-1.615	-29.644	-25.866*	1.052 .	-149.520	25.642	0.051		16.78	0.194	-0.012	522.0
6	44.435	-0.203	-0.736	-2.761	-6.809	3.421	-1.883	-33.334	-27.398*	1.298	-154.553	29.300	0.051	37.243	16.94	0.196	-0.031	523.9

Table B.2.6: **Combined performance rank.** N_{orgs} evaluated: 62

	β_0	Industry SIC code						Emplys.	Senders/ Emplys.	$\langle k \rangle$	C	L	Q	G	RSS	R^2	Adj. R^2	AIC
		3	4	5	6	7	8											
1	49.43*	1.125	2.386	-0.752	8.871	3.814	4.772	-5.043							13.24	0.064	-0.058	505.8
1a	54.74*	1.182	3.398	-0.177	10.143	4.906	5.072	-5.738	-2.721						13.34	0.068	-0.072	507.5
2	59.73**	1.356	2.565	-1.019	9.613	4.462	4.985	-5.720	-1.381	-0.237					13.32	0.088	-0.070	508.1
3	42.98	1.422	3.446	-1.410	9.613	4.642	5.351	-3.435	0.191	-0.230	32.380				13.4	0.096	-0.082	509.6
4	22.34	0.485	1.850	-1.353	9.239	3.276	4.462	16.554	9.212	-0.793 .	114.89	-22.098			13.19	0.140	-0.049	508.5
5	-13.353	1.382	1.847	0.025	9.949	3.708	5.626	18.912	9.943	-0.774 .	98.618	-11.830	-0.053		13.16	0.016	-0.044	508.9
6	92.97	0.023	0.423	0.536	9.209	3.624	6.316	32.747*	15.964 .	-1.619*	124.48 .	-26.663	-0.048	-125.166	12.95	0.204	-0.011	507.7

Table B.2.7: **Income per employee, Model 2 across different values of τ.** N_{orgs} evaluated: 55

τ	β_0	Industry SIC code						Emplys.	Senders/ Emplys.	$\langle k \rangle$	C	L	Q	G	RSS	R^2	Adj. R^2	AIC
		3	4	5	6	7	8											
0.1	2.949**	-0.079	0.098	-0.063	-0.552	-0.628*	-0.524	0.186	0.338	0.013*					0.566	0.286	0.143	104.4
1	2.794**	-0.069	0.092	-0.087	-0.571	-0.627*	-0.551	0.222	0.409	0.017 .					0.575	0.262	0.114	106.2
2	2.714**	-0.068	0.106	-0.088	-0.577	-0.635*	-0.554	0.241	0.443 .	0.023					0.577	0.256	0.107	106.6
5	2.690*	-0.066	0.113	-0.103	-0.594	-0.653*	-0.562	0.254	0.483 .	0.035					0.581	0.246	0.095	107.4
10	2.719*	-0.062	0.109	-0.118	-0.609	-0.666*	-0.570	0.257	0.506*	0.049					0.585	0.237	0.085	108.0
20	2.779*	-0.057	0.100	-0.131	-0.622	-0.676*	-0.575	0.256	0.519*	0.067					0.588	0.229	0.075	108.6

Table B.2.8: **Return on Equity, Model 4 across different values of τ.** N_{orgs} evaluated: 60 (2 missing values)

τ	β_0	Industry SIC code						Emplys.	Senders/ Emplys.	$\langle k \rangle$	C	L	Q	G	RSS	R^2	Adj. R^2	AIC
		3	4	5	6	7	8											
0.1	36.025	0.614	-0.572	-1.451	-4.992	3.807	-0.837	-28.857	-25.556*	0.758 .	-134.157	40.231 .			16.86	0.168	-0.022	521.9
1	43.383	0.694	-1.070	-1.232	-5.659	3.954	-0.458	-28.285	-25.610*	1.085*	-169.474 .	36.470*			16.71	0.183	-0.004	520.8
2	42.254	-0.457	-1.066	-1.151	-5.966	3.417	-0.081	-20.096	-21.804*	1.099	-147.975	24.984			17.01	0.153	-0.041	523.0
5	18.564	-1.467	-2.498	-2.286	-6.793	2.274	-0.703	-1.086	-12.662	0.450	-64.615	5.908			17.45	0.109	-0.095	526.0
10	-4.742	-0.871	-4.135	-2.382	-7.648	1.281	-1.693	7.502	-7.487	-0.450	10.745	-0.367			17.52	0.102	-0.104	526.5
20	-22.760	0.627	-5.388	-1.279	-8.409	0.202	-1.333	9.759	-5.670	-2.121	103.452	-1.901			17.25	0.130	-0.070	524.6

B.2.2 Random forests using informal network structure to predict organizational productivity

These tests, currently in preliminary form, will be included in full in the draft submitted for publication.

