NORTHWESTERN UNIVERSITY

The Formation and Growth of Collaborative Online Organizations

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Media, Technology, and Society

By

Jeremy Foote

EVANSTON, ILLINOIS

December 2019

Abstract

The Formation and Growth of Collaborative Online Organizations

Jeremy Foote

Collaborative online organizations have emerged as a new means of production, characterized by self-organizing volunteers making contributions to information-based public goods. There has been a large amount of research into understanding how large-scale collaborative online organizations like Wikipedia, Linux, and OpenStreetMap work, and we are starting to understand them. However, thousands of new attempts at volunteer collaboration are created every day, with most never gaining more than a handful of contributors. I argue for an approach that studies these new, small organizations based on "open systems" theories, which treat organizational outcomes as the result of individual-level decisions. From this perspective, environments shape and are shaped by the decisions people make. I draw from literature on related contexts like social movements, work groups, and voluntary organizations to theorize about how technological affordances and the state of the system influence why people create and participate in early-stage groups. I describe the four projects which make up this dissertation, and which draw on this approach. Taken together, these projects provide evidence that decisions to participate in or found an online organization are influenced by past experiences

and by the environment of existing organizations. They also find that social structures are less important for predicting behavior than previous theory would predict. Finally, they suggest the importance of low participation costs and switching costs in understanding both individual decisions and higher-level outcomes. I conclude with a discussion of the implications of these findings and an explanation of opportunities for future work.

Acknowledgements

I feel blessed and lucky to have been surrounded by such a brilliant and generous group of instructors, collaborators, and fellow students. I am particularly thankful for my advisor, Dr. Aaron Shaw. From my first quarter at Northwestern, he has worked to help me to do the research that I wanted to do. He has been generous and talented at identifying and helping me to overcome roadblocks along the way, and has pushed me to think broadly and carefully. He has been supportive and insightful in helping me to hone my skills as a researcher, a writer, and a member of the academic community.

Also deserving particular thanks is Dr. Mako Hill. With Aaron, he founded and leads the Community Data Science Collective. Aaron and Mako have created a wonderful lab culture that is inclusive and supportive. Mako has also been a wonderful collaborator and has often helped me to recognize and challenge my assumptions.

Indeed, every project of this dissertation is collaborative and I have found a wonderful set of collaborators. Drs. Darren Gergle and Noshir Contractor have given invaluable direction and feedback for individual projects and for the direction of this dissertation as a whole. While many members of the Community Data Science Collective have provided feedback for various pieces of this dissertation, I owe special thanks to Nathan TeBlunthuis and Sneha Narayan, each of whom helped me to think through computational, statistical, and theoretical challenges throughout this project.

I'm also grateful to Northwestern University for their generous funding and support; pieces of this work were also funded by grants from the NSF (IIS-1617468 and IIS-1617129) and the Army Research Office (W911NF-14-10686). I'm also thankful to Mako Hill, Trevor Bolliger, and Jason Baumgartner for providing Wikia and reddit data and for the many people who contributed to the open source software used in this dissertation, such as R, Python, pandas, and ggplot2.

Finally, I'm so thankful for the support of my family. My parents taught me to be curious and have been endlessly encouraging. My best and last expression of gratitude is for my wife, Kedra Foote. She has been so supportive of my goals and my research and has done so much to enable me to do this work. She knows me better than anyone but believes in me anyway. She and our kids—Rebecca, William, Owen, and Andrew—have kept me grounded and focused on what is most important.

Dedication

For Kedra, Rebecca, William, Owen, and Andrew.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

At a time when the Internet is most often viewed as a source of bigotry, polarization, and antagonism, cooperative online organizations provide some reason for hope. Projects like Wikipedia and Linux continue to motivate hundreds of thousands of people to contribute to public goods without financial remuneration. Their outputs are products which provide enormous benefit to the world; understanding how they work and how to increase the scope of public goods production is an important social problem.

Most of the research on collaborative online organizations has focused on large, successful projects: how people are motivated, how they self-organize, and whether they really produce high-quality outputs (Benkler, Shaw, & Hill, 2015). Among many other important findings, this research has shown that these organizations are characterized by low-cost contributions and fluid boundaries.

These conditions lead to a dynamic system, where organizations grow and shrink as people move between them. While we are starting to get a sense of how large individual collaborative online organizations work, we know very little about how the population of organizations works as a whole. We have only a limited understanding of basic questions such as: Why are there so many small collaborative online organizations? Why do some organizations succeed and grow while most do not?

The limited attention to these population-level questions has primarily focused on predicting growth and survival of individual communities. For example, researchers have looked

at the features of an organization and its members (Cunha, Jurgens, Tan, & Romero, 2019; Kairam, Wang, & Leskovec, 2012; Schweik & English, 2012) as well as the features of the environmental conditions in which an organization exists (TeBlunthuis, Shaw, & Hill, 2017; H. Zhu, Kraut, & Kittur, 2014) to predict individual community outcomes.

In this set of empirical projects, I take a distinct approach by considering organizations and populations of organizations as the result of individual decisions. In this dissertation I ask and begin to answer: In a complex social and communication environment with low barriers to entry and exit, how do people decide whether and where to allocate their efforts?

For most of these papers, I focus on the early stages of an organization: a time when uncertainty is greatest and decisions among organizations are most salient. In this first chapter, I make a longer argument for studying individual decisions, summarize the research on collaborative online organizations, and discuss the importance of studying the early stages of organizations explicitly. I argue that online platforms and computational methods provide us with a unique opportunity to study decision-making and "systems" of public goods. I then describe the four projects that make up this dissertation, each of which takes a different approach toward understanding the forces at work in early-stage decision-making and group processes.

Overall, this work provides evidence that online organizations cannot be understood in isolation. The socio-technical constraints of online organizations—in particular, the low transaction costs—cause many of the processes of individual decision-making and group growth to be deeply influenced by the environment in which a project exists. Our results paint a picture of people navigating a complex environment, making decisions based on past and current experiences and relationships, and leading to a system of online organizations which is interdependent, contingent, and difficult to predict. These findings have implications for how we

conceive of and study these organizations, but also for how we understand opportunities for cooperation. I discuss these conclusions and implications in more detail in the final chapter.

1.1. Collaborative Online Organizations

I use the term *collaborative online organizations* (or COO) to refer to groups that form and organize online to meet collective goals. In other words, they are the organizations that engage in commons-based peer production (Benkler, 2006) and open collaboration (Levine & Prietula, 2013). While many of the COO that I study are very small, I follow the liberal definition of an organization given by Farace, Monge, and Russell (1977) as groups of two or more people who take in inputs from the environment and work interdependently to create an output that meets shared goals. I focus most closely on groups like wikis or OpenStreetMap that have as their explicit goal the production of a single, public artifact.[1]

The outputs of these organizations are public goods. Public goods share two properties: people can't be excluded from using the good and use by one person doesn't diminish use by another; canonical examples include clean air and public parks. Classic theories of collective action and public goods suggest that self-interested people would not contribute to them because they can enjoy the benefits whether or not they pay the price of contributing (Olson, 1965). And yet, people do contribute to COO in huge numbers. In his groundbreaking book on peer production, Benkler (2006) explains how internet technologies reduce transaction costs in such a way that people can contribute to online public goods without strong incentives, arguing that social motivations are often sufficient. Benkler argues that when socially motivated contributions can be successfully harnessed, the societal benefits can be enormous.

[1]The other platform that we study—reddit—can also be characterized as having a shared artifact of a ranked list of posts and comments

For example, Wikipedia is now nearly two orders of magnitude larger than Encyclopedia Britannica ever was, and receives billions of pageviews per month,[2] providing richer information to a vastly larger population than a market-based encyclopedia ever could. Finding the conditions which encourage and allow for the production of more public goods should be a key goal of social scientists. I believe that a better understanding of the processes that produce collaborative online organizations can help us to meet that goal.

The success of COO in motivating and organizing volunteer labor is surprising and has spurred research across communication, economics, management, and sociology. While it is easy to see how the reduced transaction costs identified by Benkler (2006) would make contributing to online public goods more appealing, other constraints of COO would seem to make large-scale collaboration much more difficult. Research on the sociology of teams has long shown that incentives, hierarchies, and social relationships are key drivers of cooperation yet these projects typically operate without financial or market incentives, are organized without formal hierarchies, and are composed of pseudonymous strangers with weak social ties, who communicate almost exclusively via text (Benkler et al., 2015).

A body of research has grown around understanding how large collaborative online organizations overcome these problems. For example, researchers have shown that COO do not present a pure collective action problem. While the artifacts of COO are public goods available to everyone, there are also "selective incentives" (Olson, 1965) to participation: benefits that accrue to contributors that don't accrue to consumers. These include learning, enjoyment, mastery, and social relationships (Lakhani & Wolf, 2005; Nov, 2007; von Hippel & von Krogh, 2003; von Krogh, Haefliger, Spaeth, & Wallin, 2012). Another strand of research has

[2]Wikipedia and Britannica sizes are from https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons and pageview data is from https://perma.cc/FP9E-DWZQ

focused on how work is structured and power is enacted. While Benkler's (2006) original work for the most part ignored the role of structure and organization, research has found COO to be complex spaces. For example, researchers have found that there is a pathway from peripheral to sustained participation (Arazy et al., 2017; Bryant, Forte, & Bruckman, 2005; Preece & Shneiderman, 2009) which can be influenced by a newcomer's early experiences (Halfaker, Geiger, Morgan, & Riedl, 2013; Narayan, Orlowitz, Morgan, Hill, & Shaw, 2017). Others have found that work practices are structured and organized via norms, policies, and emergent hierarchies (Butler, Joyce, & Pike, 2008; Keegan, Gergle, & Contractor, 2013; Lakhani & von Hippel, 2003; Shaw & Hill, 2014; H. Zhu, Kraut, & Kittur, 2013) and that experienced members take on leadership and maintenance roles, both formal and informal (Arazy et al., 2017; Matei & Britt, 2017; H. Zhu, Kraut, & Kittur, 2012b). In other words, this research has found that large collaborative online organizations behave like organizations: they have complex roles, rules, bureaucracy, and hierarchies.

However, even large COO differ greatly from traditional organizations like firms. In order to succeed, projects need to have granular, modular outputs where contributions from disparate people can be aggregated without too much complexity (Benkler, 2006). Large projects require a core group of dedicated contributors (Butler et al., 2008; Matei & Britt, 2017) as well as complex socio-technical systems to integrate newcomers (Halfaker et al., 2013), deal with vandals (Geiger & Halfaker, 2013), and identify project needs.

1.2. Systems of collaborative organizations

While large-scale projects are successful at producing public goods, and we are starting to understand how, we understand much less about the dynamics of the larger population of

COO. The technological and institutional innovations that make large-scale COO possible—particularly the extremely low costs of entry and exit—mean that organizations have very fluid membership (Faraj, Jarvenpaa, & Majchrzak, 2011) and participants with diverse motivations (Nov, 2007). The same conditions which make it easy for someone to contribute to a COO also make it easy for them to contribute to a different COO instead, or to start a completely new COO. We know very little about how people decide which COO to contribute to and have difficulty predicting which organizations will succeed in their goals. For example, while Wikipedia has been successful at creating an encyclopedia, efforts at wiki-based textbooks and wiki-based news have been much less successful.[3] Is the difference in these outcomes due to differences in the nature of the task, the usefulness of available tools, differences in recruitment strategies, or differences in the competitive environment (e.g., these efforts now exist in a world with Instagram and Facebook, not to mention an already-successful Wikipedia)? Given the large-scale benefits of successful public goods, it is important to understand better which COO succeed, which ones fail, and why they fail.

One simple approach to studying community outcomes is simply to compare the COO that succeed with those that fail. This approach has been fruitful, identifying the importance of leadership and well-articulated goals (Schweik & English, 2012), inequality of participation, the speed of growth (Cunha et al., 2019), and the social network of current participants (Kairam et al., 2012) as strong predictors of group growth and longevity. A related strain of research looks at the role of design changes in influencing which COO grow, arguing that design choices can help to increase commitment, onboard newcomers, and recruit new users (Kraut,

[3]Indeed, even other attempts at creating online encyclopedias failed (Hill, 2013)

Resnick, & Kiesler, 2012). These perspectives focuses on understanding the individuals in a group and their relationships, resources, and actions as drivers of community outcomes.

Another approach borrows from biology to understand the ecological relationship of COO. In this tradition, initially applied to firms and volunteer organizations, people and their time are resources that organizations fight over (Cress, McPherson, & Rotolo, 1997; Hannan & Freeman, 1977; McPherson & Rotolo, 1996). This work often looks at how the location of an organization in the resource space influences whether it grows and survives and argues that relationships between groups are often more important than what happens within them. Ecology approaches have only recently been applied to online organizations. The few papers that exist show promising evidence that online organizations also experience competition and mutualism (TeBlunthuis et al., 2017; Wang, Butler, & Ren, 2013; H. Zhu, Kraut, & Kittur, 2014), something we would expect given the ease with which people can move between COO.

In sum, the two primary approaches to understanding community-level outcomes focus on understanding how within-group features and between-group relationships influence a population of COO. This dissertation builds on these approaches in two ways. First, by treating the existence of groups as contingent rather than as given. We work to shed light on why and how new COO are created and grow. Second, we treat individual decisions as the key driver of community-level outcomes, and develop approaches to try to understand how those decisions shape and are shaped by higher-level dynamics.

1.2.1. Individual decisions and open systems

We study the influences on early-stage COO using a systems-based approach. I use the term system in the same sense as an earlier generation of organizational communication scholars, who were very interested in studying organizations as "open systems" (Farace et al., 1977; D. Katz & Kahn, 1966; Rogers & Agarwala-Rogers, 1976). An open system is theorized to take in inputs from the environment such as information or materials, process them, and produce an output. Systems are made of subsystems; a firm may be made up of departments which are made up of work groups. Systems at each level influence and are influenced by the other levels. For example, the structure of a firm influences the communication load experienced by work groups and influences how they do their work (Farace et al., 1977). This perspective focuses on understanding processes, feedback loops, and the interrelationships between levels of analysis.

Typical organizational communication research from this tradition is at the level of work groups or departments, usually within firms. Some of the descendants of this approach include network analysis and collective intelligence, which explicitly recognize the importance of relations between people and across levels of analysis in understanding group and system outcomes (Kittur, Lee, & Kraut, 2009; Monge & Contractor, 2003; Welles & Contractor, 2015; Woolley, Aggarwal, & Malone, 2015).

However, each of these literatures, as well as the group literature and management literature, typically stops at the border of the group, treating the existence and composition of a group as given and static. In the context of COO, groups and their memberships emerge from the bottom up rather than being determined by a CEO or management team. Group membership is fluid and boundaries are incredibly porous; understanding the dynamics of group

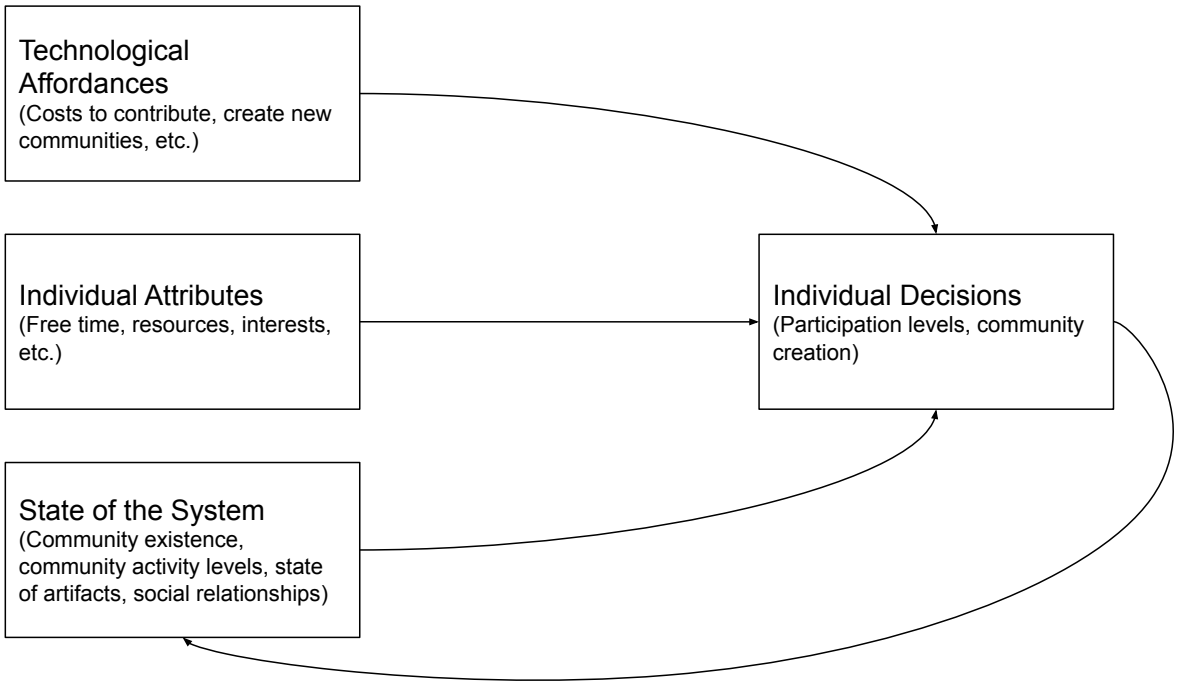


Figure 1.1. A conceptual model of how people decide whether and how to participate in collaborative online organizations. Decisions are influenced by attributes of the person, affordances of the platform, and the state of the system. Each decision then changes the state of the system in small ways, potentially influencing the decisions of others.

membership is key to understanding which groups meet their goals. And understanding the dynamics of group membership means understanding how individuals make decisions. From this perspective, COO participants are perpetually confronted with the question of how to allocate their efforts, with thousands of projects to choose from and few technical barriers to moving between them. We ask how people make these decisions and in particular how their experiences, relationships, and the current state of the system influence the decisions they make. We then use the theoretical and methodological approaches borrowed from these literatures to study how individual-level decisions recursively produce and reproduce higher level outcomes.

This perspective is visualized in Figure 1.1. The core of the model is individuals deciding how to interact with a set of collaborative online organizations. They can contribute to an existing COO, communicate with other COO participants, or create a new COO. These decisions are influenced by three primary factors: attributes of the person, affordances of the COO platform, and the state of the system.

Attributes of the person include things like their preferences, technological skills, and free time. How much someone participates, what COO they participate in, and what types of contributions they can make are obviously related to these personal attributes.

The second factor is the affordances of the COO platform. COO participants are volunteers with no financial incentives and extremely low barriers to entry and exit. They are engaged in producing a shared, continuously updated electronic artifact that is also a public good. They do this (typically) using pseudonyms and computer-mediated communication. While there is little research on early-stage COO explicitly, we can borrow from research in other contexts to predict how people's decisions might be influenced in this context. For example, by making it dramatically easier for people to move between COO or to start new COO, technological affordances increase not only the amount of participation but can change who participates and what types of COO they create.

The final factor is the state of the "COO system." Using the open systems perspective, aspects of the larger environment can influence how an individual or group acts.[4] For example, strong social connections within a given COO might encourage someone to participate more. They might also learn about new communities through social ties. Another influential aspect

[4]This link between macro and micro phenomena has long been important to sociologists. My approach draws heavily on "Coleman's boat," outlined in Coleman (1990), which focuses on connecting macro level features to individual decisions.

at the COO-level is the state of the artifact: for example, a project that feels nearly complete may have difficulty recruiting participants. At the population level, the existence and relative activity levels of COO may influence participation decisions. When COO cover the same topic, for example, people must choose how to allocate their effort between them.