B.3 Appendix: Robustness of the results

Here we have pursued exploratory, not confirmatory, analysis with a number of researcher degrees of freedom (Hofman et al. (2017)), but structure potentially shifts meaningfully with choice of network definition τ (De Choudhury et al. (2010) and Hofman et al. (2017), Chapter 5).

These tests, currently in preliminary form, will be included in full in the draft submitted for publication.

Appendix C

Appendix: Empirical network construction: computational perspectives on weak ties, stability, and densification

C.1 Appendix: Empirical network construction

C.1.1 Further results on weak ties

Expanding on weak ties: embedded ties are stronger than bridges Expanding on the result that neighborhood overlap is lowest on weak ties (Figure 5.6, Section 5.4.1), we show that embedded ties (those with shared neighbors) can be weak *or* strong, but that bridges (those without) are only weak. This is a subtly different setting than the previous result. Range characterizes how far apart a pair would be if the edge between them was removed. For pairs that share mutual contacts (Jaccard coefficient nonzero), their range is necessarily two: there exists a path of length two through their mutual contacts. We consider pairs that do not share mutual contacts (Jaccard coefficient zero) to be bridges, as the relationship between them brings together two otherwise distinct parts of the graph. More precisely, range r_{ij} is calculated as the distance between neighbors i and j if that edge was deleted (so by definition, it has minimum 2, which happens if i and j are in a triangle together).

Individuals in social networks, particularly organizations, that participate in these structural bridges are understood to be brokers: individuals for whom information must pass through, who can intermediate between different parts of an organization, and who have access to unique opportunities across networks. Brokers are understood to enjoy better outcomes within organizations and in general social systems (e.g., Burt 1992, 2004, 2010). On the other hand, ties that are embedded in

the same local networks share common connections, access to information, and are more likely to share strong ties.

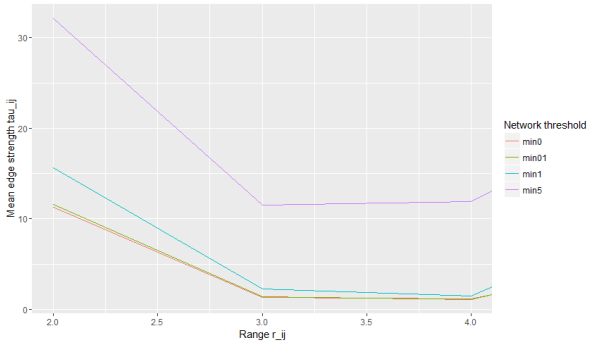

Figure C.1.1: Conditional on an edge existing, compare the range r_{ij} of the edge (i.e., the distance between those neighbors, if that edge was deleted) to the average edge strength $\langle \tau_{ij} \rangle$, over different values of τ (Bottom to top, $\tau_{\min} = 0, 0.1, 1, 5$). Taken over the six month aggregate network ($w = T = 6$ months).

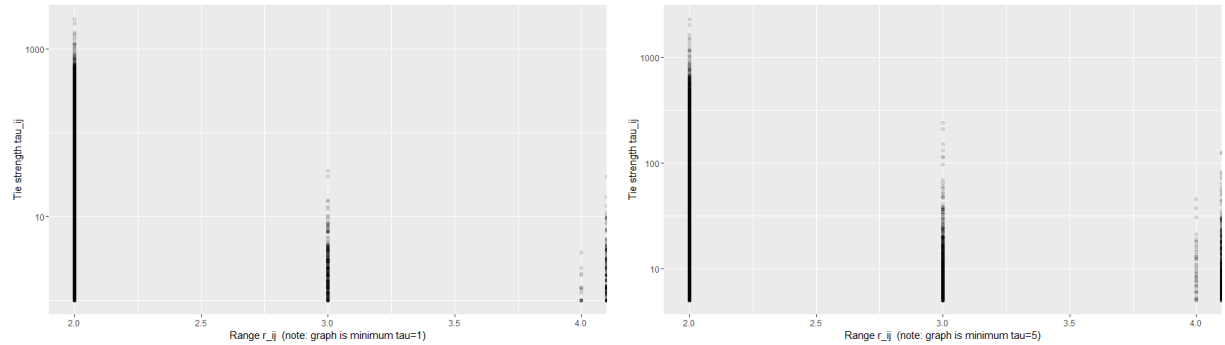

Figure C.1.2: While embedded ties can be weak or strong, bridges are always weak(er). Conditional on an edge existing, compare the range r_{ij} of the edge (i.e., the distance between those neighbors, if that edge was deleted) to the average edge strength $\langle \tau_{ij} \rangle$. Taken over the six month aggregate network for $\tau \geq 1$ and $\tau \geq 5$ ($w = 6$ months).

Figure C.1.1 compares the range of a pair (i, j) to the average tie strength τ_{ij}. The average is taken over all observed pairs with that range, and we compare several network definitions ($\tau = 0, 0.1, 1, 5$). Figure C.1.1 shows that embedded dyads, i.e., pairs that take part in at least one triangle ($r_{ij} = 2$), have higher relationship strength than otherwise-disconnected pairs. Pairs that have range greater than two, i.e., bridges, have comparably weak relationship strength. Figure C.1.2

explicitly plots the distribution of tie strengths for each edge. Figure C.1.2 shows the distribution of points from which Figure C.1.1 is generated for the $\tau = 1$ and $\tau = 5$ networks (top and bottom). We note that almost all observed edges are in a triangle, and therefore have range two.

Granovetter 1973 argued that if three individuals share two strong connections, the third will likely be closed but will be weak or strong. (The probability that this link is formed, if it does not already exist, is a related question that we do not test here, but see, e.g., Kossinets and Watts 2006; Ugander et al. 2013.) That is, triads will be closed with a tie of any strength but that "no strong tie is a bridge" (Granovetter 1973). We find evidence that supports this: ties within a triad may be weak or strong, but almost all bridges are weak. This further supports recent work that ties are infrequent and local and infrequent and distant (Quintane and Carnabuci 2016).

C.1.2 Further results on network stability

Figure C.1.3 shows the fraction of the most-central nodes (top 10% for each of the centrality measures) that are remain most central in future months. We demonstrate the short half-life of the most central nodes using a small handful of diverse organizations.

In contrast, Figures C.1.4 and C.1.5 show how many of the initially central nodes (again, the top 10%) are in the observed snapshot by month and by week, respectively. That is, these figures show what percentage of nodes in each snapshot are in the top 10% were also originally in the top 10%, regardless of whether they ever left.

These two types of figures show something qualitatively different. The first two considered, Figures C.1.3 and 5.7 show the "half-life" of the most central nodes. The second two show whether, and how often, the most central nodes become most central again.

C.1.3 Densification

Following our result that average degree does not vary with size, we also note (but do not pursue further) that we do find that distances increase by $O(\log S)$ (Figure C.1.6). Furthermore, we find that these results are robust across network definition (reciprocity strength τ). This is well