The state of the system is not only an input into decisions but is an outcome of these decisions. One person's decision to start a new COO or to make a contribution changes the state of the system and influences how others decide what to do. Community-level outcomes such as COO size, longevity, and influence result from the aggregation of interdependent decisions made by individuals.

1.2.2. Analytical approach for studying systems of COO

Earlier researchers often complained about the difficulty of collecting the necessary data to test theories about interacting systems. For example, Rogers and Agarwala-Rogers (1976) discuss the need for longitudinal data to explore mutually causative communication processes but bemoan the expense and the risk of respondents being influenced by the data-gathering process (p. 19) while Monge, Farace, Eisenberg, Miller, and White (1984) claim that organizational communication processes were well-theorized but not empirically validated in large part because of the perceived difficulty of collecting and analyzing data.

Collaborative online organizations provide exactly the sort of data these researchers dreamed of. Many COO track granular measures of contribution and communication activity over time. In addition, they are an ideal setting for the type of inter-organizational systems research I have proposed because not only do people often move between organizations, but

COO platforms track activity of the same people across communities. This rich observational data can be used to track the decisions that individual people make as they navigate and co-create a system of COO. For example, Tan (2018) looked at redditors' previous activity patterns to predict whether and when they would join a new community. We combine this sort of observational approach with survey data to get a richer sense of how people make decisions in COO. Finally, we complement these two empirical methodological approaches with agent-based simulations. Agent-based simulations let us take theories about individual decision-making and simulate the macro-level implications of agents acting according to those theories. We use simulation to explore how different models of exposure to new communities and participation decision-making result in different distributions of participation and community size.

1.2.3. Early-stage collaborative organizations

We apply this theoretical and analytical framework to study the creation and early-stage behavior of collaborative online organizations. While about half of all new firms are still in business after five years (Artinger & Powell, 2016), the median open source project never gains a second contributor (Schweik & English, 2012). From open source software to wikis to online discussion communities, most COO are small and short-lived, with very few contributors and very few contributions. If we want to understand which COO succeed and why, as well as how to help more beneficial COO to succeed, then we need to focus on new and small communities. Despite the prevalence of small COO, the vast majority of research focuses on the few large projects.

This dissertation focuses on understanding what happens at the early stages of a COO's lifecycle. While it is tempting to assume that the factors which predict the success of an established community will apply to a new community, there are reasons to expect that early-stage communities differ in their goals (Tuckman & Jensen, 1977), their composition (Rogers, 1962), and their structure (Shaw & Hill, 2014). Individuals in these communities are thus faced with a unique set of decisions. It is likely that these early-stage decisions have positive feedback loops (e.g., having lots of early members makes it easier to recruit more members; Resnick, Konstan, Chen, and Kraut, 2012), giving these decisions outsized importance. The fact that so many new COO fail suggests that these early-stage decisions are difficult to get right.

Below I give a summary of four projects, each of which examines decision-making through the lens of open systems. This conceptual and methodological approach to studying systems of COO evolved iteratively through the course of pursuing these projects, and the papers are presented in chronological order. In the first project, focused on the relationship between social structure and retention and productivity, we treat groups as independent and only consider group-level features and group-level outcomes. In the second project, we hone in on COO foundings, with a focus on how environmental factors influence founding decisions. In the third project, we start to tie individual decision-making to community-level outcomes by studying how experiences and relationships influence whether an individual decides to start a new COO and whether it grows. Finally, the last project seeks to explicitly connect individual decision-making strategies about joining and leaving COO to system-level outcomes through simulation. A brief summary of each project is given below.

1.3. Project 1: Communication networks do not explain the growth or survival of early-stage peer production projects

In the first project, we build on a rich research tradition exploring relationships between intra-group dynamics and two group outcomes: productivity and longevity. Productivity has received a lot of attention in the COO literature, although much of it has focused on large COO. We know that these large projects rely on a dedicated core who do much of the administrative and policy-enforcing work (Crowston & Howison, 2005; Cunha et al., 2019; Matei & Britt, 2017) and a larger group who self-select into various roles (Arazy et al., 2017; Keegan, 2015; Welser et al., 2011). Much of this work is supported by autonomous bots who help to control vandals and enforce norms (Geiger & Halfaker, 2013; Halfaker et al., 2013). At an individual level, people are motivated to contribute in order to learn, to socialize, to have fun, or as a part of their identity (Nov, 2007; H. Zhu et al., 2012b).

There has also been some work on why individuals continue their participation in large-scale COO. Many of these studies focus on newcomers, finding that those who have negative early experiences are likely to leave (Halfaker et al., 2013; TeBlunthuis, Shaw, & Hill, 2018).

It is not clear how these motivations and tools work in early-stage COO, when there are far fewer people. For example, there is evidence that a more open system of production is replaced by more bureaucracy and structure as a project grows and ages (Shaw & Hill, 2014). There are also reasons to believe that the set of people who join early-stage communities may be different, and have different motivations than later participants. For example, early adopters are likely to be less risk-averse than others Rogers and Agarwala-Rogers (1976).

For insight into how productivity and retention work in early-stage COO we can also turn to the work group literature. Research on group processes in work groups suggests the

importance of coordination and social integration in supporting productivity (Ilgen, Hollenbeck, Johnson, & Jundt, 2005). Coordination and social integration happen through social and communication networks, and structures of communication are often used to measure how integrated and connected groups are. Consistent with these theories, work groups which are integrative, with high density and low hierarchy, are more successful along multiple dimensions (Balkundi & Harrison, 2006; Cummings & Cross, 2003).

The constraints of online peer production, such as text-based communication and low barriers to exit, should lead integrative structures to be even more important in predicting group outcomes. Empirically, studies of participants in large, successful projects suggest the importance of socialization in large COO (Bryant et al., 2005; Preece & Shneiderman, 2009) and theories about group formation suggest that integrative communication structures are vital in newly formed groups as they negotiate group goals and establish norms (Tuckman & Jensen, 1977).

The context for this study is the online wiki hosting platform Wikia.[5] Wikia hosts tens of thousands of wikis on various topics, ranging from popular media topics to recipes to academic job postings. Like many work groups, wiki contributors are engaged in understanding, synthesizing, and representing knowledge through the creation of an artifact. This similarity in task makes it reasonable to assume that wiki communities have similar sorts of coordination and socialization needs, and that theories from work groups should apply in this context.

We create measures of how integrative the communication networks are in over 1,000 early-stage wiki communities. We use these measures to predict two outcomes reported as important to wiki founders (Foote, Gergle, & Shaw, 2017): the quantity of information added

[5]Wikia has recently changed their name to Fandom.

to the wiki and the longevity of the community. Specifically, we create networks based on co-editing on "talk pages" and use these networks to measure hierarchy, centralization, density, and integration. We use a negative binomial regression model to predict how much content a community will add in the first 700 edits and use a Cox proportional hazards model to predict how long until a community goes 30 days without having multiple editors.

Contrary to theory and findings in the work group literature, and contrary to our expectations, we find that the structure of a community has little relationship with its productivity or its longevity. None of our measures show a clear relationship with either outcome, and when taken together they did not statistically improve model fit over a model with only controls.

Our model of decision-making provides provide a few possible explanations for these surprising findings. The first is that the affordance of a shared artifact (the wiki) can act as a partial substitute for communication. The wiki content can perform many of the functions that communication performs in face-to-face communication, such as coordinating activity and socializing group members. In this way it acts as a kind of "boundary object" (Lee, 2007). As people edit and view each other's work, the artifact can becomes a location for the creation, propagation, and debating of norms. It may thus partially replace explicit communication as a tool for negotiation, exercising power, and socialization.

Shared, editable artifacts can also produce "stigmergic coordination." Stigmergy is a means of coordinating group behavior by changing aspects of the environment rather than through explicit communication (Bolici, Howison, & Crowston, 2016). A classic example is an ant leaving a pheromone trail. In many COO contexts people can do things like create links to empty wiki pages or write class methods that don't do anything as a way to signal to others about what future work should be done.

A second explanation for our findings is based on a systems-level argument. In a work team, there are plenty of opportunities for individuals to do less or worse work; communication and social integration may help people to see the group as part of their identity. When that happens, people are more willing to contribute to group goals (Van Knippenberg, 2000). In the context of COO, the barriers to entry, exit, and formation are low, so there are many different groups that someone can contribute to. There is therefore a strong self-selection process, and those who choose to join a given group are likely much more dedicated to the topic than the general population. If someone already has a strong connection to a COO's goals then social structures may have a lower marginal influence on their group identity or willingness to contribute.

Both of these ideas present implications for public goods production to be tested in future work. For example, if artifacts do so much of the heavy lifting, then this suggests that projects that do not support stigmergy may have trouble coordinating their work. Our results also suggest that promoting more integrative social structures may not be effective in improving group outcomes, but that COO should concentrate on identifying people who are already interested in the topic.

1.4. Project 2: Motivations and Goals of Wiki Founders

Project 2 and Project 3 focus on the decision to start a new collaborative online organization. Thousands of new COO are created every day, yet there has been very little empirical research into how or why people start them. Indeed, the volume of new COO present a theoretical puzzle: In a system where failure is so common, why do people continue to start new communities?

One possible explanation is given by Kraut et al. (2012). In their influential book on designing online communities, they acknowledge the lack of empirical evidence around foundings but theorize that while founders and early users don't get financial rewards, they may gain additional benefits and power compared to later users if a community grows large. These benefits thus incentivize participation in new communities.

These sorts of conditional benefits for founders parallel the structure of rewards in firms, where entrepreneurs receive outsized benefits if and only if a firm succeeds. Entrepreneurship has been well-studied and provides a useful baseline for what we might expect to see in COO. Like COO founders, entrepreneurs are working to recruit a group of people to meet a collective goal. A major strand of the entrepreneurship literature has studied which people are more likely to become entrepreneurs and which entrepreneurs are more likely to succeed (e.g. Anderson & Miller, 2003; Backes-Gellner & Moog, 2013; Lazear, 2004).

In both of the papers studying founders, we start with these findings about entrepreneurs and ask how founders might differ in this context, where recruits are volunteers, founders do not receive financial rewards for success, and the difficulty, costs, and risks of founding a new organization are so much lower. Conceptually, we build on Project 1 by 1) focusing on individuals rather than groups as the level of analysis and 2) explicitly exploring the way that past experience and one's perception of the system influence decision-making.

Founders have an important role in the system of COO. We can think of the process of creating new COO as the way that the COO system explores the possibility space. If experienced users or men or high-SES people are much more likely to become founders then we may be missing out on areas of the space of possible COO that would be successful.

In addition, group foundings are difficult to study in many contexts and rely on retrospective reports. Having a data-rich environment where we can observe foundings as they happen is a novel opportunity to advance our understanding of founding processes in contexts that are more difficult to measure.

In the second project of this dissertation, presented and published at CHI (Foote et al., 2017), we use a survey to understand the way that founders approach their decision to start a new community. We asked founders about the purpose and goals for their communities, about their experience and motivations for becoming a founder, as well as their perception of whether they were starting generalist or specialist communities.

As we have argued earlier, much of the research on peer production focuses on the largest projects. When researchers do study projects of different sizes, they often use community size as a dependent variable, assuming that all projects are intended to grow large and that those which do not grow are failures. However, the incredibly low costs of creating a community lead us to predict that there may be a more diverse set of goals that people have for their communities.

In order to understand the perspectives of founders as they were founding a new community, we worked with Wikia to send a survey to people immediately after they started a new wiki. This project used survey responses from over 300 founders. We discovered that founders had a mix of motivations. While the most common motivations were around the desire to spread information about a topic and build community, many respondents reported starting the wiki as a way to learn about the software or to work on personal projects such as fanfiction.

Founders also had varied goals. When we asked them to rank their goals, by far the most common top goal was “high quality information about the topic.” Many others chose an active or long-lasting community as their top goal. Contrary to the assumptions of much research on COO, only 15 out of 331 respondents chose having a large number of contributors as their top goal.

These findings support our overall model of how people make decisions (Figure 1.1). We found strong evidence that affordance of lower founding costs leads to a large amount of diversity in the types of projects that can be attempted. We not only found evidence of diversity in goals but also evidence of just how low the costs of founding are. When we asked how long people had been thinking about starting their community, 46% of respondents had only thought about it for "a few hours" or "a few minutes".

We also found evidence that founders are aware of and influenced by the larger system of COO and seek to find topical niches where their projects are likely to be successful. Because costs are so low, these niches can be incredibly narrow. When we asked founders about the potential audience for their projects, 50% chose the response that “only a handful of people” would be interested in the topic of their wiki. When we asked how many contributors they expected in the first 30 days, the median response was 4. Respondents also explained that many of them started a wiki about a topic explicitly because a wiki about that topic didn’t currently exist. They intentionally sought out niches with low competition. And so, technological affordances and the state of the system influenced individual decisions, encouraging people to create small, niche communities.

These findings also have implications for how we understand online community outcomes. We should not assume that growth is the goal. By ignoring other aspects of success which are

more important to founders (and likely to community members) we are missing out on COO which are successful without growth. A related point is our finding that small communities are intentional and diverse. They are not simply a sample of larger COO which happened to get unlucky. Rather, they are attempting to do different things, in different ways. These small-scale, temporary organizations deserve additional scholarly attention.

1.5. Project 3: The Behavior and Network Position of Peer Production Founders

The survey approach in the second project gave us data about why founders became founders and how they perceived the purposes of their COO. However, this approach cannot tell us anything about all of the people who don't become founders. In the third project we use digital trace data to explore how past experience and position in social networks relates to a person's likelihood of founding a community, as well as whether or not their communities grow. As in the first two papers, we use theories and research from existing contexts (primarily entrepreneurship) to guide the questions that we ask, and to compare how the constraints of COO alter which people become founders, how much influence they hold, and what kind of groups they create.

There are a number of theoretical and empirical reasons to believe that starting a new community is very different from joining an existing community. Empirically, we see that in contexts like entrepreneurship, those who begin new businesses are different from those who don't: for example, they have more diverse experience (Backes-Gellner & Moog, 2013) and experience working with other entrepreneurs (Nanda & Sørensen, 2010). While there is almost no direct research on COO founders, theories suggest that people who join early-stage communities have different motivations than later participants (Resnick et al., 2012).

In this paper, which was presented and published at iConference (Foote & Contractor, 2018), we again use digital trace data from Wikia. We record different types of data from three different periods in time. First, we created behavioral and network measures for over 60,000 users who made at least one edit in March or April of 2009. We then recorded the first three contributors to all of the new wikis created in May and June of 2009. Finally, we looked at the number of contributors that participated in each of these new wikis from their founding until January 2010. We used this data to compare the activity and network position of Wikia founders to non-founders and we identify differences between these populations. We also explored the relationship between these measures and the number of contributors a wiki gains.

We find that the vast majority of founders are very new users who don't even appear in our dataset (i.e., had zero edits in March or April). Among those founders who are not brand new users, we find that founders, despite the differences in constraints, are similar to entrepreneurs. Like entrepreneurs, they are more likely to have diverse experience and experience with newer communities. They also have more experience than non-founders, and we find some evidence that they are on the periphery of the social networks on Wikia.

When we look at the relationship between founder measures and community growth, there is a lot more noise and founder measures explain very little of the variation. There is some slight evidence that founders whose communities grow are more central in social networks, but the overall finding is that past behavior and network position do not explain community growth very well.

One key constraint of COO is that founders have few tools for incentivizing others to participate. Unlike entrepreneurs, who can use financial incentives to meet their goals, founders

must motivate volunteers who can easily leave at any time. On the flip side, founders also receive fewer personal benefits from the success of a project. Unlike firms, online communities can be seen as public goods (Olson, 1965) and founders are not typically enriched by success. Taken together, these constraints help to explain why founders may not have as much influence over their communities.

COO platforms also have very low barriers to creating new communities. This helps to explain our finding that COO were very likely to have been started by new users. Low barriers also lead to changes in the purposes and goals of a project, as we learned in Project 2. When a new firm is started, it requires capital investment and financial risk. Therefore, one of the goals of the firm must be to grow large enough (in size and market capture) to repay those investments. By reducing these constraints, online platforms allow for communities to be formed which draw from a larger set of purposes and goals. Thus, we may be underestimating the influence of founders when we only focus on growth as an outcome.

Finally, we discussed how a biased process (e.g., if only experienced users started new communities) could lead to a system that does not explore the entire space of possible projects. Research suggests that peripheral COO users are more diverse than long-time participants (Shaw & Hargittai, 2018). This being the case, our finding that most wikis are started by new users suggests that the process of founding is relatively inclusive and the system of COO explores the space relatively broadly. We expect that this is again a function of low costs: when the barriers to project creation are low enough then new users can provide new ideas. Of course, there is still plenty of room for biases to influence foundings. For example, while founders are likely to be new users, it is possible and indeed probable that women and girls are less likely to

learn about or become users in the first place (Shaw & Hargittai, 2018). It may also be more difficult for certain people or types of projects to gain contributors once they are founded.

1.6. Project 4: Social exposure and participation processes in online communities

The final project is the most explicit attempt to model the process outlined in Figure 1.1. In this paper, versions of which were presented at ICA and IC2S2, we attempt to explicitly model the relationship between individual decision-making and higher-level systems of COO.

Social computing researchers have developed a number of theories about how individuals are exposed to new online communities and COO and how they decide whether or not to participate in a given community (Panciera, Halfaker, & Terveen, 2009; Preece, Nonnecke, & Andrews, 2004; Ren et al., 2012; Resnick et al., 2012, e.g.). These theories are rich descriptions of individual-level behavior but rarely consider the higher-level implications.

We argue that one reason for this is the difficulty in moving from individual-level models to group-level or population-level outcomes and argue that agent-based simulation can be used for this purpose. We can create agents who act according to social computing theories and simulate their behavior. If their behavior captures important aspects of how people behave in the real world, then simulated populations should exhibit the same macro-level patterns of behavior seen in real COO platforms. In this paper we focus on two outcomes which are closely related to exposure and participation decisions: the distribution of community sizes and the distribution of the number of communities which each person participates in, which we call the "participation rate."

We attempt to model participation decisions and exposure in ways that capture the main aspects of previous research. This is easier for participation decisions, as Resnick et al. (2012)

already created a mathematical model to describe how people decide whether to join a community. They claim that people estimate the future size of the community in order to estimate the benefits they will receive (e.g., information, friendship, etc.). If these benefits outweigh the costs (e.g., learning new norms, new software, etc.) then they will join the community. We model exposure using a simple word of mouth model, whereby agents learn about new communities through people who belong to their current communities. We simulate each of these models working independently and then simulate a model where both of these mechanisms are working together.

We show that while neither theory by itself is a good match for the empirical data, the combined model produces heavy-tailed distributions for community size which look quite similar to what we see empirically. This suggests that word of mouth exposure combined with participation rules that prioritize quickly growing communities are a plausible partial explanation for skewed community sizes.

This combined model also produces the most skewed participation rate, although even in this model participation is not nearly as skewed as in real online platforms. We identify a few possible sources of additional inequality, including heterogeneity in free time or interests as well as untheorized cumulative advantage mechanisms whereby people who are already active in many communities are likely to participate in more. We conclude with a discussion of the promise of agent-based simulation for social computing research.

1.7. Key findings

These projects present a novel approach to thinking about how collaborative online organizations work. Rather than treating organizations as the focus of study, we show how a focus

on individual decisions can illuminate both individual behavior and help us to understand the mechanisms behind community outcomes.