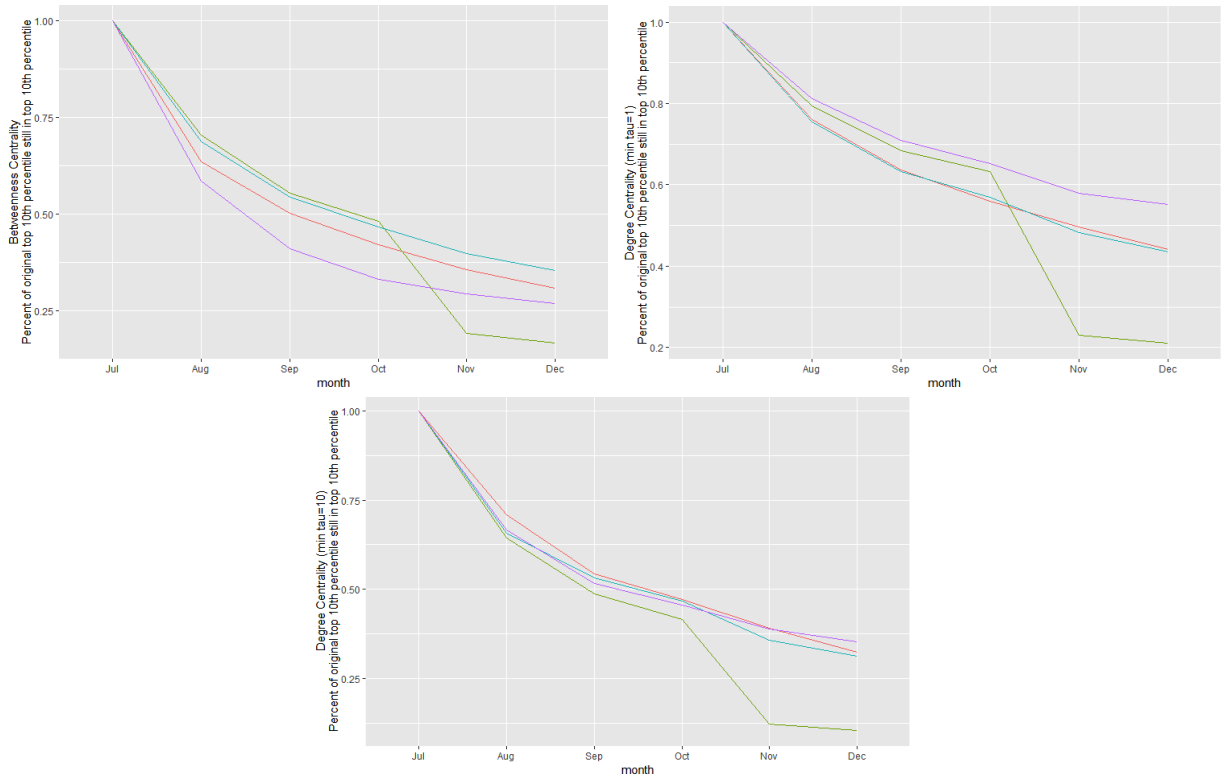

Figure C.1.3: Stability of highest-betweenness and highest-degree individuals over time, per month. Each line represents a different organization (only four are shown—preliminary analysis), connecting observations taken for each month ($w = 1$ month, $\delta = 1$ month). For each network snapshot, we compare the top 10% of individuals by betweenness ($\tau = 1$), degree ($\tau = 1$) and degree ($\tau = 10$) to those who were in the top 10% in the first snapshot. The y-axis plots the percentage that have *remained* in the top 10% since the initial observation.

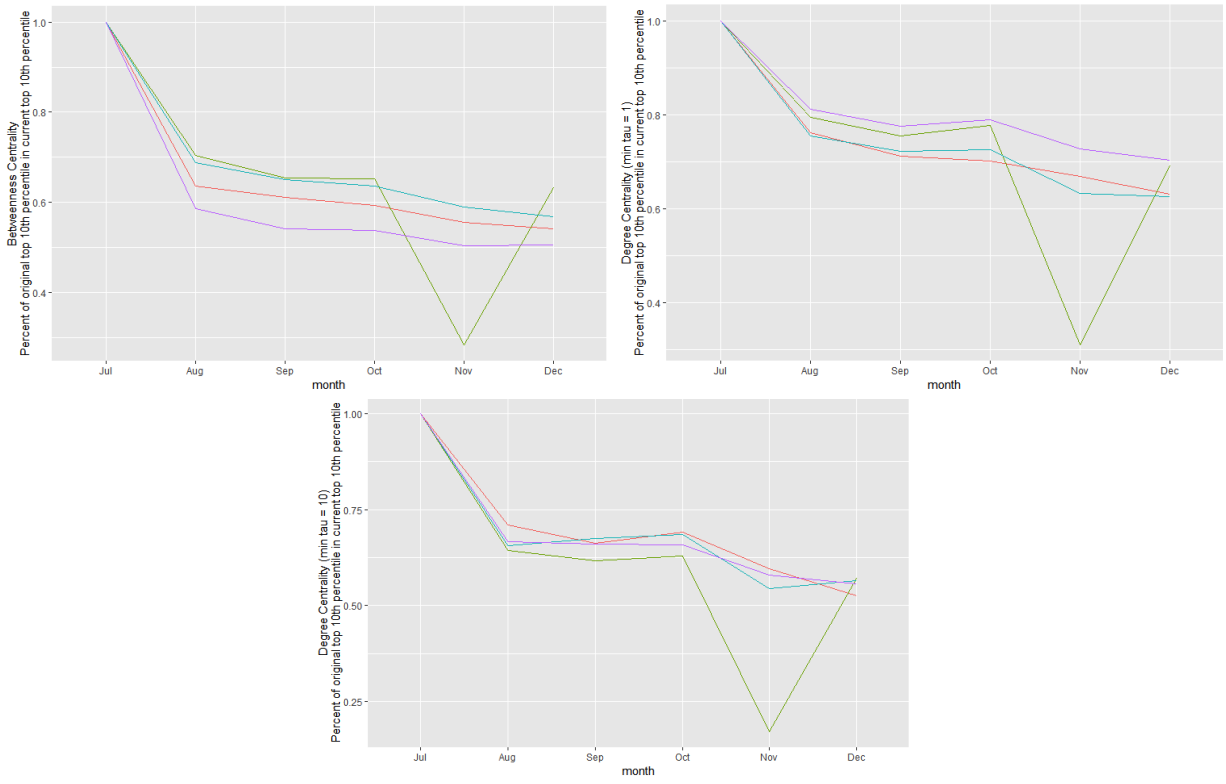

Figure C.1.4: Consistency of highest-betweenness and highest-degree individuals over time, per month. Each line represents a different organization (only four are shown—preliminary analysis), connecting observations taken for each month ($w = 1$ month, $\delta = 1$ month). For each network snapshot, we compare the top 10% of individuals by betweenness ($\tau = 1$), degree ($\tau = 1$) and degree ($\tau = 10$) to those who were in the top 10% in the first snapshot. The y-axis plots the percentage that were *originally* in the top 10% in the initial observation *and* are in the top 10% for the observed snapshot.)