This approach led to a unique set of research projects which challenge a number of assumptions and findings about collaborative online organizations and human behavior. First, we find that people's participation decisions are influenced both by their past experience as well as by the current state of the system. Second, we learned that small, temporary communities are intentional and meet needs that that neither traditional organizations nor large-scale COO can. Third, we find evidence that—unlike work groups—early-stage COO are not held together through social ties. Rather, they are motivated by a shared interest and goal and coordinate much of their work through a shared artifact. Finally, we theorize that the low costs to create and participate in COO are behind many of the dynamics that we observed, from how individuals decide whether to start a new COO to how rich-get-richer dynamics emerge in populations of communities.

The following four chapters contain full papers for each of the projects. The final chapter summarizes and expands on the key insights from this dissertation and charts out future directions for research.

CHAPTER 2

Communication networks do not explain the growth or survival of early-stage peer production projects

Jeremy Foote
Northwestern University

Aaron Shaw
Northwestern University

Benjamin Mako Hill
University of Washington

Communication enables coordination and social integration in collaborative groups. In the contexts of work groups and teams, prior research finds that more dense and integrated communication structures support better performance. We explore the relationship between communication structure and group performance in a population of early-stage peer production wiki communities engaged in the collaborative production of shared information resources. We theorize that there is an especially strong need for coordination and social integration in small, newly formed online communities, and that communities with relatively more integrative communication networks will be more successful. We test this theory by measuring communication network structure and group outcomes in a population of 1,002 nascent wikis.

Contrary to prior literature and our expectations, we find a very weak relationship between communication structure and collaborative performance. We suggest a number of explanations, including the role of shared artifacts in coordinating work and integrating newcomers.

2.1. Introduction

Work teams perform best when members of the team are socially integrated and when the team is efficient in coordinating its activities. A large body of research has shown that socially integrated and effective coordination are enabled by and reflected in patterns of communication. High performing teams have communication networks which are more dense, less hierarchical, and more integrative (Balkundi & Harrison, 2006; Cummings & Cross, 2003; Reagans & Zuckerman, 2001).

Nearly all of this research has studied work groups consisting of paid employees, typically coordinating face-to-face on projects which they are assigned to. The Internet has made new ways of organizing production possible that produce enormously valuable information goods without firms, payment, task allocation, or face-to-face interactions. What do productive communication structures look like for these groups?

Commons-based peer production communities like those that produce Wikipedia and open-source software are engaged in complex information resource creation and organization. However, they are composed primarily of unpaid volunteers with minimal formal hierarchy who rarely meet face-to-face. If successful work groups require dense, integrative networks even when working and talking face-to-face, it seems like integrative structures would be even more important for group performance when communication happens only via text and when

group members can leave at any time without financial or career consequences. These patterns of communication should be particularly important in newly formed groups, which engage in establishing norms and priorities (Tuckman, 1965).

We use an extensive population of new peer production projects—1,002 online wikis from Wikia—to identify the network communication structures that are associated with higher rates of community productivity and longevity. Surprisingly, we find that network structure does not predict the performance of networked organizations producing networked artifacts. We find that neither productivity nor longevity have a discernible relationship with measures of integrative network structure. We discuss two ways in which the affordances of peer production might explain the results we find. First, early-stage wikis may require less coordination and social integration than work groups. Second, peer production groups can coordinate via artifacts in ways that traditional work groups may not be able to. This "stigmergic" coordination allows them to coordinate work and priorities without direct communication. Our findings suggest that these affordances allow peer production groups to work together and survive without the social structures previous research has suggested as necessary for small groups to function.

2.2. Background

2.2.1. Coordination & Social Integration in Organizations

Many groups and organizations use communication to engage in complex coordination and information-sharing activities that advance their collective goals. An input-process-output (I-P-O) model of team functioning provides a framework for understanding this, where the *inputs*

are the attributes, knowledge, and resources of team members, the *process* includes the interaction between team members, such as information sharing, social influence, and communication, and the *output* includes the work produced as well as changes to the team members or their relations to one another (Ilgen et al., 2005; Mathieu, Maynard, Rapp, & Gilson, 2008).

Two of the most important aspects of the 'process' portion of the I-P-O model are coordination and social integration. Coordination involves linking together different actors towards collective tasks (Van De Ven, Delbecq, & Koenig, 1976). The amount and type of coordination needed depends on the tasks a group is engaged in. For example, more interdependent processes require more coordination (Tushman, 1979; Van De Ven et al., 1976). One type of coordination occurs at the organization level, allowing teams to build a shared understanding of their goals and what needs to be done (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000). Another type of coordination occurs at the resource level, as team members learn who has various resources or capabilities and how to access these resources (Kotlarsky, van den Hooff, & Houtman, 2015; Wegner, 1987).

While different communication network structures can work better for different group or task types, dense, fairly egalitarian, well-integrated networks allow information to flow efficiently (N. Katz, Lazer, Arrow, & Contractor, 2005). These structures may be especially important in groups collaborating over computer-mediated communication channels since these groups often have more difficulties coordinating their work (Gibson & Gibbs, 2006; P. J. Hinds & Bailey, 2003; Kotlarsky et al., 2015) and creating shared understanding (P. J. Hinds & Mortensen, 2005).

Building shared understanding requires social integration, which is the second important group process. Virtual teams, work groups, and other collaborative groups are social organizations whose ability to process inputs relates to their degree of social cohesion. Communication is especially important in helping to create social integration (Gibbs, Kim, & Ki, 2016). Through communication, new members learn about group norms and expectations and start to identify as a member of the group (Scott, 2007). When group members see group membership as a key part of their identity they are more willing to sacrifice their own goals in order to contribute to the group's goals (Van Knippenberg, 2000). This social integration tends to support effective and long-lasting collaboration.

Inclusive, integrative communication structures allow for and reflect effective patterns of coordination and social integration. A large body of empirical research by communication, organization, and psychology scholars has studied the relationship between communication patterns and group performance in firm-based work groups and distributed or virtual teams. Overall, this research finds that highly integrated communication structures correlate with better task performance (Balkundi & Harrison, 2006; Cummings & Cross, 2003; P. J. Hinds & Kiesler, 2002; Reagans, Zuckerman, & McEvily, 2004).

It is much less clear how these results might apply to other types of teams. Because nearly all of the previous research has studied firm-based teams, Cummings and Cross (2003) call for a better understanding of "…informal organizations where there are performance goals, yet there are not constraints from formal hierarchy or other cultural norms that impose structures" (p. 209). Our work takes up that call.

2.2.2. Communication networks in commons-based peer production

Specifically, we consider the relationship between communication networks and performance in the context of commons-based peer production (CBPP). Coined by Benkler (2002), "commons-based peer production" describes a mode of organizing the creation of publicly accessibly information goods through the mass aggregation of many small contributions. The most well-known examples of CBPP are Wikipedia and open-source software like GNU/Linux and Firefox. These projects have been able to coordinate the efforts of hundreds of thousands of volunteers to create products that often outcompete market-based alternatives (Benkler, 2006).

CBPP projects differ from firm-based work groups along a number of important dimensions, including organizational structure, membership composition, and communication tools (Benkler, 2016). In terms of organizational structure, CBPP projects rarely have clear, formal governance or management hierarchies when they begin. While a small number of users have technological tools (e.g., for banning vandals), these users do not typically make assignments or organize tasks. In terms of membership, CBPP participants are primarily volunteers who can come or go at will, rather than paid employees. Finally, communication in these projects typically happens through text-based, computer-mediated channels rather than face-to-face, which makes coordination both more important and more difficult (DeSanctis & Monge, 1998; Gibson & Gibbs, 2006; Wiesenfeld, Raghuram, & Garud, 1998).

At first blush, it seems like this open, participatory structure would lead to similarly egalitarian structures of communication and participation resembling the integrative networks characteristic of effective work groups and virtual teams. However, prior research indicates that many of the largest CBPP projects possess extremely unequal levels of participation with a group of core contributors who make most of the contributions, supplemented by a "long

tail" of more peripheral contributors who each add very little (Keegan et al., 2013; Matei & Britt, 2017; Schweik & English, 2012). Similar patterns characterize the communication networks of these mature projects, with sparse networks which have many contributors on the periphery (Crowston, Wei, Li, & Howison, 2006) and high levels of hierarchy (Crowston & Howison, 2006). These patterns likely stem from the incredibly low barriers to entry and exit in CBPP which allow for incredibly low-commitment contributions, a dynamic which simply doesn't exist for firm-based groups.

However, this research suffers from a few limitations. First, it is descriptive rather than inferential. That is, it does not typically seek to understand the relationship between differences in network structures and project outcomes. These studies provide evidence that CBPP is less integrative than work groups, but that does not mean that *relatively* less integrative CBPP projects are more or less successful. There have been exceptions to this limitation, and a few papers have compared community structures across projects. In this research, equality of participation and network density typically have a negative relationship with performance and community growth (D. Hinds & Lee, 2009; Qin, Cunningham, & Salter-Townshend, 2015; Roth, Taraborelli, & Gilbert, 2008).

Even these papers often suffer from the second limitation, which is a focus on relatively mature communities that have already produced substantial resources. While some work includes smaller communities (e.g. Kittur & Kraut, 2010; Roth et al., 2008; Schweik & English, 2012) or considers early stages of peer production projects directly (e.g. Foote et al., 2017), only one small-scale study has explicitly examined the structure of early communities (D. Hinds & Lee, 2009). The typical focus on popular, mature projects may explain the prevalence of hierarchical, sparse, and unequal networks in the CBPP literature. As we discuss more extensively

in the next section, there are a number of reasons to believe that early-stage communities differ from large-scale projects in ways that predict a greater need for integrative networks.

2.2.3. Communication in nascent peer production communities

Our approach looks at a population of projects when they are at the same early stage. While the largest CBPP communities boast millions of participants, most CBPP communities never grow larger than a few members (Schweik & English, 2012). If we want to understand the way that CBPP works—and when it doesn't—we should focus much more attention on early projects. Studying only the relatively rare large and successful projects "selects on the dependent variable." How might we expect coordination and social integration processes to occur in early-stage CBPP and what sorts of communication structures would reflect productive and long-lasting communities? The empirical evidence, primarily from large-scale CBPP projects, suggests that dense communication networks are weakly, but negatively, associated with performance (Qin et al., 2015; Roth et al., 2008). While it is tempting to assume that these findings are common to all CBPP, there are a number of theoretical and empirical reasons to believe that early-stage projects have a different composition of members and different objectives than mature projects.

First, organizational theories suggest that groups go through different stages with different communicative needs. For example, Tuckman (1965) claims that early groups go through stages of orientation and norm creation before they can successfully cooperate. How, when, and whether these group formation stages happen in CBPP is not well understood, but empirical work suggests that CBPP communities do engage in different types of work at different

stages. For example, Schweik and English (2012) find that open source software projects experience phases of growth, maturity, and decline. Distinct patterns of leadership, governance arrangements, and collective behavior characterize each phase and explain variation in project outcomes. Related research has found that as they grow communities become more structured (Kittur & Kraut, 2010; Shaw & Hill, 2014; TeBlunthuis et al., 2018). This theoretical and empirical evidence suggests that early-stage communities have different sets of needs such as norm-formation and goal-setting, which should require inclusive, broad conversations among group members.

Second, early-stage projects are likely composed of contributors who differ from later users. CBPP projects have often been characterized as public goods and subject to theories of collective action. According to these theories, early contributions often have a lower benefit-to-cost ratio (Marwell & Oliver, 1993; Monge et al., 1998). For example, in our context wiki founders must choose a name for the wiki, create navigation pages, and recruit contributors. None of these activities directly contribute to the good of a shared information artifact. In public goods like these, we would expect that more interested and more resource-rich individuals will choose to make early contributions (Marwell & Oliver, 1993). Indeed, empirical work provides some evidence that contributors to new CBPP communities are more active and experienced than others (Foote & Contractor, 2018). Related CBPP research shows the importance of having a core group like this to coordinate work and integrate newcomers (Kittur & Kraut, 2010). Those who have been integrated are more willing to contribute toward group goals (Kittur, Pendleton, & Kraut, 2009) and are more likely to continue participating in the community (Halfaker, Geiger, & Terveen, 2014).

These theories suggest that members of nascent online communities are a more dedicated set of participants engaged in work that requires more interdependence than mature projects. Therefore, we expect early-stage CBPP projects to benefit from integrative communication structures similar to those found in effective work groups. We use these insights to present a number of hypotheses about the relationship between communication structures and project success.

Scholars have measured success in CBPP communities in different ways, from membership growth to artifact quality to contribution amount (Howison, Inoue, & Crowston, 2006; Roth et al., 2008; Schweik & English, 2012; H. Zhu, Kraut, & Kittur, 2012a). We focus on two outcomes which wiki founders reported as most important to them: the creation of information (*productivity*) and building a long-lasting community (*longevity*) (Foote et al., 2017). These outcomes also match up with the processes we have hypothesized as important, with coordination processes relating more closely to productivity and social integration relating more to community longevity. Communication structure can also be characterized in many different ways. We borrow from the communication and management literature on teams and networks to identify four features of communication structure: *centralization*, *density*, *hierarchy*, and *periphery size*. We discuss how each feature relates to coordination and social integration and make hypotheses about how each might relate to productivity and longevity in early-stage CBPP projects.

2.2.4. Centralization

Centralization refers to variance in how many others each person in a network is connected to. When a few people are much more connected than others, centralization is high. While

higher variance in *contributions* correlates with CBPP community productivity (Kittur & Kraut, 2010; Qin et al., 2015), theories of group formation suggest that highly centralized communities will have difficulty in forming shared norms and coordinating who knows what.

H1a: Centralization will have a negative relationship with productivity.

By definition, a highly centralized community has many members who are are not well connected to others and therefore have less opportunity to become socially integrated. People who are not integrated are likely to leave the community.

H1b: Centralization will have a negative relationship with longevity

2.2.5. Hierarchy

Hierarchy in networks refers to how circular communication flows are. Hierarchy is high when communication only moves one way. The intuition is that a network is hierarchical if commands and communication move in one direction but not the other. While hierarchy may reflect organization, it typically indicates problematic coordination and socialization processes, such as contributors who aren't willing to talk to each other. Hierarchy has been found to have a negative association with work group outcomes (Balkundi & Harrison, 2006; Cummings & Cross, 2003). While absolute hierarchy in CBPP is quite high due to low exit costs (Crowston & Howison, 2006), theory and empirical evidence from work groups lead us to predict that:

H2a: Hierarchy will have a negative relationship with productivity.

H2b: Hierarchy will have a negative relationship with longevity.

2.2.6. Core membership

Peripheral contributors bring valuable information into CBPP communities (Gorbatâi, 2014; Ransbotham & Kane, 2011) and large peripheries may signal the popularity and importance of the project, encouraging others to contribute (Zhang & Zhu, 2011). On the other hand, productive work groups have most of their members in the core, allowing them to integrate ideas from all team members. We expect that nascent communities more closely resemble these smaller teams, and the increased need for coordination and consensus at this early stage of CBPP lead us to predict that a strong core is vital for coordinating and encouraging contributions.

H3a: Core membership size will have a positive relationship with productivity.

Group longevity requires a subset of members who continue to participate over time, and this willingness to participate is increased through social integration (Van Knippenberg, 2000). Core membership is one of the closest analogs to social integration: groups where many members have relationships with many others are likely to be socially integrated and committed to the project.

H3b: Core membership size will have a positive relationship with community longevity

2.2.7. Density

In dense networks, many people are connected to each other. It seems like density should always contribute to coordination and social integration. Indeed, in work groups density is generally, albeit weakly, associated with better performance (Balkundi & Harrison, 2006; Reagans & Zuckerman, 2001). However, in both early-stage open source software (D. Hinds &

Lee, 2009) and Wikipedia "project" groups (Qin et al., 2015), density has a negative relationship with productivity. One explanation is that maintaining a strong, dense social network may take time away from actually contributing to the artifact (Qin et al., 2015). This empirical evidence from such a similar population of communities leads us to hypothesize that, especially after controlling for hierarchy and core membership, higher density will be costly to coordination and productivity.

H4a: Density will have a negative relationship with productivity.

On the other hand, if high density is an indicator of "too much" socializing, then we would expect that members with high density will be more socially integrated. While they may be less likely to contribute to group productivity, theory suggests that they will also be less likely to leave the community and the project will be less likely to dissolve.

H4b: Density will have a positive relationship with longevity.

2.3. Data & Measures

Our data comes from a population of online communities working to produce wikis that are hosted by the firm Wikia. The term "wiki" refers to both the type of software which facilitates the collaborative creation and distribution of information resources as well as the resource produced. While Wikipedia is the most well-known instance of a wiki, there are hundreds of thousands of other wiki communities, with varied goals, topics, and community structures.

Wikia was founded in 2004 by Jimmy Wales, a founder of Wikipedia, and Angela Beesley, an active Wikipedia contributor. Like Wikipedia (and unlike other wiki hosting sites), all Wikia wikis are publicly viewable and publicly editable. Wikia holds the largest sample of

publicly accessible and editable peer production wikis, many of which have widely divergent topics and participation rates. This diversity and variance help ensure that the results of our analysis have broader generalizability beyond analyses of a single community or a smaller scale comparison. Additionally, the dataset we collected in April 2010 is the most complete set of Wikia wikis available. In 2010, Wikia began deleting small, inactive wikis from the site thereby making it difficult to evaluate questions about early-stage peer production projects.

In preparing the dataset for analysis, we applied several inclusion criteria. Before constructing our measures, we removed all edits to these communities marked as having been made by bots. While our focus is on new, small communities, in reality most communities never grow large enough to be considered communities at all. Of over 76,000 wikis that we have data for, only 1,268 had at least 700 non-bot edits to article pages[1] as of April 2010. At the point when 700 edits had been made, the median number of participants in the communication network is 10, a similar size to large work groups. We calculate our measures based on a "data snapshot" of all of the edits made to a community at the point it received its 700th edit.[2]

2.3.1. Dependent Variables: Productivity and Longevity

To measure productivity, we count the total number of words added by contributors in the first 700 non-reverted edits. The completeness and quality of the wiki is one of the primary goals of early-stage contributors to CBPP (Foote et al., 2017) and this is one measure of how quickly the wiki as an artifact is progressing. Because we look at the same number of edits for each wiki (rather than, for example, the number of edits per month), this is also a rough

[1]Wikis have a number of "namespaces", such as talk pages, user pages, and administrative pages. The primary content of the wiki, such as articles, is typically located in the main namespace.
[2]In separate robustness checks reported in the appendix, we ran models using snapshots at 500 and 900 edits and found similar results.

measure of the per-edit effort that participants put into the project. To measure longevity, we count the number of months from the 700th edit to the start of the first 30-day period in which the wiki was edited by fewer than two people.

2.3.2. Network Measures

Our measures of communication structure are based on behavioral traces of interaction captured in the logs of each wiki. This provides a more objective measure of interaction than the survey-based self-reports typical in studies of communication structure in work groups (e.g., Cummings & Cross, 2003). Both a strength and a weakness, this approach gives us a highly accurate representation of the actual communication structure but does not capture the cognitive understanding that users have of the network.

To construct measures of communication network structure, we first create communication networks for each wiki. These networks attempt to capture interactions within the spaces dedicated to conversational activities. These are the "talk" pages connected to every article, user, or policy page on each wiki. Talk pages are used to discuss the page to which they are connected (see Figure 2.1). For example, a talk page connected to an article is typically used to discuss ways that the article can be improved, although they can also be used as places to socialize or discuss a topic. We create communication networks by looking at the edits to all of the talk pages for each wiki. Although talk page edits are not always explicitly directed toward others, we consider recent editors of the same page as the recipients of an edit. We create a directed tie in our network from every talk page editor to the contributors of the previous five edits to that page. Additionally, we create a directed tie when any contributor edits another's

Figure 2.1. Example of a talk page from the Lord of the Rings wiki. Users discuss the article's topic and how to improve the article.