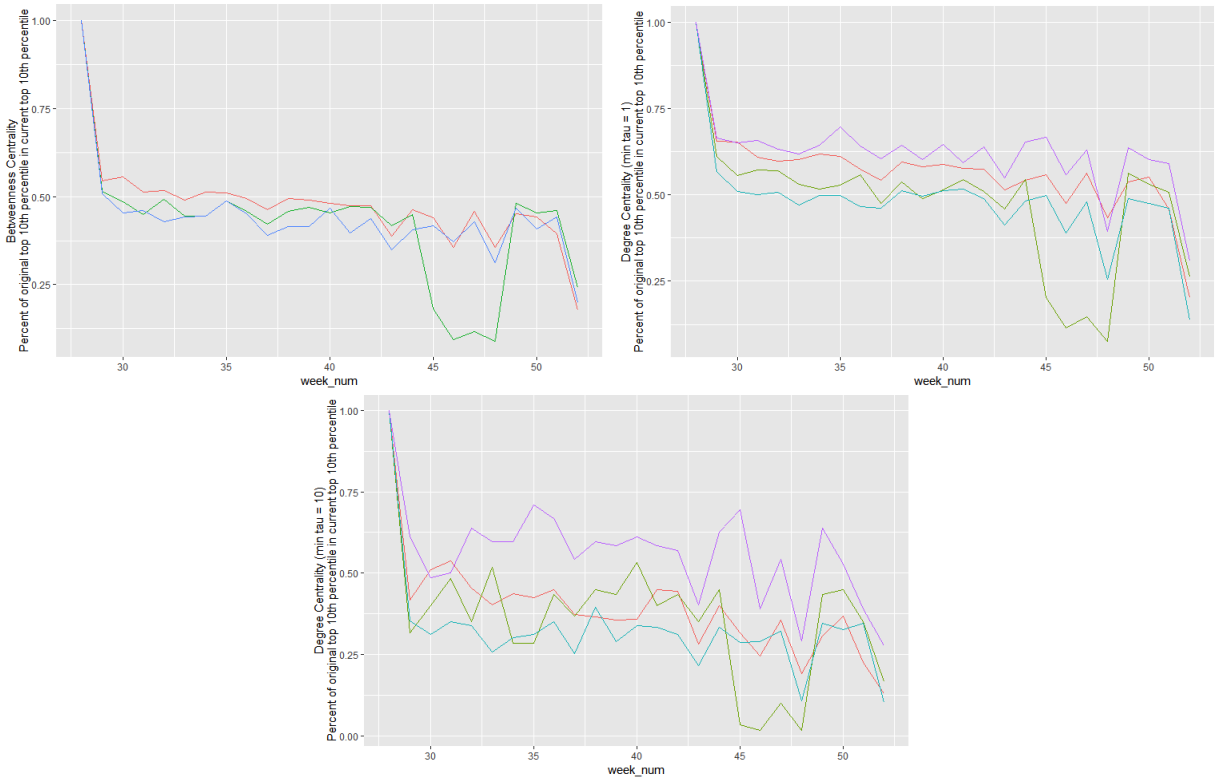

Figure C.1.5: Consistency of highest-betweenness and highest-degree individuals over time, per week. Each line represents a different organization (only four are shown—preliminary analysis), connecting observations taken for each week ($w = 1$ week, $\delta = 1$ week). For each network snapshot, we compare the top 10% of individuals by betweenness ($\tau = 1$), degree ($\tau = 1$) and degree ($\tau = 10$) to those who were in the top 10% in the first snapshot. The y-axis plots the percentage that were *originally* in the top 10% in the initial observation *and* are in the top 10% for the observed snapshot.