"user talk" page. This is a public talk page connected to a user's account, typically used for interpersonal communication. We attempt to measure communicative relationships rather than temporary interactions by only keeping edges that represent at least two interactions (i.e., for a tie from i to j to exist, i would have to edit a page after j or write on j's user talk page at least twice). We create a network for each wiki at the point that it had 700 edits.

We limit our analysis to communities which have at least four people in their communication network, dropping 202 communities. Our rationale is that the network measures we apply (described below) are not meaningful for smaller networks. We also remove 64 wikis where the structure was so simple that our network measures could not be computed. This leaves us with a total of 1,002 wiki networks, which we use to construct the following measures of network structure.

Density is the number of ties between the N nodes in a given network divided by the number of possible ties if every node was connected to every other node.

We follow Cummings and Cross (2003) by using the hierarchy measure defined by Krackhardt (1994). We calculate *hierarchy* as the ratio of paths in the graph which are not cyclical. For a given path from person v_i to person v_j, the path is cyclical if there also exists a path from v_j to v_i.

We use a scaleable method of identifying core group membership, based on the same theoretical construct in the work group literature (Cummings & Cross, 2003) and related to measures of closure (e.g. Shen, Monge, & Williams, 2014). To measure core-periphery structure, we calculate the "coreness" for each node. This is the largest k subgraph the node is in for which all members of the subgraph have at least k ties with other members of the subgraph (Seidman, 1983).[3] We identify contributors as being central (i.e., non-peripheral) if their coreness is at least 3, and measure the proportion of contributors who are central in each community, which we call the *core member ratio*. When this measure is high, it means that many of the users are part of a cohesive group, where all of the members of the group have interactions with others in the group.

[3]More detail about this measure is provided in the appendix.

We use two measures of centralization, *indegree centralization*, which is a measure of inequality in how much each person is talked to, and *betweenness centralization*, which is a measure of inequality in how often a person is in the shortest path between other people (Freeman, Roeder, & Mulholland, 1980). These two measures capture two related but distinct ways in which a community might be centralized. Following Qin et al. (2015), for indegree centralization we take the gini coefficient of the distribution of indegree over nodes and for betweenness centralization we take the gini of betweenness centrality scores.

2.3.3. Controls

Many network measures have a relationship with the size of the network. For example, large networks typically have lower density, since the number of possible ties scales exponentially. Therefore, we include a control measure for *network size*. We also include *talk edits* and *edge weight*, two controls for the overall amount of activity between members, in order to distinguish the structure of communication from the amount of communication. In order to help control for time-varying factors that may have played a role in shaping the growth patterns of particular wikis we include *months since founding*.

The popularity of wiki topics is a potential confound for the relationships we seek to identify. For example, popular topics may attract more peripheral contributors than niche topics. Without access to a direct measure of popularity, we rely on several proxies. First, we include measures of how quickly the community is producing content. Specifically, we measure *days to 350 (700) edits*. We also control for the number of *total editors* and *active editors* (editors with at least ten edits). Finally, we include *contribution inequality*, a measure which has been found

to be an important predictor of performance in previous work on CBPP (Kittur & Kraut, 2010).

2.4. Analysis

We estimate the relationship between project structure and project outcomes by regressing our project-level measures of productivity and community survival on our project-level measures of network structure after the first 700 edits. Because our measures of productivity are all over-dispersed counts, we use negative binomial linear regression to model productivity. To estimate community survival, we fit Cox proportional-hazards models. In each model, we added polynomial terms for significant controls until model fits failed to improve based on likelihood-ratio tests.

Our appendix shows summary statistics for each of our variables. As is typical in online communities, almost all of our count measures are highly skewed. To address this, we use the natural log of *momentum* measures, the *number of editors*, and the *network size* measures in all of our models. We also use the log of *density*.

2.5. Results

Figure 2.2 shows the results of the full model used to predict productivity, with 95% bootstrapped confidence intervals.[4] Surprisingly, we find that none of the network measures have a significant relationship with productivity, and that including the network measures does not improve the fit of the model based on AIC. The only possible relationship is for density (*H4a*), which barely falls outside of our confidence interval ($\beta = 0.32$, CI $[-0.022, 0.687]$), but has a significant relationship with productivity in one of our two robustness check datasets. If a real

[4]Full regression results are available in the appendix.

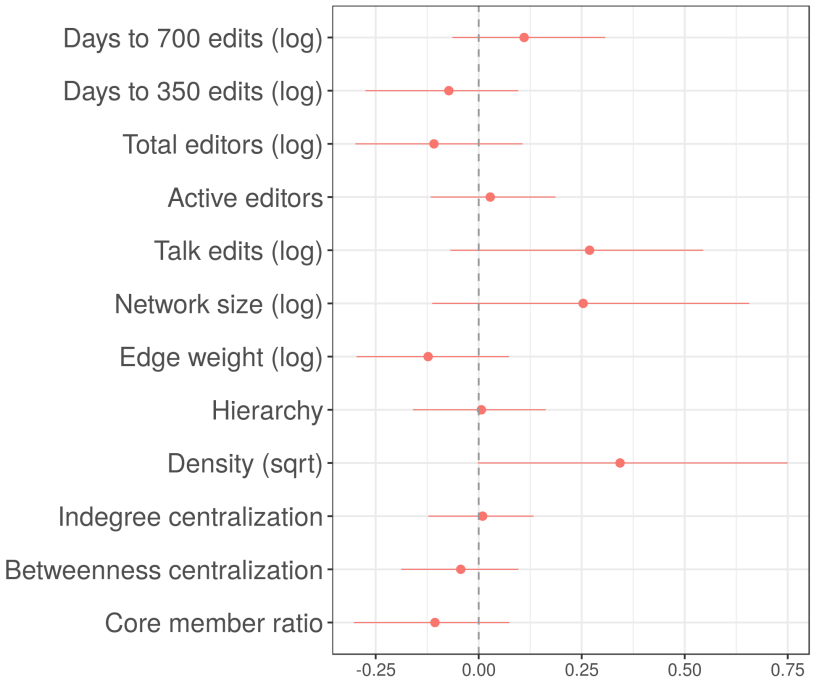


Figure 2.2. Scaled regression coefficients predicting the number of non-reverted words added in the first 700 edits. Polynomial control terms are excluded for clarity. Error bars represent bootstrapped 95% confidence intervals.

relationship exists, it is very weak. We did not find evidence for *H1a*, *H2a*, *or H3a*, as neither centralization measure (indegree centralization: $\beta = -0.01$, CI $[-0.469, 0.366]$; betweenness centralization: $\beta = -0.25$, CI $[-1.052, 0.435]$), hierarchy ($\beta = 0.01$, CI $[-0.314, 0.361]$), nor the core membership ratio ($\beta = -0.19$, CI $[-0.529, 0.137]$) explain variation in productivity. A number of our controls were positively associated with productivity, namely the number of nodes in the talk network, how long ago the wiki was founded, and the gini of edits per editor.

Because interpreting these results can be difficult, table 2.1 visually depicts the relationships between communication structure and productivity using actual networks as examples. The first column shows a visualization of the communication network of a wiki at the first

quartile of each measure, the second shows a network at the third quartile, and the third column provides the bootstrapped 95% confidence intervals of the model-predicted change in productivity when moving from the first to the third quartile. We see that while wikis vary considerably in their communication structure, the expected influence of these measures is slight and highly variable.

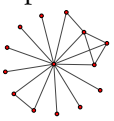
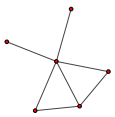
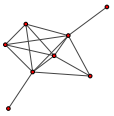
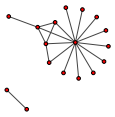
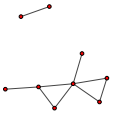
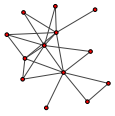
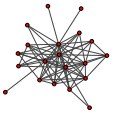
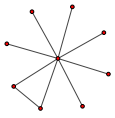
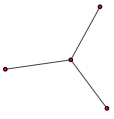
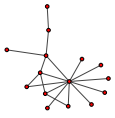
	25th percentile	75th percentile	Δ Productivity [95% CI]
Density			-1.7% – 68.0%
Hierarchy			-10.0% – 12.5%
Core Ratio			-27.6% – 8.5%
Degree Centralization			-8.8% – 7.2%
Betweenness Centralization			-11.6% – 5.1%

Table 2.1. Networks at the 1st and 3rd quartile of each measure; 95% confidence interval of predicted change in productivity when moving form 1st to 3rd quartile.

Our second set of results, predicting the longevity of a community, is presented in Figure 2.3.[5] Again, the model including network measures surprisingly does not provide a better fit

[5]A model allowing for time-varying measures produced similar results

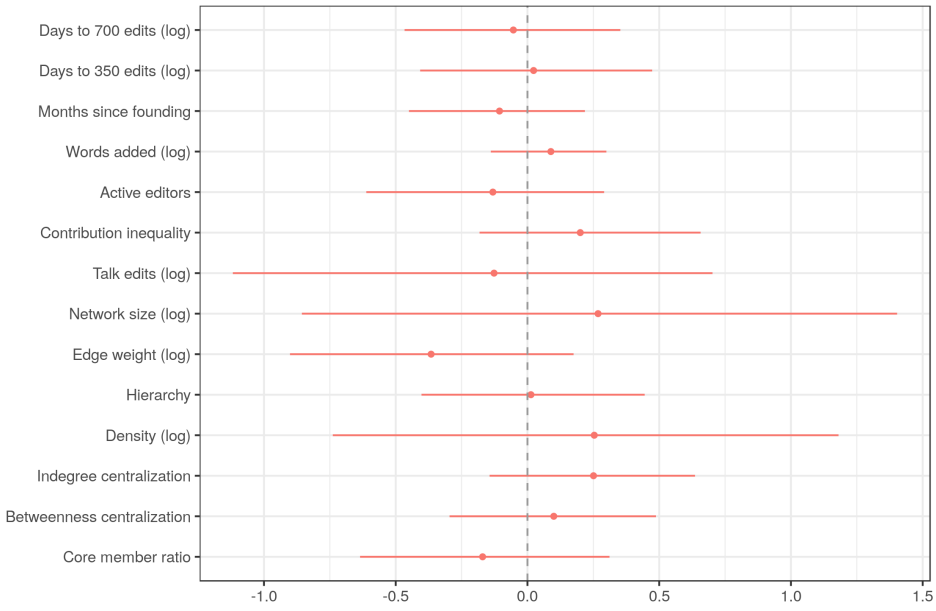


Figure 2.3. Scaled regression coefficients predicting the length of a wiki's survival. Polynomial control terms are excluded for clarity. Error bars show bootstrapped 95% confidence intervals.

than the controls-only model based on AIC and none of the network measures have a significant relationship with the likelihood of survival. Indeed, we see that the only significant predictor of survival is the number of different editors who contributed to the project.

2.6. Discussion

We hypothesized that nascent CBPP projects would benefit from structured conversation in order to coordinate their work and in order to create a long-lasting community. However, our results provide only very weak evidence that early communication structure predicts a group's productivity and no evidence that communication structure relates to longevity. These results are surprising, but our sample is large enough that we should be able to detect

even weak effects. Studies of work groups that have presented statistically significant estimates of the relationship between communication structure and performance have typically used datasets of less than 100 teams (Balkundi & Harrison, 2006). Even with a larger dataset and more statistical power, one explanation for a null finding might be that the relationship in peer production is simply noisier and more difficult to detect. This seems unlikely because our measures of network structure have comparable variance to similar measures from the work group literature. For example, our measure of hierarchy ($\mu = 0.47$; SD $= 0.24$) has a similar distribution to hierarchy measured by Cummings and Cross (2003) in work groups ($\mu = 0.57$; SD $= 0.28$) and our measure of density ($\mu = 0.24$; SD $= 0.14$) is smaller in this context but has similar variance to the density measured by Wise (2014) in work groups ($\mu = 0.34$; SD $= 0.23$). As a result, we believe that our sample of over $1,000$ projects should have been more than adequate to identify relationships between network structures and group outcomes similar in magnitude to those found in previous studies of work groups and that any relationships that exist in our data are likely much weaker.

Overall, these findings suggest two surprising features of early-stage CBPP projects: First, they do not require structured communication in order to coordinate their work, and second, social integration does not increase project-level productivity or longevity. This is different from finding that CBPP projects are unstructured: there was a large amount of variation along each of our measures of network structure. What we found was that there is no strong relationship between structural measures and either of our outcomes. We discuss possible explanations for these findings below.

Figure 2.4. User DolAmroth123 moves information into a pre-defined template. This shows other users where and how information should be organized on this wiki without any explicit coordination or communication.

2.6.1. Stigmergic coordination

CBPP projects and wikis in particular may require less communication in order to coordinate for a few reasons. The first is that work is organized around a shared, continuously updated artifact. The pages on Wikipedia, for example, are updated as soon as an editor makes an edit. This affordance allows CBPP contributors to use the artifact itself to perform "stigmergic" coordination. Stigmergy is a concept that comes from research on social insects like ants and refers to coordination that happens through modifying the environment (e.g., an ant's pheromone trail) rather than through direct communication. Bolici et al. (2016) describe how this concept can apply to CBPP. For example, the state of the artifact can act as a signal for resource matching. Those with special knowledge or skills don't necessarily need to communicate in order to be told about a need – they can simply see a need and contribute (Benkler, 2006; Kane & Ransbotham, 2016).

More directly, users can leave explicit or implicit signals of needs in the artifact itself (D. Hinds & Lee, 2009). Changes to wikis can coordinate work, negotiate norms, and integrate newcomers. Some edits are explicitly stigmergic in that they provide trails for others to follow. For example, by creating a link to a non-existing page, a contributor can signal the need for someone to create that page. Figure 2.4 shows a more subtle example of how edits can coordinate and provide norm instruction. Wikia provides this interface, which shows how a page has changed over time. In this example, the user *DolAmroth123* has taken the "(aged 41)" content added by *Roberto4554* and moved it into its own spot in a pre-defined template. Without explicit communication, this edit by *DolAmroth123* both enforces a norm about how information should be organized and teaches *Roberto4554* (and others who see the page) about how to use the template feature of the website.

Our findings about the relative unimportance of structured communication, combined with theories of stigmergic communication and coordination, suggest that there may exist a tradeoff between social structure and project structure. When the structure of a project is explicit and the goals are well-defined, as in many early-stage CBPP projects, then there are few social interdependencies. Many simple coordination tasks can be performed through the wiki itself and thus do not require complex social structures. This theory suggests an explanation for findings in the CBPP literature that projects tend to become more structured and hierarchical over time (Halfaker et al., 2013; Shaw & Hill, 2014; TeBlunthuis et al., 2018). In contrast with work groups, the work of a typical CBPP project is fairly well-delineated in early stages and higher-level coordination and decision making about the purpose or goals of the organization may simply not be needed. As projects grow and become more complex, it is more difficult to signal needs through the artifact and structured coordination is needed.

2.6.2. Limits of social integration

We hypothesized that projects with highly integrative networks would encourage high levels of contributions and that communities with high hierarchy, for example, would not be as productive. However, we found no relationship between these measures and productivity. Even more surprisingly, we find no evidence that communication structure has any relationship to a community's longevity. This suggests that either social integration does not matter very much in the context of peer production or that wiki talk page networks are not capturing the socialization and socializing that is happening. For example, users might be socializing face-to-face or using other online platforms. While we cannot dismiss the latter out of hand, it is evident that talk pages are being used by communities for communication and coordination. Looking at our data, we find that communities make a median of 89 edits to talk pages. This is a large enough number of contributions to suggest that a substantial portion of communication happens through these channels.

How can we reconcile findings about the importance of social integration in work groups (Cummings & Cross, 2003) and large peer production projects (Bryant et al., 2005; Halfaker et al., 2013) with our finding that communication structure does not help to predict nascent CBPP project survival? One possible explanation is that the effect of social integration increases over time. In larger projects like Wikipedia new contributors are less motivated by social motivations (Bryant et al., 2005). In a new wiki project everyone is a new contributor, so early social structure may not matter much.

Another interpretation is that socialization is less important in voluntary, non-professional communities like CBPP. If contributors are committed to the goals of the project, then the

marginal influence of social integration will be low. In short, in work group settings strong social cohesion may be needed to encourage everyone to contribute. However, when everyone is already eager to contribute, social motivation may be less important.

And what about the theory that early groups would need to communicate more in order to create norms and goals? Our results suggest that in CBPP this sort of negotiation may be unnecessary. On Wikia, as in many other CBPP contexts, new communities are either explicitly or implicitly part of a larger ecosystem. It is likely that new communities start with a set of default norms, based on acceptable behavior in related projects. In addition, wikis have a fairly explicit, visible, and unchangeable goal from the moment they are created. For example, the wiki at lordoftherings.wikia.com will be focused around Lord of the Rings content; there is very little room for negotiation of high-level goals. Finally, unlike in a work team where participants have contractual obligations to the organization, the cost of exit from peer production is incredibly low. Instead of "storming" and "norming" (Tuckman, 1965), those who disagree with the goals of a wiki may be more likely to create a new project or simply not edit at all. By changing the relative cost of leaving, CBPP may encourage people to choose "exit" rather than "voice" (Hirschman, 1970). A complementary implication is that the affordances of CBPP may make it difficult for projects to succeed when goals are not clear and must be negotiated. Indeed, Hill (2013) argues that a well-defined goal helped Wikipedia succeed where other attempts to build an online encyclopedia fell short.

Overall, our results do not support our hypotheses and contradict the theoretical predictions we derived from prior literature. Instead of telling the story of work groups operating in a new environment, our findings suggest that early-stage CBPP projects operate very differently. While we believe that stigmergy and a highly motivated set of participants are plausible

explanations for our findings, more empirical work is needed to understand why structures of communication in early-stage wikis do not predict project-level outcomes.

2.7. Threats to Validity & Limitations

While our study benefits from using a large population of wikis across a number of different domains, it is still possible that our findings may reflect idiosyncrasies of the software platforms and user interface configurations of Wikia wikis. Other online communities have different goals, norms, and affordances. For example, wiki talk pages are hard to find and use. In some cases, users may not even be aware of all messages added to the site by others and so some ties may not represent direct communication. Interfaces that make communication more visible and intuitive might lead to different results.

Additionally, the analysis presented here is based on digital trace data taken from a single platform and is subject to the sorts of caveats that apply to other similar studies. For example, we do not capture interactions that occur "off-wiki" either in face-to-face meetings, or through other technological systems. A related problem is that Wikia is a pseudonymous system. The same individual might edit from multiple accounts or sometimes edit anonymously but we treat each of these separate accounts as a different individual – potentially undermining the precision of several of our measures.

Finally, this paper is a cross-sectional look at wikis at one point in their lifecycle. Many theories and findings about groups suggest that they adapt and change over time (including our finding that nascent groups do not resemble large CBPP projects). Future work should investigate the ways that communication networks evolve as communities grow and age.

2.8. Conclusion

In a large-scale study of over 1,000 online wikis engaged in peer production, we find that communication network structures are poor predictors of both community productivity and longevity. A number of theories and empirical research papers suggest that these sorts of structures should support effective groups in any context. Our findings run contrary to this earlier research and call into question whether integrative social structures are always necessary for collaboration.

Our findings suggest that the relationship of communication structure to effective collaboration and organization is not universal but contingent. While all groups require coordination and undergo social influence, groups composed of different types of people or working in different technological contexts may have different communicative needs. We suggest two primary explanations. First, volunteers who already have an interest in a project may require less social integration than employees. Second, the technical affordances and narrow scope of wikis allow participants to coordinate via stigmergic clues in the artifact itself.

While the presence of stigmergic coordination in CBPP projects is fairly well-known, our results suggest that affordances which allow for stigmergic coordination might reduce the need for social structures and can influence the ways that groups interact. When project goals are clear and narrow, a shared artifact can act as a tool for communicating and negotiating norms and priorities, reducing the need for explicit communication and coordination.

2.9. Acknowledgements

This work relied on use of the University of Washington's Hyak computing cluster and was supported by NSF grants IIS-1617468 and IIS-1617129. Earlier versions of this paper received very helpful feedback from participants at the INSNA Sunbelt, American Sociological Association, and the International Communication Association conferences.

2.10. Appendix

This appendix provides additional information on how our measures were created, summary statistics, and full regression models for the models used in the paper as well as models created at 500 edits and 900 edits.

2.10.1. Measures

The following table gives a description of each measure. We then describe a few measures in more technical detail than allowed for in the text.

2.10.1.1. Gini coefficient. We use gini coefficients to measure both of our *centralization* measures as well as the control for *contributor inequality*. The gini coefficient is a measure of the inequality of a distribution, and is defined as

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n}x_i}$$

where n is the number of people, and x_i and x_j are the measures (betweenness centrality, indegree, or contribution amount) for group members i and j.

Dependent Variables	
Measure	**Description**
Productivity	Total non-reverted words added
Survival	Number of months following the 700th edit that a wiki remains active
Network Measures	
Density	Proportion of possible ties that exist
Core member ratio	Ratio of editors with coreness greater than 2
Hierarchy	Ratio of paths which are reciprocal
Indegree centralization	Gini of incoming edges per node
Betweenness centralization	Gini of betweenness centrality
Control Variables	
Total Editors	Logged number of people with at least one edit
Active editors	Number of people with at least 10 edits
Days to 350 (700) edits	Logged days from the first edit to the 350th (700th) edit
Contribution inequality	Gini of contributions per person
Talk edits	Total number of edits to talk pages
Months since founding	Number of months ago that the wiki was founded
Network size	Logged number of people in the network
Edge weight	Mean weight of ties in a network

Table 2.2. The first column lists each measure used in our models. The second column provides a descriptions of how each measure was calculated.

2.10.1.2. Core member ratio. To calculate the core member ratio, we take advantage of the social network analysis concept of *coreness* and *k-cores* as initially developed by Seidman (1983). The *k*-core of a network is the maximal subgraph of that network such that all of the nodes in the network have at least *k* connections to other nodes. For example, a 2-core is the largest possible set of nodes such that all of the nodes are connected to at least two other nodes in that set (Figure 2.5). For each node, their coreness number is the maximum *k* for which they

are part of that *k*-core. The core member ratio reported in the text is the proportion of nodes which had a coreness measure of at least 3.

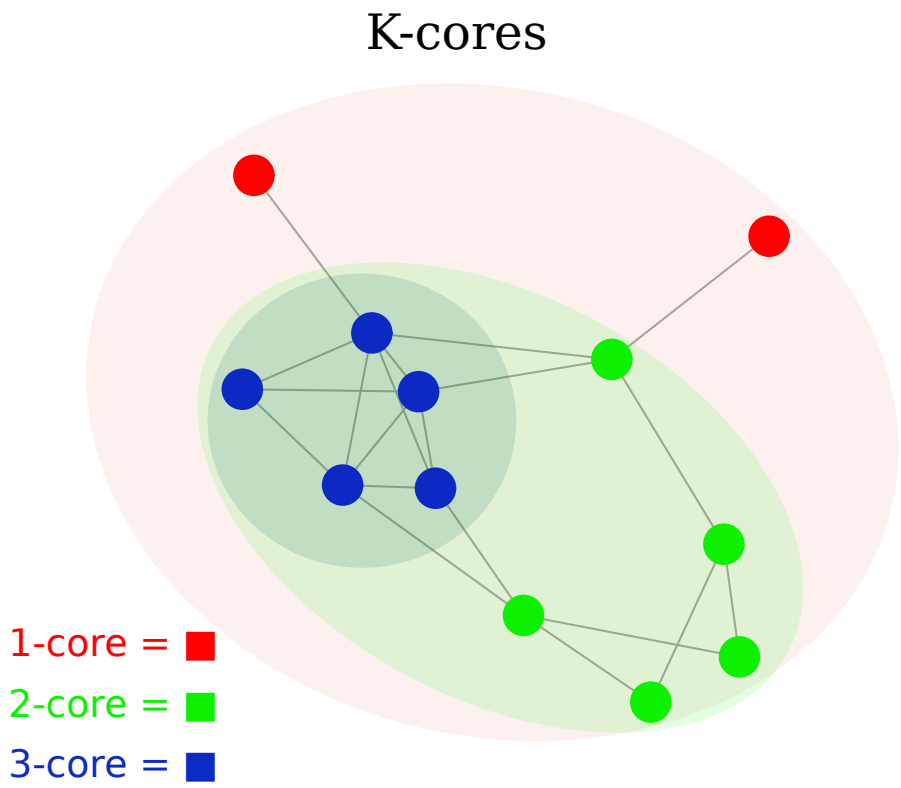


Figure 2.5. K-cores of a random graph. Each colored subgraph represents a k-core and the coloring of each node is its coreness measure (the highest *k* for which it is part of that k-core)

2.10.2. Summary Statistics and Correlations

Table 2.3 shows summary statistics for each of our variables. Figure 2.6 shows correlations between variables.

Table 2.3. Summary statistics for each of our measures. We use logged versions of highly skewed measures.

Statistic	Mean	St. Dev.	Min	Median	Max
Months survived	15.59	14.52	0.00	10.40	64.53
Words added (thousands)	239.69	234.61	13.67	171.18	3,690.69
Days to 700 edits (log)	5.10	1.36	0.00	5.37	7.36
Days to 350 edits (log)	3.63	1.40	0.00	3.74	6.73
Total editors (log)	4.52	0.79	2.08	4.58	9.08
Months since founding	33.10	16.86	0.60	30.48	73.07
Active editors	9.59	4.14	1	9	27
Contribution inequality	0.87	0.07	0.41	0.88	0.98
Talk edits (log)	4.56	0.86	2.40	4.50	8.84
Network size (log)	2.42	0.53	1.61	2.40	4.44
Edge weight (log)	1.63	0.39	1.10	1.54	4.24
Hierarchy	0.47	0.24	0.00	0.45	1.00
Density (log)	−1.59	0.55	−3.83	−1.59	−0.05
Indegree centralization	0.44	0.15	0.00	0.44	0.82
Betweenness centralization	0.80	0.10	0.38	0.81	0.97
Core member ratio	0.39	0.29	0.00	0.44	1.00

2.10.3. Regression results

Tables 2.4 and 2.5 show full regression results for the models used in the paper. Note that AIC increases in both models when network measures are added, indicating that adding this set of measures does not improve model fit.

2.10.4. Robustness results

As robustness checks, we include regression results when using a cutoff of 500 and 900 edits in tables 2.6 to 2.9. The overall results are very similar to our primary results at 700 edits. One difference is that the relationship between density and productivity is far from being significant

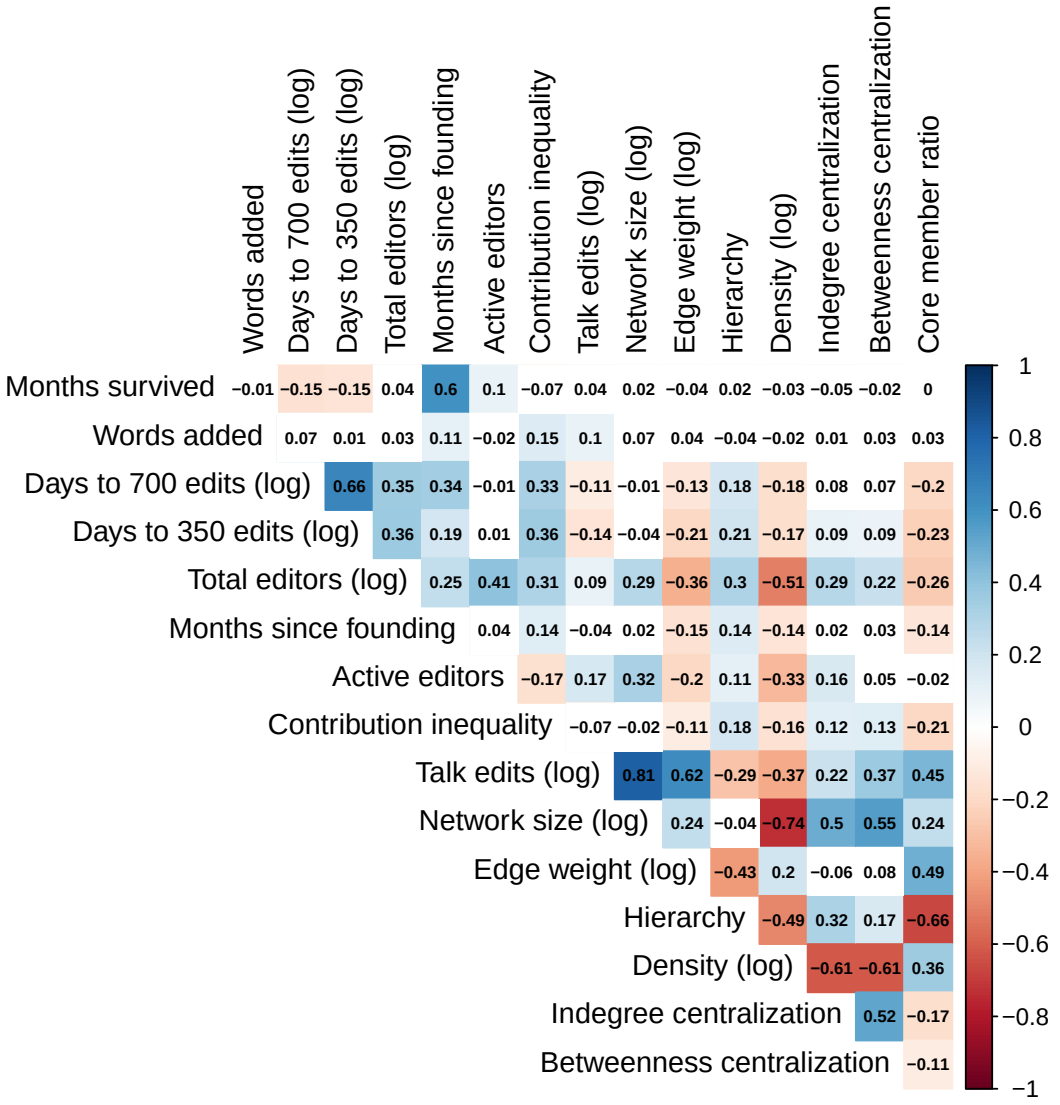


Figure 2.6. Correlation coefficients for each of the variables

at 500 edits but is significant (and stronger) at 900 edits. This provides some evidence of a change in the importance of density as a predictor for productivity as a project grows.

	Controls	Network Measures
Intercept	12.031*	12.219*
	[11.515; 12.545]	[11.446; 13.014]
Days to 700 edits (log)	0.033	0.039
	[−0.031; 0.104]	[−0.025; 0.111]
Days to 350 edits (log)	−0.027	−0.026
	[−0.110; 0.035]	[−0.101; 0.034]
Total editors (log)	−0.086	−0.067
	[−0.188; 0.052]	[−0.181; 0.086]
Active editors	0.005	0.002
	[−0.013; 0.022]	[−0.016; 0.019]
Contribution inequality	4.570*	4.798*
	[1.881; 6.497]	[2.134; 6.771]
Contribution inequality^2	4.259*	4.493*
	[2.289; 6.757]	[2.580; 7.021]
Months since founding	2.554*	2.552*
	[0.380; 4.536]	[0.459; 4.530]
Months since founding^2	1.926*	1.577
	[0.152; 4.247]	[−0.096; 3.740]
Talk edits (log)	0.130*	0.149
	[0.062; 0.192]	[−0.038; 0.316]
Network size (log)		0.281
		[−0.109; 0.703]
Edge weight (log)		−0.162
		[−0.382; 0.073]
Hierarchy		0.012
		[−0.314; 0.361]
Density (log)		0.323
		[−0.022; 0.687]
Indegree centralization		−0.012
		[−0.469; 0.366]
Betweenness centralization		−0.253
		[−1.052; 0.435]
Core member ratio		−0.188
		[−0.529; 0.137]
AIC	26501.883	26502.283
BIC	26555.890	26590.658
Log Likelihood	−13239.942	−13233.141
Deviance	1076.480	1075.633
Num. obs.	1002	1002

* 0 outside the confidence interval

Table 2.4. Negative binomial regression predicting productivity at 700 edits. Model 1 includes only controls, Model 2 adds communication network measures. Bootstrapped 95% confidence intervals are in brackets.

	Controls	Network Measures
Days to 700 edits (log)	−0.033	−0.020
	[−0.176; 0.105]	[−0.172; 0.130]
Days to 350 edits (log)	0.015	0.008
	[−0.140; 0.176]	[−0.145; 0.169]
Months since founding	−0.002	−0.003
	[−0.012; 0.007]	[−0.013; 0.006]
Words added (log)	0.072	0.089
	[−0.153; 0.279]	[−0.139; 0.300]
Active editors	−0.008	−0.016
	[−0.065; 0.041]	[−0.074; 0.035]
Total editors (log)	−2.335*	−2.451*
	[−3.467; −0.986]	[−3.598; −1.177]
Total editors (log)^2	0.205*	0.202*
	[0.060; 0.326]	[0.062; 0.329]
Contribution inequality	1.652	1.508
	[−1.322; 4.982]	[−1.368; 4.942]
Talk edits (log)	−0.182*	−0.074
	[−0.353; −0.014]	[−0.652; 0.409]
Network size (log)		0.253
		[−0.808; 1.324]
Edge weight (log)		−0.464
		[−1.145; 0.222]
Hierarchy		0.028
		[−0.856; 0.946]
Density (log)		0.230
		[−0.670; 1.070]
Indegree centralization		0.830
		[−0.476; 2.107]
Betweenness centralization		0.523
		[−1.548; 2.553]
Core member ratio		−0.295
		[−1.101; 0.540]
AIC	2940.785	2944.548
R^2	0.069	0.079
Max. R^2	0.950	0.950
Num. events	235	235
Num. obs.	1002	1002
Missings	0	0
PH test	0.003	0.000

* 0 outside the confidence interval

Table 2.5. Cox proportional hazard results predicting hazards at 700 edits. Model 1 includes only controls, Model 2 adds communication network measures. Bootstrapped 95% confidence intervals are in brackets.

	Controls	Network Measures
Intercept	11.861*	12.463*
	[11.348; 12.382]	[11.591; 13.292]
Days to 500 edits (log)	0.074*	0.079*
	[0.013; 0.151]	[0.020; 0.158]
Days to 250 edits (log)	−0.020	−0.028
	[−0.104; 0.047]	[−0.107; 0.038]
Total editors (log)	−0.149*	−0.141*
	[−0.271; −0.034]	[−0.264; −0.018]
Active editors	0.008	0.003
	[−0.017; 0.034]	[−0.022; 0.030]
Contribution inequality	4.964*	5.029*
	[2.405; 7.672]	[2.241; 7.616]
Contribution inequality^2	6.107*	6.228*
	[3.655; 8.973]	[3.900; 9.054]
Months since founding	2.124*	2.148*
	[0.043; 4.652]	[0.067; 4.686]
Months since founding^2	2.085	1.933*
	[−0.017; 5.172]	[0.032; 4.971]
Talk edits (log)	0.122*	0.109
	[0.050; 0.206]	[−0.097; 0.296]
Network size (log)		0.162
		[−0.283; 0.669]
Edge weight (log)		−0.057
		[−0.300; 0.191]
Hierarchy		−0.209
		[−0.555; 0.153]
Density (log)		0.038
		[−0.351; 0.457]
Indegree centralization		0.060
		[−0.422; 0.484]
Betweenness centralization		−0.801
		[−1.737; 0.068]
Core member ratio		−0.170
		[−0.568; 0.212]
AIC	27466.344	27468.412
BIC	27520.919	27557.715
Log Likelihood	−13722.172	−13716.206
Deviance	1144.099	1143.276
Num. obs.	1055	1055

* 0 outside the confidence interval

Table 2.6. Negative binomial regression predicting productivity at 500 edits. Model 1 includes only controls, Model 2 adds communication network measures. Bootstrapped 95% confidence intervals are in brackets.

	Controls	Network Measures
Days to 500 edits (log)	−0.018	−0.003
	[−0.145; 0.108]	[−0.128; 0.123]
Days to 250 edits (log)	0.117	0.112
	[−0.020; 0.245]	[−0.029; 0.237]
Months since founding	0.007	0.007
	[−0.001; 0.016]	[−0.001; 0.016]
Words added (log)	0.015	0.010
	[−0.196; 0.210]	[−0.200; 0.213]
Active editors	−0.031	−0.043
	[−0.091; 0.027]	[−0.106; 0.021]
Total editors (log)	−1.781*	−2.015*
	[−2.918; −0.232]	[−3.227; −0.500]
Total editors (log)^2	0.142	0.152
	[−0.047; 0.249]	[−0.022; 0.272]
Contribution inequality	0.829	0.869
	[−1.177; 3.337]	[−1.222; 3.588]
Talk edits (log)	−0.144	0.017
	[−0.315; 0.021]	[−0.459; 0.438]
Network size (log)		0.197
		[−0.829; 1.192]
Edge weight (log)		−0.633*
		[−1.248; −0.010]
Hierarchy		−0.515
		[−1.326; 0.300]
Density (log)		0.183
		[−0.654; 0.987]
Indegree centralization		0.959
		[−0.041; 2.020]
Betweenness centralization		0.014
		[−1.648; 1.718]
Core member ratio		−0.387
		[−1.115; 0.452]
AIC	3548.740	3549.444
R^2	0.064	0.076
Max. R^2	0.967	0.967
Num. events	280	280
Num. obs.	1055	1055
Missings	0	0
PH test	0.000	0.000

* 0 outside the confidence interval

Table 2.7. Cox proportional hazard results predicting hazards at 500 edits. Model 1 includes only controls, Model 2 adds communication network measures. Bootstrapped 95% confidence intervals are in brackets.

	Controls	Network Measures
Intercept	12.348*	12.216*
	[11.826; 12.899]	[11.449; 12.938]
Days to 900 edits (log)	0.035	0.040
	[−0.029; 0.112]	[−0.024; 0.114]
Days to 450 edits (log)	−0.005	−0.002
	[−0.099; 0.052]	[−0.080; 0.052]
Total editors (log)	−0.128	−0.099
	[−0.231; 0.006]	[−0.213; 0.061]
Active editors	0.010	0.009
	[−0.007; 0.024]	[−0.008; 0.024]
Contribution inequality	3.976*	4.139*
	[1.415; 6.010]	[1.374; 6.156]
Contribution inequality^2	5.014*	5.268*
	[3.128; 7.605]	[3.408; 7.804]
Months since founding	2.802*	2.828*
	[0.766; 4.800]	[0.956; 4.824]
Months since founding^2	1.628	1.239
	[−0.088; 4.107]	[−0.391; 3.434]
Talk edits (log)	0.114*	0.112
	[0.039; 0.182]	[−0.081; 0.274]
Network size (log)		0.413*
		[0.031; 0.862]
Edge weight (log)		−0.185
		[−0.410; 0.065]
Hierarchy		0.064
		[−0.303; 0.439]
Density (log)		0.484*
		[0.117; 0.903]
Indegree centralization		−0.170
		[−0.714; 0.239]
Betweenness centralization		0.269
		[−0.503; 1.089]
Core member ratio		−0.312
		[−0.677; 0.038]
AIC	24857.478	24852.454
BIC	24910.630	24939.429
Log Likelihood	−12417.739	−12408.227
Deviance	991.555	990.434
Num. obs.	927	927

* 0 outside the confidence interval

Table 2.8. Negative binomial regression predicting productivity at 900 edits. Model 1 includes only controls, Model 2 adds communication network measures. Bootstrapped 95% confidence intervals are in brackets.