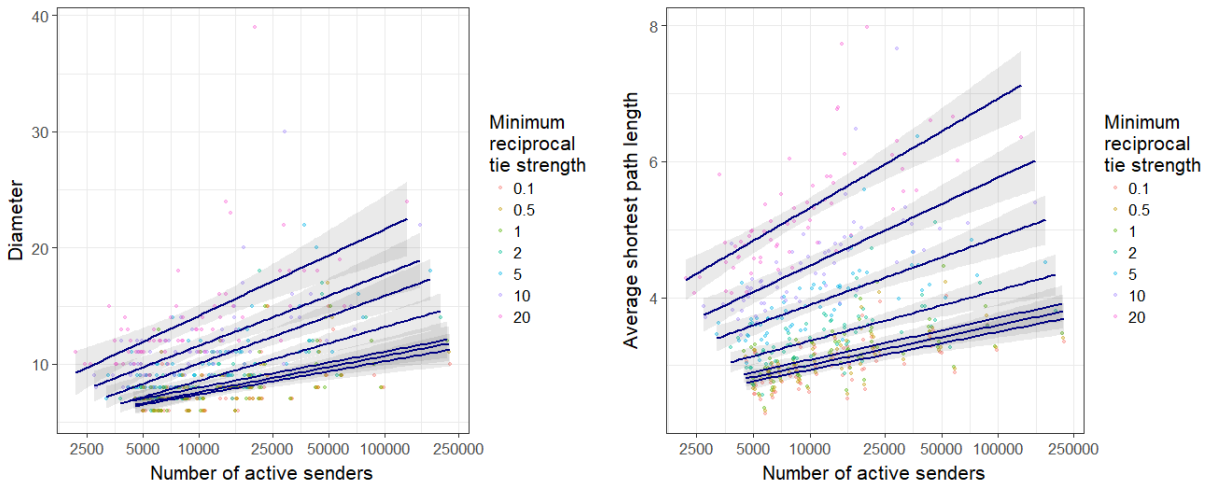

Figure C.1.6: Top, diameter across all networks. The diameter is the length of the longest shortest path between two senders compared to the size of the network. Bottom, the average shortest path between two senders in an organization, across all networks. The lines shows the function of best fit—here, growing (not shrinking) as $O(\log S)$. This matches many models from random graph theory. The model of best fit was chosen by AIC. For the average shortest path length, $O(\log \log S)$ cannot be rejected either.

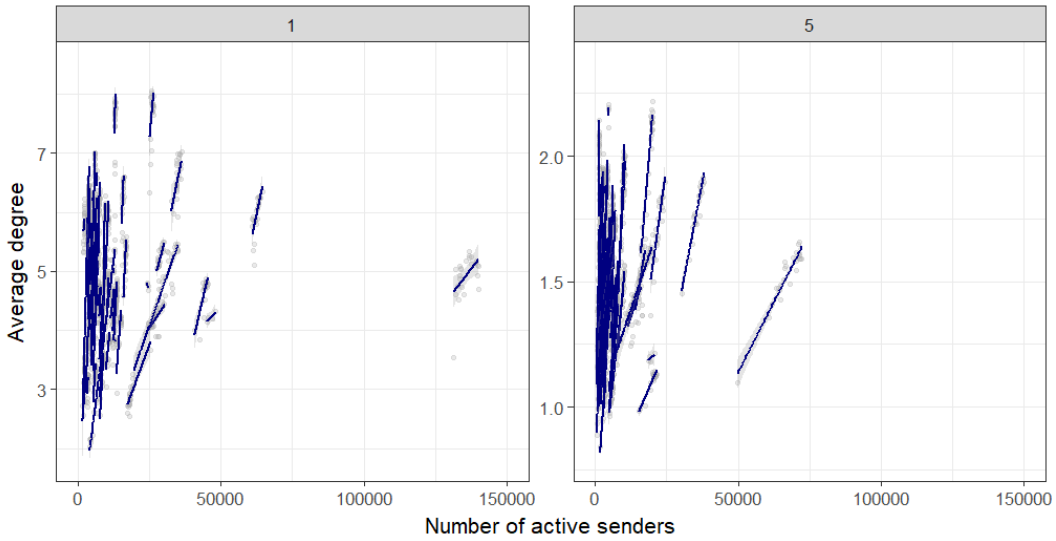

Figure C.1.7: **Average degree (number of contacts) within an organization increases with the number of observed senders.** Observations are taken across $w = 1$ week periods over $T = 6$ months, minimum $\tau = 1$, left, and $\tau = 5$, right. We use the 24 week periods for which we have complete coverage. This figure contains the same data as Figure 5.10 but for more variables and over all organizations.

aligned with many models from random graph theory, as well as previous empirical work (Backstrom et al. 2012; Jacobs et al. 2015a), but contrasts the implications of the densification literature if comparing across comparable networks.

Further details on densification

Figure C.1.7 shows average degree increasing with the number of active senders observed, across all organizations, time windows of one week, and minimum reciprocal tie strengths of 1 and 5.