	Controls	Network Measures
Days to 900 edits (log)	−0.044	−0.030
	[−0.220; 0.123]	[−0.208; 0.151]
Days to 450 edits (log)	0.032	0.024
	[−0.160; 0.202]	[−0.181; 0.200]
Months since founding	−0.001	−0.002
	[−0.012; 0.010]	[−0.015; 0.008]
Words added (log)	−0.073	−0.068
	[−0.328; 0.203]	[−0.342; 0.215]
Active editors	−0.028	−0.029
	[−0.081; 0.020]	[−0.080; 0.022]
Total editors (log)	−2.342*	−2.295*
	[−3.512; −1.128]	[−3.405; −0.862]
Total editors (log)^2	0.204*	0.186*
	[0.074; 0.328]	[0.044; 0.298]
Contribution inequality	1.385	0.841
	[−2.248; 5.461]	[−2.898; 4.975]
Talk edits (log)	−0.151	0.391
	[−0.352; 0.027]	[−0.159; 0.877]
Network size (log)		−0.701
		[−1.943; 0.570]
Edge weight (log)		−0.474
		[−1.196; 0.289]
Hierarchy		−0.349
		[−1.494; 0.725]
Density (log)		−0.203
		[−1.281; 0.763]
Indegree centralization		0.610
		[−0.856; 2.056]
Betweenness centralization		−0.444
		[−2.800; 1.827]
Core member ratio		−0.778
		[−1.724; 0.159]
AIC	2315.564	2318.623
R^2	0.070	0.081
Max. R^2	0.922	0.922
Num. events	188	188
Num. obs.	927	927
Missings	0	0
PH test	0.028	0.112

* 0 outside the confidence interval

Table 2.9. Cox proportional hazard results predicting hazards at 900 edits. Model 1 includes only controls, Model 2 adds communication network measures. Bootstrapped 95% confidence intervals are in brackets.

CHAPTER 3

Starting Online Communities:

Motivations and Goals of Wiki Founders

Jeremy Foote
Northwestern University

Darren Gergle
Northwestern University

Aaron Shaw
Northwestern University

Why do people start new online communities? Previous research has studied what helps communities to grow and what motivates contributors, but the reasons that people create new communities in the first place remain unclear. We present the results of a survey of over 300 founders of new communities on the online wiki hosting site Wikia.com. We analyze the motivations and goals of wiki creators, finding that founders have diverse reasons for starting wikis and diverse ways of defining their success. Many founders see their communities as occupying narrow topics, and neither seek nor expect a large group of contributors. We also find that founders with differing goals approach community building differently. We argue that community platform

designers can create interfaces that support the diverse goals of founders more effectively.

3.1. Introduction

The people who found new organizations are in a unique position to exert influence on the way that the organization develops. For example, research on firms has shown that the attributes of a company's founder, such as their education, industry experience, and position in a social network, can influence the firm's growth and survival (Cooper, Gimeno-Gascon, & Woo, 1994; Ostgaard & Birley, 1996). Founders can also influence the fate of online communities. Communities which have active, well-connected, experienced founders are more likely to experience continued activity (Kraut & Fiore, 2014).

Prior research on firm and community founders makes an implicit assumption: that all founders are hoping for their organizations to grow and survive. This may be a reasonable assumption for entrepreneurs, but online community founders are likely to have more varied motivations and project goals. Learning more about founders' motivations and perspectives on community success is important for understanding how this new form of organizaing works.

To learn more about why founders start communities, we surveyed users of the online community hosting site Wikia.com immediately after they founded a new wiki. We focus on founders' *motivations* for founding new wikis and their *project goals*. Motivations encompass more immediate incentives in comparison to project goals, which are more forward-looking and relate to how founders will evaluate the success of their community.

While some founders seek to build large, productive online communities, only 18% of our respondents rank content growth as the top goal. The most common primary goal is the creation of high-quality information. Overall, motivations and goals are very diverse, but many point to the creation of communities which are modest in scope and in aim. Wikis are often created on a whim, with nearly half of respondents reporting that they had considered starting their wiki for only a few hours or less before founding it. The primary goal that a founder has for a project is related to how and whether they engage in community building activities, suggesting a path whereby community outcomes may relate to founder goals.

Our results suggest that online community researchers and community platform designers can do more to understand and support attempts to build smaller and shorter-term communities, and that focusing on growth or longevity creates a limited view of success in online communities.

3.2. Background

Prior research on motivations in online communities has focused primarily on *contributors*, finding that online community participants have heterogeneous motivations. For example, Wikipedia editors have motivations like enjoyment, learning, and ideology (Nov, 2007; Schroer & Hertel, 2009). Similarly diverse motivations have been found for contributors to open source software (Hertel, Niedner, & Herrmann, 2003) and other online communities (e.g., Lampe, Wash, Velasquez, & Ozkaya, 2010). We expect that community founders also have diverse motivations, but lack empirical evidence of what those motivations are.

Existing research also suggests diversity in how community members assess whether an existing community is successful. For example, open source programmers use many different

measures to determine success, such as user satisfaction, developer satisfaction, code quality, and popularity (Crowston, Annabi, & Howison, 2003; Schweik & English, 2012). Just as contributors see current success differently, we anticipate that founders will have different understandings of future success.

We might also expect to see diversity in founder goals due to different levels of experience on a specific platform. As users become more familiar with a tool or ecosystem of projects (e.g., Wikipedia, Wikia, GitHub), they may recognize opportunities for community and interaction that are not visible at first (Bryant et al., 2005). As users gain experience and their perspective of the existing projects in a particular domain changes, their goals for new communities may also change.

Differences in the motivations and goals of founders may also lead to different community outcomes. For example, the way participants in a community understand a project helps explain their subsequent levels of participation (Antin, Cheshire, & Nov, 2012; von Krogh et al., 2012), and greater founder activity is associated with community survival (Kraut & Fiore, 2014).

Founders' approach to community building presents one mechanism that could produce divergent outcomes. The people in charge of a community can take practical steps to help the community grow and be productive, such as attracting newcomers, encouraging contributions, and regulating bad behavior (Kraut et al., 2012). The way that founders apply these tactics (or whether they apply them at all) may depend on the goals they have for the community. For example, founders who create a community as a place to communicate with a small group of friends will be unlikely to take steps to attract newcomers, but may have plans for encouraging contributions.

3.3. Method

To better understand community founder motivations, we conducted a survey of founders on the Wikia platform. We chose Wikia as it is a widely used platform for the creation of topic-based wikis. Users on Wikia can start new wikis about any topic they choose, and users are free to edit any wikis on the site. Wikis represent a distinct type of online community, based on the creation and maintenance of a shared knowledge artifact. There are, of course, many other types of online communities, whose founders may have a different set of motivations and goals.

We used an online survey to gather data from founders because it facilitated the collection of a broad range and large number of responses. We worked with representatives at Wikia to distribute the survey to founders of new communities. After users initiate and complete the new wiki creation process, an automated message is posted on the user's "Message Wall" (a threaded discussion interface for user-to-user interactions). By default, messages posted on a user's Message Wall are also emailed to them. During our study, an invitation to participate in our online survey was included at the bottom of this automated message. The data included in this paper is from surveys taken between 5 April and 31 August 2016. By surveying founders immediately after they created a new wiki, we capture the motivations and goals they had at the time of founding, before any community outcomes could distort their memory of events.

The survey consisted of 70 items, including items about demographic background, Internet use skills,[1] and past experience on Wikia and other online communities. Our items about motivations included an open-ended question ("What made you want to start this wiki?") and

[1]We used a modified version of the 10 item instrument developed by Hargittai and Hsieh (2012).

Topic	Survey Items
Motivation	"What made you want to start this wiki?" "On a scale of 1-5, how much did each of these motivations contribute to your decision to start the wiki? 'Playing around / Learning how the software works,' 'As a joke,' 'A place to organize personal material,' 'To communicate with friends,' 'To provide publicity for myself or my company,' 'A new product was released that I was interested in,' 'There wasn't an existing wiki about the topic,' 'Exisiting wikis about the topic were of poor quality,' 'I wanted to spread information about the topic,' 'I disagreed with how existing wikis were being run,' 'I wanted to build a community,' 'I wanted to be more involved in governing the conversation around this topic,' 'Other (please explain)' "
Project Goals	"Think about how you will assess whether or not your wiki is successful, and rank which of the following are the most important measures of success for you: 'A large number of contributors', 'A large amount of information about the topic,' 'High quality information about the topic,' 'A community that remains active for a long time,' 'A highly active community,' 'Meeting new people,' 'Other (please explain)' "

Table 3.1. Key survey topics and abbreviated questions

a set of thirteen Likert-scale items based on the contributor motivation literature.[2]. To discover goals, we asked participants to think about how they would determine whether their wiki is successful, and to rank the importance of seven success metrics. These included both instrumental and terminal goals, which were drawn from prior literature (e.g., Crowston et al., 2003; Schweik & English, 2012) and expanded to incorporate additional community-oriented and interaction-oriented goals.[3] For both motivations and goals, we allowed respondents to enter free text "Other" responses. The first author reviewed the text responses and selected characteristic examples that we report below. Finally, to determine a founder's plans for community building, we asked them whether they would implement community building strategies suggested by Kraut et al. (2012).

[2]Items were ranked from 1 ("Not a motivation") to 5 ("A primary motivation") If a respondent rated more than half of the motivation items then we imputed "Not a motivation" for any missing items.

[3]We started with seven project goals (see Table 3.1), and collapsed two items representing facets of community and two measures representing facets of growth into one measure each.

3.4. Results

During the period of data collection, 46,828 total new wikis were created by 35,749 users. A total of 720 respondents started the survey. We removed 91 respondents who failed our attentiveness check, were obviously fraudulent,[4] or who reported being minors. This left us with 629 respondents. Respondents were allowed to skip items they did not want to answer. For example, 312 founders filled out the motivation section and 254 ranked their goals. In order to maximize the number of participants included in each analysis, we used responses that were complete for that portion, even if the respondent skipped other portions. We report the *N* for each analysis in the table captions.

3.4.1. Founding Motivations

To simplify the analysis of motivations, we reduced the dimensionality of our thirteen motivation items. Factor analysis suggested that there were five dimensions:[5] spreading information and building community, problems with existing wikis, fun and learning, creating and publicizing personal content, and the "Other" item. We summarize the distribution of responses in Table 3.2. The factor representing the desire to build a community and to spread information about a topic has the highest overall popularity, although each of the motivations was identified as a primary motivation by a number of respondents.

While some additional motivations appeared in the free-text responses, none of these was reported by more than a few users. This suggests that our original items represent typical

[4]For example, those who reported having been born before 1915.

[5]A KMO test (Kaiser & Rice, 1974) showed that exploratory factor analysis was appropriate (MSA=.69), and a scree plot suggested four factors was optimal. All of the items loaded onto at least one factor with a cutoff of .3, with the exception of the "Other" item, which we treat as its own factor. For cross-loaded items, we included the item in the primary factor, all of which loaded at .4 or higher. Combined, the factors explained 34% of the variance.

	Mean	SD	Primary Motivation
Info and Community	3.12	1.04	70
Bad Existing Wikis	1.89	1.28	28
Fun and Learning	2.32	0.96	15
Personal Material	2.71	1.26	43
Other	1.84	1.59	60

Table 3.2. Distribution of motivations. The final column shows how many respondents chose that as a primary motivation (mean rating > 4), N=312.

motivations of respondents. That said, some of the new motivations identified are interesting. A few respondents wrote about teaching others about a topic. For example, one wrote, "I am a huge fan of the books and I wanted to create a wiki where other fans can read about and learn about the series." Another expressed an interest in gathering feedback: "I run a gaming community that is currently in development, I want users to be [...] able to contribute and add information about [...] our game." Another reported using the wiki to organize and store content: "[I] wanted a place to organize content as [my] web serial was written." Many examples are inconsistent with the goal of building a large and active community with many contributors.

Motivations need not be particularly strong because the barrier to founding a new community is so low within a platform like Wikia. Among our respondents, only 32% had planned to start a wiki for a few weeks or longer, while 46% of respondents reported that they had only thought about it for "a few hours" or "a few minutes" beforehand. Founding new communities appears to often be impulsive and not the result of delibration or careful planning.

3.4.2. Project Goals

Next, we look at the goals that founders have for their projects. Each of the project goals was ranked as the top goal by a substantial proportion of respondents. The creation of high-quality information was the most prevalent top goal, with 47% of respondents selecting that option. The next two most popular goals, "community" and "growth", were nominated as the top goal by 20% and 18% of respondents, respectively.

While many studies focus on the longevity and activity of a community as markers of success, this finding suggests that many wiki founders care most about creating a high-quality repository of information rather than community growth or other outcomes.

Indeed, most founders do not expect that their projects will become large and popular. When we asked founders about the potential audience for their projects, 50% chose the response that "only a handful of people" would be interested in the topic of their wiki. When we asked how many contributors they expected in the first 30 days, the median response was 4.

One additional goal emerged from the "Other" free text responses. Multiple respondents expressed the desire for external usefulness and popularity, independent of community-building. For example, one respondent's goal was, "Seeing the material actually used by [...] groups and individuals." Another's was, "People that actually benefit from my information." Various other goals were expressed by one or two respondents, such as strengthening relationships.

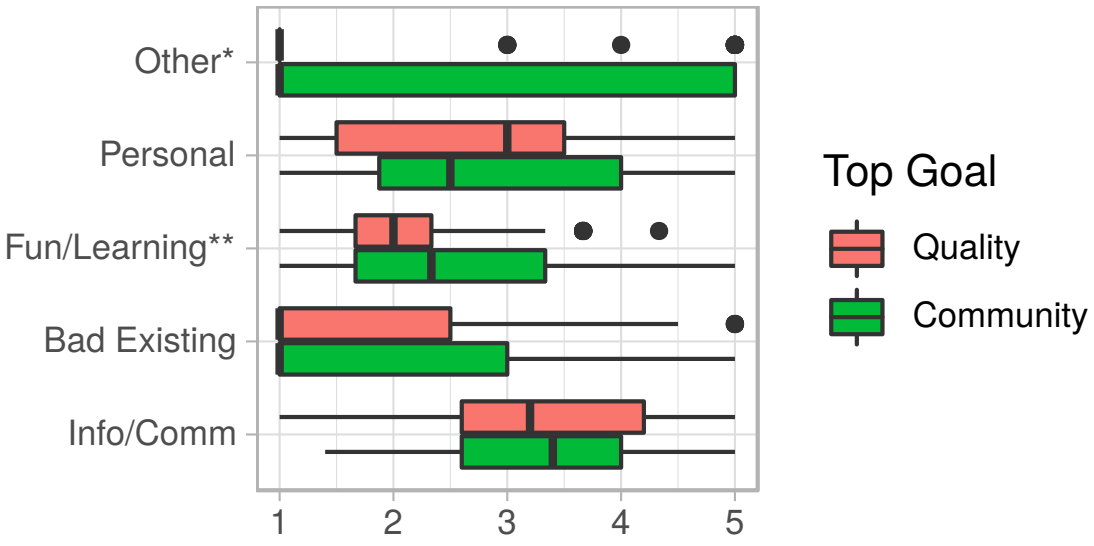


Figure 3.1. Differences in the motivations for respondents on a scale of 1 ("Not a motivation") to 5 ("A primary motivation"), based on their top goal for the wiki, N=248. * $p < .05$, ** $p < .01$

3.4.3. Comparing Founders with Different Goals

We next explore whether there are differences between founders with different project goals. We group respondents based on their top-ranked goal and compare those who selected each of the two most common top goals ("Information Quality" and "Community").

Figure 3.1 compares how motivations differ between those who chose each of the two top goals. t-tests at $\alpha < .05$ indicate that those whose top goal was community-based ("a community that remains active for a long time" or "a highly active community") were more likely to be motivated by fun and learning motivations and "other" motivations than those whose top goal was "high quality information about the topic."[6]

[6]The differences in the "Other" motivation in Figure 3.1 and the "Recurit Members" strategy in Figure 3.2 were only marginally significant after applying a Holm-Bonferroni correction and should be interpreted with caution.

These differences in goals may result from pre-existing differences between founders, such as demographic attributes or experiences. For example, perhaps long-term users value community more since they are more likely to perceive the role of community on a site (Bryant et al., 2005). However, we performed χ^2 tests and found that community-oriented founders do not differ from quality-oriented founders by gender ($p = 0.36$), employment status ($p = 0.56$), how often they visit Wikia ($p = 0.21$), or how long they have been active on the site ($p = 0.5$). A t-test of founders' self-reported technical knowledge also showed no significant difference between the groups ($p = 0.1$). Based on these basic measures, we find no relationship between a founder's attributes or experience and her project goals.

We also investigate whether having different goals changes the way that founders approach managing their community, finding mixed results. Prior research has shown that groups whose founders add content to the site and who are engaged and active are more likely to survive (Kraut & Fiore, 2014). We asked founders about the time they were planning to devote to their wiki, but a χ^2 test shows no difference between community-oriented and quality-oriented founders in the ratio of those planning to dedicate at least 3 hours per week ($p = 1$).

However, we do find differences in the specific plans founders have for community building, encouragement, and regulation (Kraut et al., 2012). As with motivations, we use t-tests to compare the responses based on whether a founder's top goal is community or information quality. The results appear in Figure 3.2.

Those who rate community their top priority are more likely to have plans to recruit members and to encourage contributions. Both groups are equally likely to plan to welcome new members and regulate behavior.

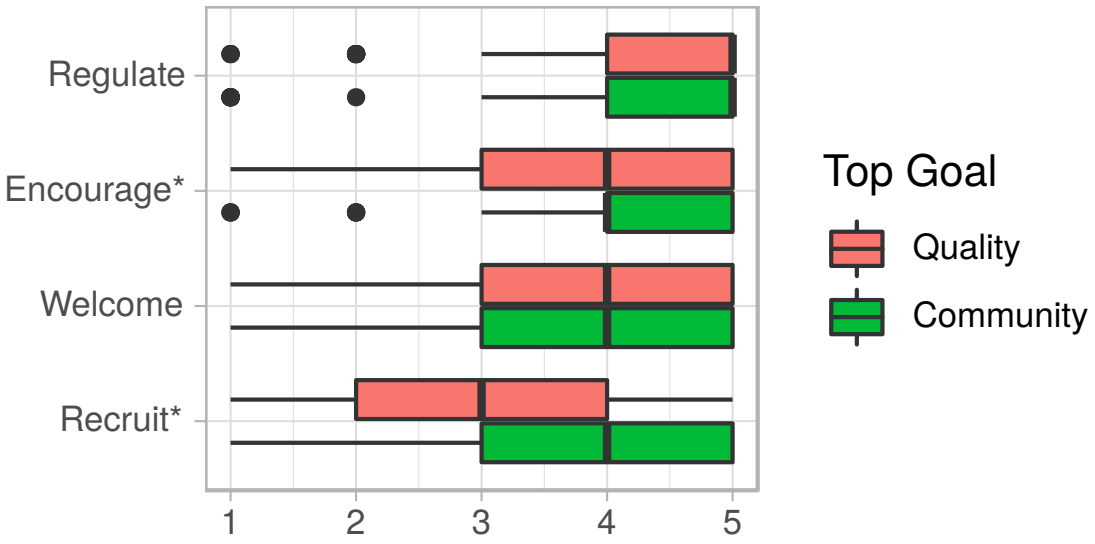


Figure 3.2. Differences in whether a respondent plans to implement each community building strategy, on a scale of 1 ("Definitely not") to 5 ("Definitely yes"), based on their primary goal for the wiki, N=251. * $p < .05$

3.5. Discussion

The results of our survey provide evidence that wiki founders have varied motivations for starting their communities and varied goals for what those communities will eventually become. Even along dimensions that many founders care about (e.g., information quality and community activity), founders differ in how they prioritize these goals. While some respondents sought to build large, popular communities, we learn that most founded their wikis with very little foresight or planning and with modest goals. Instead of aspiring to lead and influence the masses, many founders seek to collect or distribute information about niche topics among a small group.

We suggest that the prevailing focus on community growth and longevity obscures the importance of small, niche communities. Small communities are not necessarily failures or cautionary tales. They may be meeting their founders' goals and deserve to be studied on their own terms. On Wikia, where communities are built around a shared, public artifact, many founders see creating that artifact as more important than building a community. Other community platforms like reddit or Facebook Groups have their own affordances and their founders likely have a different distribution of motivations and goals. However, growth-based measures of success are unlikely to fully represent founders' objectives in any of these venues.

By learning more about the founders of communities, platform designers gain insight into how people aspire to use a system and can better support those goals. For example, designers could add dashboards to measure and display the metrics of success that founders care most about. They could also build tools to support those goals. In the case of Wikia, we find a number of founders use wikis to create and distribute personal material. Wikia might enhance the user page interface or develop templates to facilitate this behavior.

3.6. Limitations & Future Directions

Future studies can assess the generalizability of these findings on Wikia and other platforms. Running a similar study across a set of different online community platforms would allow us to learn whether founders on other platforms have similar motivations and goals. Also, as with any survey, non-response may bias our results. However, we anticipate that a more representative sample would strengthen our key findings about the modest goals of Wikia founders, since the least dedicated founders would presumably be less likely to take the survey.

This study indicates that researchers and designers can do more to attend to founders. We examined how founders' motivations and goals relate to their plans for community building, but future research should also analyze how motivations and goals relate to the actions founders take. In this way, different motivations and goals may explain variations in community outcomes. On the other hand, future research may find that founders have limited influence and communities can transcend their founders' intentions as well as their actions.

3.7. Acknowledgements & Access to Data

The authors would like to thank Trevor Bolliger at Wikia for his help deploying the survey. We are also grateful for the comments and suggestions of our reviewers. Financial support for this work came from the National Science Foundation (grant IIS-1617468) and Northwestern University. A replication dataset has been archived in the Harvard Dataverse and is available at http://dx.doi.org/10.7910/DVN/YG9IID. This paper was published by the ACM at https://doi.org/10.1145/3025453.3025639.

CHAPTER 4

The Behavior and Network Position of Peer Production Founders

Jeremy Foote
Northwestern University

Noshir Contractor
Northwestern University

Online peer production projects, such as Wikipedia and open-source software, have become important producers of cultural and technological goods. While much research has been done on the way that large existing projects work, little is known about how projects get started or who starts them. Nor is it clear how much influence founders have on the future trajectory of a community. We measure the behavior and social networks of 60,959 users on Wikia.com over a two month period. We compare the activity, local network positions, and global network positions of future founders and non-founders. We then explore the relationship between these measures and the relative growth of a founder's wikis. We suggest hypotheses for future research based on this exploratory analysis.

4.1. Introduction

The surprising success of online peer production (OPP) projects like Wikipedia and open source software has shown that groups of motivated volunteers can successfully create high-quality goods without formal hierarchy. Scholars across a number of disciplines have studied why these projects work, sparking new research on the role of firms, intellectual property rights, and individual motivations in producing shared goods (Benkler, 2006; Levine & Prietula, 2013; Raymond, 1999; von Krogh & von Hippel, 2006). In this project, we focus on one aspect of OPP projects that has escaped much scholarly attention: its founding.

Founders have been shown to be influential in the similar context of entrepreneurship. Researchers have found both that people differ in their propensity to become an entrepreneur (Backes-Gellner & Moog, 2013; Dobrev & Barnett, 2005; Lazear, 2004; Nanda & Sørensen, 2010; Zhao, Seibert, & Lumpkin, 2010) and that the attributes and experiences of founders relate to a firm's success (Cassar, 2014; Eisenhardt & Schoonhoven, 1990; Jo & Lee, 1996; Ostgaard & Birley, 1996; Stam, Arzlanian, & Elfring, 2014). There are a number of reasons to think that OPP founders differ from entrepreneurs, such as the lower costs, risks, and benefits of founding. Learning about OPP founders can help us to understand why projects grow (or don't) in this increasingly important context.

We use measures of editing behavior and measures of social capital and social integration from over 60,000 contributors to Wikia.com to explore how founders differ from non-founders as well as how founders of high-growth projects differ from founders of low-growth projects. Our exploratory results suggest that compared to non-founders, founders are typically novelty-seeking and have low social capital. However, founders who are successful at

creating larger communities have a different set of attributes: they have less diverse experience but are actually more integrated in social networks, suggesting that they may use their social capital to recruit others.

4.2. Related Work

4.2.1. Entrepreneurship

Researchers have studied both who decides to become an entrepreneur and what attributes and experiences of entrepreneurs correlate with firm success. They have found that people with diverse experience and skills are more likely to become entrepreneurs (Backes-Gellner & Moog, 2013; Lazear, 2004; Wagner, 2003), as are those who have worked with other entrepreneurs (Nanda & Sørensen, 2010). Given that someone has chosen to start a new firm, founders with more experience (Cassar, 2014; Eisenhardt & Schoonhoven, 1990; Jo & Lee, 1996) have more successful firms, as do those with larger and more diverse social networks (Stam et al., 2014).

4.2.2. Peer Production

At first blush, online peer production projects seem very dissimilar to firms. They are composed of volunteers, without formal hierarchies and without paychecks. Indeed, much research on OPP focuses on how these organizations can work using structures and incentives so different from firms (Benkler, 2016, provides a survey). Surprisingly, much of this research

has found that these supposedly decentralized, leaderless organizations are actually quite structured. For example, researchers have found that people select into different social and behavioral "roles" (like jobs) that are persistent over time (Welser et al., 2011), including leadership roles (H. Zhu et al., 2012a). In other words, OPP looks more like firms than we would expect.

We argue that at its core the decision to start a new OPP project is very similar to the decision to start a new business. There are different costs to creating each type of organization, different tools for recruiting and encouraging contributors, and different benefits that accrue to founders. However, both founders and entrepreneurs organize the efforts of a group of people to create shared outputs. Indeed, the outputs of OPP projects and firms often directly compete.

Despite the similarity in context, very little research on peer production founders exists. Survey-based research suggests that OPP founders have diverse expectations and modest goals (Foote et al., 2017). In the related context of online communities, researchers found that active and well-connected founders started longer-lived groups (Kraut & Fiore, 2014). This paper extends these earlier works by comparing founders and non-founders and by exploring the relationship between founder attributes and the growth of a community rather than its survival.

Based on the entrepreneurship research, we focus on two questions:

RQ1: How do founders differ from non-founders?

RQ2: How do founders of large and small communities differ?

4.3. Methodological Approach

Our data comes from a dump of all edits to all wikis on Wikia.com as of January 2010. While wikis are used for different purposes, many of the wikis on Wikia are used for knowledge aggregation. The most popular wikis aggregate information about popular media, such as Disney movies or the Harry Potter books. These wiki projects represent an important strand of OPP but differ markedly from other OPP projects like open source software development.

We collected data at three different points in time (Figure 4.1). We gathered behavior and network measures from the 60,959 users who made at least one edit from March to April 2009. We then identified the founders of the 16,904 wikis created in May and June 2009, and measured the number of unique contributors to each of these wikis as of January 2010. We used these data snapshots to compare the activity and network positions of those who became founders in May or June and those who did not, and to predict the relative growth of a founder's wikis between founding and January 2010.

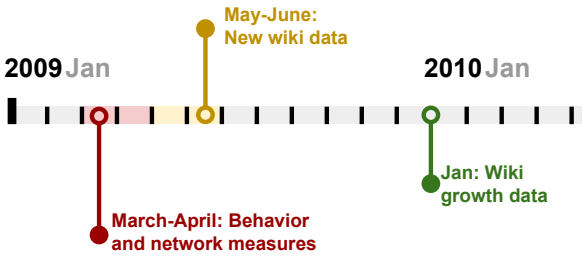


Figure 4.1. Timeline of data collection

Specifically, we measured behavior and attributes which have been found to relate to founding propensity or organization success in entrepreneurship. We created a number of *experience* and *activity* measures, such as tenure on Wikia, lifetime edits, recent edits, days since editing,

and number of days with at least one edit. For *diversity of experience*, we measured the number of wikis that a user had contributed to and the Gini coefficient of edits per wiki. Finally, we measured the user's *founding expertise* and experience by measuring how many wikis users had started in the past, how often they started new pages, how often they participated in administering wikis, and the earliest point at which they contributed to a wiki (e.g., a value of 3 means they were the third editor on a wiki).

In order to measure a user's *social capital*, we created two different types of social networks based on two kinds of activities that occur on Wikia. We used article pages to create undirected collaboration networks, where edges are formed between an editor and the previous five editors of that page.[1] Communication networks are similar but are directed: users are assumed to be talking *to* the previous five editors on a talk page or *to* the owner of a user talk page. We created unweighted networks for each wiki and created each unweighted global network by combining the wiki networks. For each of the two network types, we measured the degree centrality, betweenness centrality, and PageRank of each user. There is, of course, no direct way to measure social capital, but these are somewhat overlapping measures of how prestigious and involved a user is in the network (Wasserman & Faust, 1994). We also calculated "coreness", which is a measure of how integrated a user is (Barberá et al., 2015). We measured these values for each user in the global networks as well as in the wiki-level network for the wiki they were most active in.

We defined founders as the first three editors to a given project, since many founders start new wikis as groups (Foote et al., 2017). We measured growth as the number of unique contributors to the wiki from the time of its founding until January 2010. Because a given user

[1]Similar analyses, (e.g., Zhang & Wang, 2012) connect all editors, but a cutoff more closely approximates the social interactions we are interested in.

might start multiple wikis during our data collection, we used the median community size for all foundings that a user was a part of. We also included a control for when the median wiki was started.

Because there is so little research specifically on peer production founders we took an exploratory and hypothesis-generating approach. Our predictors had a high degree of multicollinearity, so we used elastic net regularized regression to identify candidate predictors (Zou & Hastie, 2005). We used the smaller set of predictors produced by this procedure in regression models for each of our research questions.

4.4. Results

There is one interesting result which doesn't require any modeling. We learned that founding is a rare activity for established users. Of the 60,959 users who were active in March and April 2009, only 823 founded a new wiki in May or June. That is not to say that founding is rare overall. There were 8,496 founders in May and June, meaning that nearly 90% of wikis were founded by new users.

For the rest of the analyses, we remove users with 'bot' in their username (N= 104) and low-activity users who don't appear in both the communication and collaboration networks (N = 45,775). There were also Wikia employees who we could not identify automatically, so we remove all users with edits to more than 100 different wikis (N = 24). This leaves us with a sample of 15,184 users and 470 founders.

For our first research question, we try to predict which users would become founders. Figure 4.2 shows the results of a logistic regression with the predictors identified by the elastic net

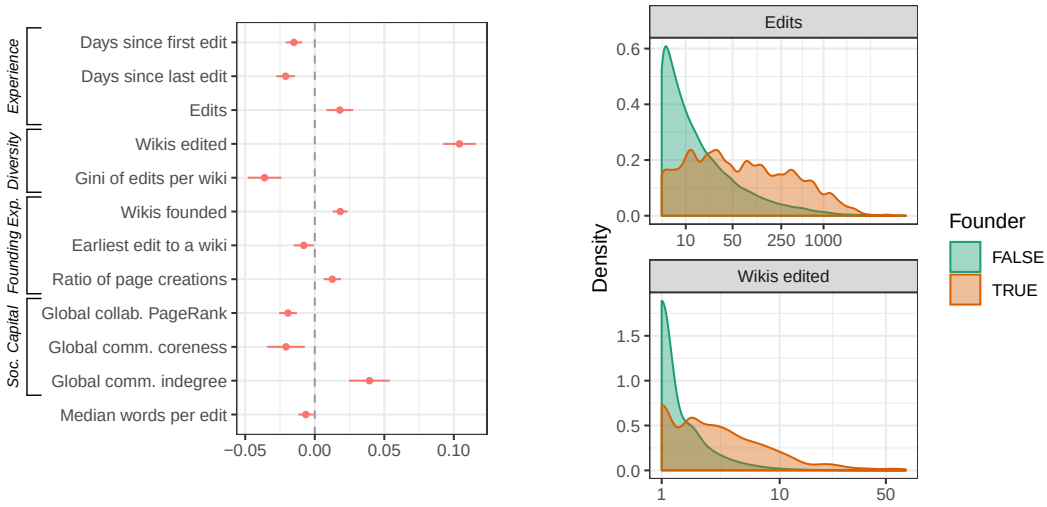


Figure 4.2. Left: Scaled coefficient estimates with 95% confidence intervals for predicting of whether a user becomes a founder ($N = 15{,}184$). Right: Density plots of the number of edits and the number of wikis edited in March and April 2009 by those who become founders in May and June 2009 (blue) and those who do not (red).

procedure. Given our earlier finding that founders are likely to be new users, it is not surprising to see that tenure has a negative relationship with founding. However, both activity (edits) and diversity of activity (wikis edited, Gini of edits per wiki) predict becoming a founder, as was found to be the case in the entrepreneurship literature. This suggests that founders are *either* new users *or* active users. Founding experience is also positively related, with creating pages, founding other wikis, and being an early participant all predicting founding. Social capital is mixed, with communication network indegree as a positive predictor while PageRank and integration (coreness) are negative predictors. Overall, the difference between founders and non-founders is quite pronounced along many dimensions. On the right side of Figure 4.2 we show a few density plots comparing the number of edits and the number of wikis edited, respectively. In both cases, there are clear differences between founders and non-founders.

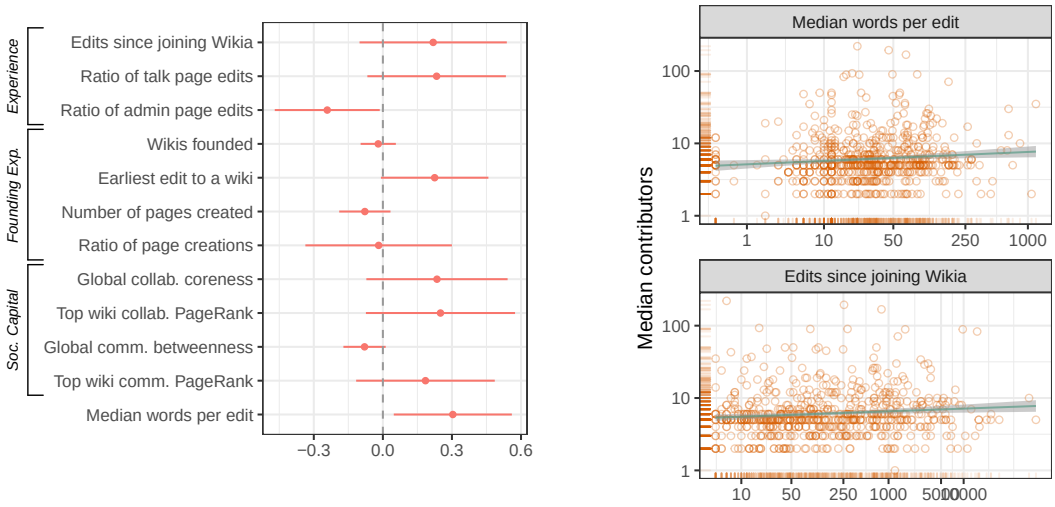


Figure 4.3. Left: Scaled coefficient estimates with 95% confidence intervals, predicting the number of contributors to a community ($N = 470$). Right: Scatterplots of the median words per edit and total edits on the x axes and median community size on the y axes.

In Figure 4.3 we show the results of a negative binomial regression predicting the median number of contributors that a founder's wikis received as of January 2010. This relationship is much noisier and much more difficult to model. The scatterplots on the right show two of the strongest predictors of growth and even these are incredibly noisy. This suggests that founder behavior and attributes do not have a strong effect on community growth. That being said, the analysis does offer a few insights. First, there is a benefit to experience in terms of editing activity, while experience as an admin has a negative relationship to community growth. Social capital is mixed, although most predictors show a positive relationship with growth. The median edit size of a founder's edits positively predicts growth. Finally, it is worth noting that the regularization step eliminated diversity measures from this model, suggesting that there is no significant relationship between diversity of experience and community growth in this dataset.

4.5. Discussion

We believe that the results of this exploratory study provide a foundation for future work. For example, founders in our study edited many different wikis and were more likely to start new pages. However, neither of these behaviors were good predictors of community growth. This suggests that many founders seek novelty and new experiences but then abandon their communities quickly to move on to the next new thing. Future research should look more explicitly at novelty-seeking behavior.

Our findings regarding social capital suggest additional hypotheses to be tested. We saw that integration in the global network was negatively related to becoming a founder but positively related to community growth. One explanation is that those who are on the periphery are discontent and thus more likely to start new communities but are ironically less able to grow their communities since growth requires spending social capital which they do not have.

Finally, we saw that a founder's typical edit size was the strongest predictor of growth. The reason for this relationship is not obvious, but one explanation is that people who make large edits typically add lots of information to a page and that wikis with more information are able to attract new contributors. Future work could test this hypothesis through experiments or causal inference.

4.6. Conclusion

There are important limitations to this work. Most obviously, we only look at founders who have previous activity on the site. Nearly 90% of the founders in our dataset were new to the site. This is consistent with research which shows that new communities can be founded as a way of learning about a site (Foote et al., 2017), but this also means that our approach

cannot tell us anything about the majority of founders. Our results regarding the propensity to found and community growth only apply to existing members and the attributes and impact of new member founders are likely to differ. In addition, while we study a very large population of projects, they are all wikis on the same platform and all of the data comes only from activity on that platform. Interviews or surveys could help us to gain a richer understanding of how founding decisions are made and whether our measures accurately represent the constructs as proposed. Finally, similar analyses in contexts such as GitHub are needed to test generalizability.

Despite these limitations, we believe that this dataset and analysis provide unique insight into an important phenomenon and establish a set of intuitions for future work on peer production and public goods production to build upon.

Acknowledgments

This work was supported by the Army Research Office (W911NF-14-10686) and the NSF (IIS-1617468, IIS-1617129). The authors also wish to thank the iConference reviewers for their very helpful comments. This paper was published by Springer as part of the Lecture Notes in Computer Science at https://doi.org/10.1007/978-3-319-78105-1_12.