Table C.1.1 shows the relationship between the observed size of a network and the average degree, under a hierarchical linear model for networks constructed with $\tau_{\min} = 1$ and 5, for $w = 1$ month. Here the data are more impoverished—we only have $T = 6$ months and therefore six observations for each organization—and we find mixed results. For $\tau_{\min} = 5$, we find results analogous to the weekly level. For $\tau_{\min} = 1$, we do not find evidence for a relationship between size and average degree. The former, $\tau_{\min} = 5$, $w = 1$ month, could be more similar to the

$\tau_{\min} = 1,\ w = 1$ week, if this reflects relationships that are weak but consistent across weeks. However, most relationships are weak (Figure 5.3), and it is not clear why this might vary across network definitions. Expanding our analysis of user activity to the monthly level could provide insight into this difference, but we leave that to future work.

Figure C.1.8 shows the variation of average degree and the number of active senders to the week of observation. Although we do observe variation about holidays, we do not find that degree or the number of users are varying meaningfully across these 24 weeks.

Figure C.1.9 compares the amount of messages sent within organization, by raw count, to the number of active senders observed. Here, a message counts as a single message, regardless of the number of recipients, and the number of active senders is divided by the total number of unique senders ever observed, such that the numbers are comparable across organizations. As this is a simple rescaling, the pattern that the total and median number of emails sent (the latter, per active user) increases with the number of active senders is robust to whether this is raw number or fraction of active senders.

	Model: $\langle k \rangle$ under $\tau_{\min} = 1$ $w = 1$, $T = 6$ month snapshots	Model: $\langle k \rangle$ under $\tau_{\min} = 5$ $w = 1$, $T = 6$ month snapshots
(Intercept)	11.54^{***}	5.78^{***}
	(0.51)	(0.54)
$S_{rescaled}$	2.42	5.23^{***}
	(1.29)	(1.13)
AIC	1195.20	370.11
BIC	1219.00	393.91
Log Likelihood	-591.60	-179.06
Num. obs.	390	390
Num. groups: Symbol	65	65
Var: Symbol (Intercept)	8.15	12.30
Var: Symbol $S_{rescaled}$	31.58	48.21
Cov: Symbol (Intercept) $S_{rescaled}$	4.05	22.60
Var: Residual	0.52	0.04

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table C.1.1: Hierarchical linear model comparing degree and observed network size for different minimum levels of τ for $w = 1$ month networks.

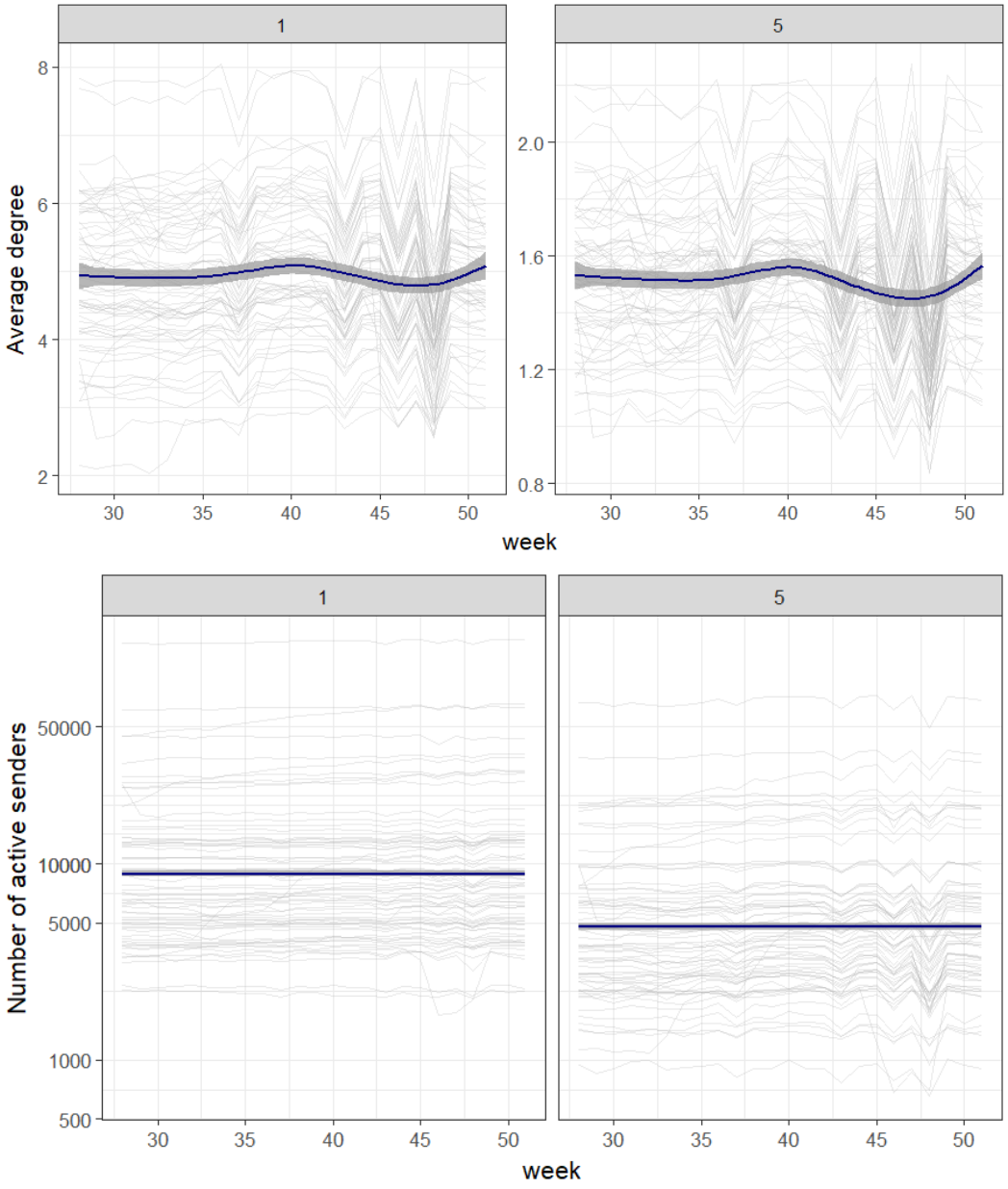

Figure C.1.8: **Top, average degree (number of contacts) within an organization across time. Bottom, number of unique active senders within an organization across time.** Observations are taken across $w = 1$ week periods over $T = 6$ months, minimum $\tau = 1$, left, and $\tau = 5$, right. We use the 24 week periods for which we have complete coverage. The dip in the later weeks reflects Thanksgiving and holiday breaks, but we otherwise do not find a meaningful variation with time.

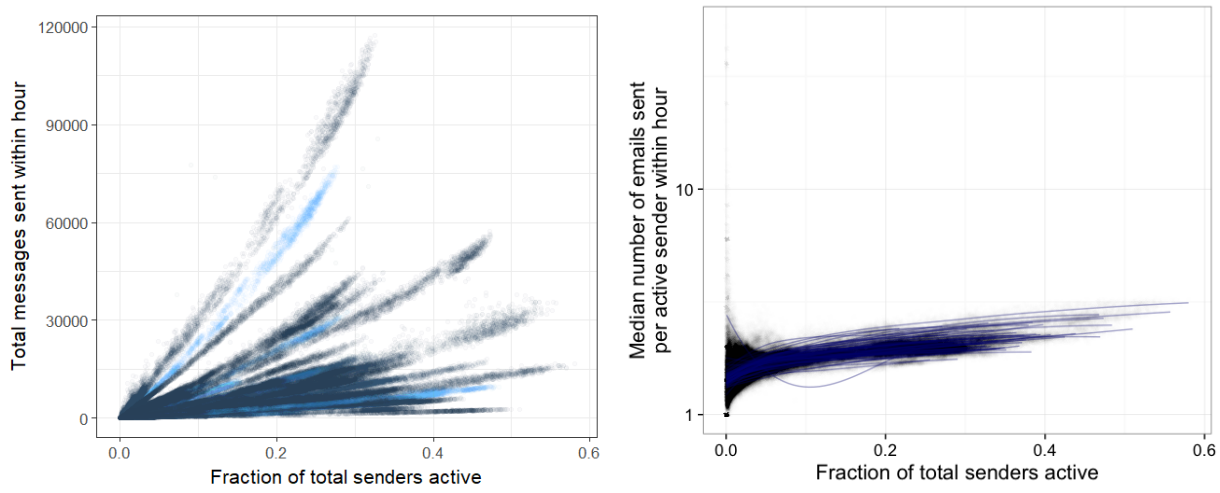

Figure C.1.9: **At the hour level, senders send more messages when more other people are active.** Top, the total number of messages sent in an hour period increases with the fraction of active senders within an organization. Bottom, conditional on a sender being active, the median number of messages sent in an hour period *per active hour user* increases with the fraction of active senders within an organization. Each point represents an observation of the median number of messages sent for a given hour ($w = 1$ hour). The fraction of active senders is given by S_{observed} divided by the total number of unique active senders ever observed ($T = 6$ months).