CHAPTER 5

Social exposure and participation processes in online communities

Jeremy Foote
Northwestern University

Benjamin Mako Hill
University of Washington

Nathan TeBlunthuis
University of Washington

Aaron Shaw
Northwestern University

Social computing researchers have developed a number of theories about how and why people join online groups. These processes have implications for the dynamics of groups and populations of groups but these higher-level implications are rarely tested. We use agent-based simulation to model the large-scale results of lower-level rules of interaction. We argue that simulation is useful for studying social computing questions and we apply this approach to test whether current theories about how people are exposed to and decide whether to join online communities result in the highly skewed community sizes and participation rates that are observed empirically. Our simulations show that theories of group joining and exposure are inadequate on their own to predict these behavior patterns. However, when both mechanisms

are combined they provide a plausible explanation for a portion of the empirical skew in community size and participation rate distributions. We suggest that further theory is needed to explain the additional skew in online community participation.

5.1. Introduction

Why do people join and participate in online communities? Why do current members stop participating? These questions are important for platform designers but also for social scientists seeking to understand why certain groups gain power and influence. Social computing researchers have studied these questions and have developed a number of theories describing how people are exposed to online communities, how they decide whether to lurk or participate, and when they decide to leave (Bryant et al., 2005; Panciera et al., 2009; Preece et al., 2004; Ren et al., 2012; Resnick et al., 2012).

These theories are often the result of individual-level analysis. They focus almost exclusively on how individual people decide to participate in a specific community and are validated at this individual level. However, these individual-level decisions have huge implications for communities and populations of communities. Indeed, community sizes and activity levels are the aggregation of interdependent decisions made by individual potential contributors. Understanding how individuals decide which communities to participate in should allow us to predict higher-level behavior.

For example, the distribution of how many communities each person participates in and the distribution of participants per community are both highly skewed. Individual-level theories are not typically informed by or validated with this higher-level behavior and often ignore

interdependence in decision-making. It is not clear whether social computing theories about how individuals are exposed to and decide whether to participate in groups would predict these empirical patterns.

Testing higher-level effects of individual-level theories can be incredibly difficult and it is unsurprising that it is rare. In this paper, we use agent-based simulation as a methodological bridge between individual-level social computing theories and population-level outcomes.

We use social computing theory to inspire formalized models of word of mouth exposure and expected utility participation rules. We show that when tested separately, simulated agents acting according to these rules do not produce distributions of participation rates or community size similar to those seen in real-world data. On the other hand, when we combine the two models—when agents learn about new communities via their current communities and join those which are growing most quickly—communities exhibit the same kind of rich-get-richer dynamics that we see in online community platforms like reddit.

A combined model also partially explains the highly skewed number of communities per person but is not as skewed as behavior on empirical community platforms. We suggest there must be additional mechanisms which contribute to highly unequal participation rates and provide avenues for future research. We conclude with a discussion about the potential for agent-based modeling in social computing research.

5.2. Background

5.2.1. Dynamics of participation in online communities

Our claim that participation theories typically ignore community-level and population-level outcomes does not mean that these outcomes are not important to social computing researchers.

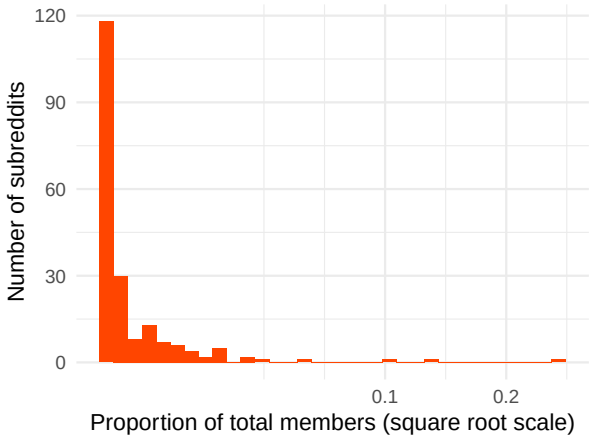


Figure 5.1. Distribution of members per subreddit in January 2017, where each member had to have made at least 5 comments. The X axis is the proportion of all participants who participated in a given subreddit.

Indeed, a persistent puzzle in online community research is the existence of highly skewed, "heavy-tailed distributions," which appear along multiple dimensions of participation (Adamic & Huberman, 2000; Johnson, Faraj, & Kudaravalli, 2014). In the context of participation decisions, the two most important dimensions are the distribution of participants per community (*community size*) and the distribution of communities per participant (*participation rate*). Both of these are highly skewed. For example, in one month on the link sharing site reddit, twelve subreddits[1] had over 100,000 unique contributors, while over 44,000 subreddits had fewer than five contributors (Figure 5.1). The number of subreddits that each user account contributed to follows a distribution with a similar shape. While most people are active in only one or two subreddits, a few people participate in dozens (Figure 5.2).

One approach to explaining these heavy-tailed distributions is fundamentally mathematical, theorizing that they arise from "cumulative advantage," where success begets success (Barabási & Albert, 1999; Merton, 1968). This approach is simple and concise, but has little

[1]Subreddits are topic-based communities created by users of the site.

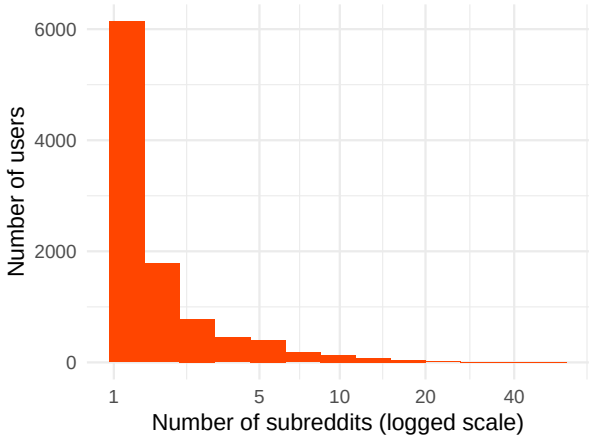


Figure 5.2. Distribution of subreddits per person from a sample of 10,000 users in January 2017, where each person had to comment at least five times to be considered a member.

explanatory power on its own. Researchers still must identify the sources of the cumulative advantage mechanisms.

Two strands of social computing research attempt to explain differences in outcomes between communities. The first treats each community as independent and models aspects of the community, its members, or their interactions as predictors of things like community size or longevity (Crowston & Howison, 2005; Crowston et al., 2006; Kittur & Kraut, 2010; Kraut & Fiore, 2014; Shaw & Hill, 2014). From this perspective, differences in outcomes arise from and are explained by things that happen within a community. For the most part, this research does not seek to explain the overall distributions of community size. In addition, this approach completely ignores the fact that individuals may participate in multiple communities and is therefore unhelpful in explaining participation rates.

A second stream of research takes an organizational ecology approach to online communities. Organizational ecologists repurpose conceptual tools from biology to understand the outcomes of groups like firms or social movements (Hannan & Freeman, 1977; Soule & King,

2008). The key claim of organizational ecology is that much of the difference in group outcomes is a result of competition between groups. Applied to online communities, researchers have shown the effect of competition on communities that overlap in membership or topic (TeBlunthuis et al., 2017; H. Zhu, Chen, et al., 2014; H. Zhu, Kraut, & Kittur, 2014). Ecological studies do take population dynamics into account and seek to explain community size distributions. However, this research typically moves agency to the organization and treats people as resources that organizations compete over. We extend this approach by looking for mechanisms of competition in the way that individuals are exposed to, navigate, and produce a system of competing communities.

In short, we have a rich literature on individual-level decision-making which rarely considers higher-level behavior and a growing literature on community dynamics which rarely considers the individual-level decisions that are the key mechanisms behind higher-level dynamics and between-group competition.

An accurate theory of how people decide which communities to participate in should predict the emergence of heavy-tailed participation rates and community sizes. In the next section, we describe social computing theories of participation and exposure in more detail and build up intuitions for their relationship to participation rates and community sizes.

5.2.2. Participation decisions

People join offline and online groups in order to advance their goals and to meet their needs for information, group identity, social interaction, or a sense of purpose (Klandermans, 2004; Resnick et al., 2012). For the most part, social computing research on this topic focuses on individuals deciding whether and how to participate in a given community. For example, a

growing body of research has studied how people begin contributing on Wikipedia, how newcomers are socialized, and what factors induce people to stop contributing (Bryant et al., 2005; Halfaker et al., 2013; Panciera et al., 2009; Shaw & Hargittai, 2018). This research suggests that in addition to having interest in a community, people estimate the impact and importance of their contributions. All else being equal, there are more benefits to being part of a larger community. For example, when the government of the People's Republic of China eliminated access to Wikipedia for their citizens, other Chinese-speaking Wikipedians responded to this reduction in community size by dramatically reducing their participation (Zhang & Zhu, 2011).

After joining a group, people develop increasing commitment to it. Attachment to a community can come through both identity and social relationships; the extent to which a person has these attachments influences whether they continue to participate (Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, & Potts, 2013; Kairam et al., 2012; Kraut et al., 2012; McPherson & Ranger-Moore, 1991; Shi, Dokshin, Genkin, & Brashears, 2017). From this perspective, people are more likely to exit groups when the costs of participation exceed the benefits.

In sum, people joining and leaving communities react to the state of the community and their expectations about what the community will look like in the future, paying particular attention to the future size of the community. This model was summarized by (Resnick et al., 2012) with a mathematical equation. They claim that participation decisions are a function of participation benefits, early adopter benefits, and startup costs. Participation benefits include social relationships, entertainment, and information, and are a function of the eventual size of the community. Early adopter benefits might include influence, reputation, or even revenue sharing. Startup costs include learning new software, learning the norms of a new community,

and building reputation and social ties. If someone believes that the *expected utility* of the participation benefits plus the early adopter benefits outweigh the costs then they will join or continue to participate.

These theories have clear implications for the distribution of community sizes. They suggest that when faced with a choice people are more likely to join larger communities. This is a clear example of cumulative advantage, the mathematical mechanism behind heavy-tailed distributions. It's not clear whether these models are strong enough to produce the extremely skewed community sizes that we see in online community platforms but it seems likely that they will at least push the distribution in that direction.

Proposition 1: Expected utility models of participation produce community sizes that are right-skewed.

On the other hand, the implications for the number of communities per person is much less clear. These theories and empirical findings are typically focused on explaining whether someone will participate in a given community and do not explicitly address how people decide how many groups to participate in. There is no evidence that they produce cumulative advantage mechanisms, such as being more likely to join a community if you already participated in many communities.

5.2.3. Exposure Processes

Exposure processes refer to how someone learns about a new online community. Exposure has been theorized as being a combination of impersonal exposure—like advertisements or search engine results—and interpersonal exposure, which occurs through social networks, with interpersonal exposure being more important (Kraut et al., 2012).

The basic idea behind interpersonal exposure is that people learn about new things via their social ties. A fairly extensive literature has emerged around product recommendations in online communities, called "electronic word of mouth" (Cheung & Thadani, 2012). This research has found that people's consumer behavior is influenced by the opinions and recommendations they get from other community members. We can consider other online communities as a type of product that may be recommended to other users. These kinds of recommendations and references to other communities are common. For example, in January 2017, 792,643 comments on reddit included links to other subreddits, which represents over 1% of all comments.

Given the prevalence and importance of word of mouth exposure, we focus on understanding the implications of this theory for higher-level behavior. At first blush, there appears to be little connection between word of mouth exposure and community size distributions. However, larger communities by definition have more members, and therefore more people out there to talk about them. This leads to more people joining them than smaller communities, leading to even more people out there talking about them: another clear example of a cumulative advantage mechanism.

Proposition 2: Word of mouth exposure mechanisms produce community sizes that are right-skewed.

It is not quite as straightforward, but word of mouth exposure also offers a potential explanation for the unequal distribution in the number of communities each person participates in (what we call the *participation rate*). If the number of communities someone hears about is partially a function of the number of communities they participate in, then those in many communities will be exposed to more new communities than others. All else being equal, this

would produce the sort of positive feedback loop that results in heavy-tailed distributions. As with the expected utility models of participation, it is not clear whether these mechanisms are strong enough to produce extremely skewed community sizes and participation rates but we posit that they will produce some skew in these distributions.

Proposition 3: Word of mouth exposure mechanisms produce individual participation rates that are right-skewed.

Even if neither of these mechanisms are powerful enough on their own to explain highly unequal empirical patterns, a model that combines both of them provides an additional feedback loop. In this combined model, people are not only more likely to be exposed to larger communities but they are also more likely to join them and therefore to expose others to these larger communities, who will also be more likely to join them, etc. This cumulative effect should provide even greater skew to both community sizes and participation rates.

Proposition 4: People who are exposed to communities via word of mouth and join them based on expected utility will produce the most highly skewed distributions of community size and participation rate.

Theories about social exposure and participation behavior were not designed to explain extremely unequal community sizes or participation rates. However, we have laid out a number of places where these theories suggest opportunities for communities and individuals to experience cumulative advantage, which is a fundamental explanation for highly skewed distributions like these. What is not clear is whether reasonable interpretations of these theories, whether tested alone or in tandem, are sufficient to explain the extreme heavy-tailed distributions that we see empirically or only enough to explain a small portion of the skew. We

use agent-based simulation to simulate people behaving according to these theories and then measure and visualize participation rates and community sizes.

5.3. Analytical Approach

5.3.1. Agent-based simulations

It is difficult to directly tie individual-level behavior to higher-level outcomes, which may be why these relationships have been largely unexplored in the social computing literature. For example, it is hard to think of an experiment or observational analysis that could show a relationship between different theories of exposure and aggregate measures of community sizes.

Many other disciplines, such as ecology, economics, and sociology, are faced with similar difficulties, often referred to as the "micro-macro divide" (Opp, 2011). In these disciplines, as in social computing, researchers study "agent-based complex systems" (Grimm et al., 2005) where properties of the larger system emerge from the interdependent decisions that agents make.

Researchers in these disciplines often turn to the use of agent-based simulation in order to bridge the micro-macro divide (Grimm et al., 2005; Wilensky & Rand, 2015). Agent-based simulations (or agent-based models) use computers to simulate the behavior of agents who can perceive their local environment, make decisions, and learn from their past experiences. Simulations thus allow researchers to isolate and study the relationships between micro-level decisions and macro-level outcomes (Grimm et al., 2005).

We begin with a null model based on agents behaving randomly. We then build models of expected utility-based participation processes and word of mouth exposure, attempting to

capture the key mechanisms of each set of theories. For each theory, we follow a "pattern-oriented" approach (Grimm et al., 2005) by comparing simulations of the theoretical model to empirical patterns of community size and communities per person which were observed on reddit in January 2017.[2]

Our goal is not to fully explain behavior on reddit; rather, we seek to understand how changes to how agents behave move the general shape of the distributions further or nearer to what we observe empirically. We therefore do not use statistical tests to compare distributions. Rather, we compare histograms to get a sense of how simulated versions of social computing theories drive changes in the skew of aggregate behavior.

This allows us to identify theories that are inconsistent with patterns observed in the real world and to refine theories and models (Grimm et al., 2005). For each model, we modify a naive formalization in a way that is consistent with the theory and with common sense but also seems likely to be a better fit to observed patterns. In other words, we ask whether there are plausible modifications that might be made to the way agents make decisions that could result in more highly skewed participation. Upon simulating the revised behavior, we again evaluate the results against benchmarks (i.e., community sizes and individual participation rates) at multiple levels of the empirically observed system. Multi-level evaluation helps ensure that our models are not simply over-fitting to a single dimension of the systems or processes in question. This approach can help us to both show the limits of theory in explaining certain higher-level behaviors as well as to provide insight into gaps in current theories.

[2]These patterns are very similar in shape to other time periods in reddit and to patterns in other online community platforms.

5.3.2. Baseline simulation model

For all of our simulation models, we begin with a population of people (agents) and a population of communities. During each "month" each person is exposed to a subset of the communities. They decide which (if any) of the communities to join, and which (if any) of their current communities to leave. Based on the empirically observed ratio of participants to communities on reddit, we model a population of 9,000 people who can be exposed to and join any of 200 communities.[3] For each run of the simulation, we stop after 24 "months" and measure the number of communities per person (*community size*) and the number of people per community (*participation rate*).

We begin with a null model, where exposure, joining, and exiting are all random. This both acts as a check that our subsequent models behave like we would expect in simplified circumstances and also shows the "main effect" of varying some of the fundamental parameters. We can also use these null, random models as a baseline to compare to more complex models.

For this set of models, in each simulated "month" every person examines the set of communities they belong to and leaves each one with probability p_l. Each month they are also exposed to a set of communities where each community is drawn from the total pool with probability p_e; they join each of these communities with probability p_j. After 24 months we examine the distributions of communities per person and people per community.

The probability of someone being exposed to a given community is not observable in the real world—we don't know how many or which online communities people are exposed to. It follows that we also do not know the probability that someone joins a community after

[3]In January 2017, there were 3,578,907 active commenters on reddit who commented in 78,201 subreddits, a similar ratio to what we use in our simulations.

exposure. We assume that both of these values are fairly low, and test p_e and p_j at the levels $\{.01, .05, .1, .2\}$. We do, however, have evidence of how often people *leave* online communities. Looking at all of the people who commented at least five times in any given subreddit in January 2017, we see that 56% of them did not comment at least five times again in February. We use that as a reasonable estimate of the baseline probability of leaving a community and set p_l to .56.

5.3.3. Expected utility-based participation

We build on this baseline model in order to simulate agents acting based on expected utility theories of group joining behavior. We base our model on the theoretical model proposed by Resnick et al. (2012) which we explained earlier. The Resnick et al. model is formalized in that it is written in equations, but the items in the expected utility equation are only intended to give broad strokes about the factors that influence decision-making processes. The authors were interested in creating a framework for thinking about how a given design choice might change general behavior; they were not attempting to design a function to predict how a given person would respond in a given situation. Our task in building agent-based simulations is to translate these general concepts into rules specific enough for agents to follow.

Our goal is to create a set of rules that represents the central aspects of expected utility theories of participation. In our first model, we do this in a naive way. We assume that people are randomly exposed to communities in the same way as in the baseline model. However, instead of choosing randomly which communities to leave and which to join, they create a combined set of their current communities and the new communities they are exposed to. They then estimate the net benefit that they will receive from participating in each community and order

their choices. They participate in each of the top p_k proportion of these communities if its expected utility is positive. Again, we expect that people only participate in a small proportion of the communities that they consider, and we allow p_k to take the values $\{.05, .1, .2\}$.

As a reminder, expected utility is calculated as participation benefits plus early adopter benefits minus startup costs. Resnick et al. (2012) treat participation benefits as a fixed quantity that can be either earned or not; we argue that it makes more sense to treat these benefits as monotonically increasing with community size. A community with 1,000 people has more information, more opportunities for friendship, etc., than a community with 100 people. However, this benefit is likely to scale sublinearly — the community with 1,000 people is not 10 times more valuable. We therefore calculate participation benefits as the log of the anticipated size of a community.

Similarly, we also treat early adopter benefits as a continuous measure. The basic idea is that the earlier you join a community, the more opportunity for influence and reputation. These benefits also increase with the eventual size of the community, but decrease based on how large the community was when you joined. We calculate early adopter benefits as the log of the future size divided by the log of the current size. Specifically, these two measures are calculated as:

$$PB = log(S_F + 1)$$

$$EA = \frac{log(S_F + 1)}{log(S_C + 2)}$$

where PB is the participation benefits, EA is the early adopter benefits, and S_C and S_F are the current size and estimated future size, respectively. Finally, we assume that startup costs (SC)

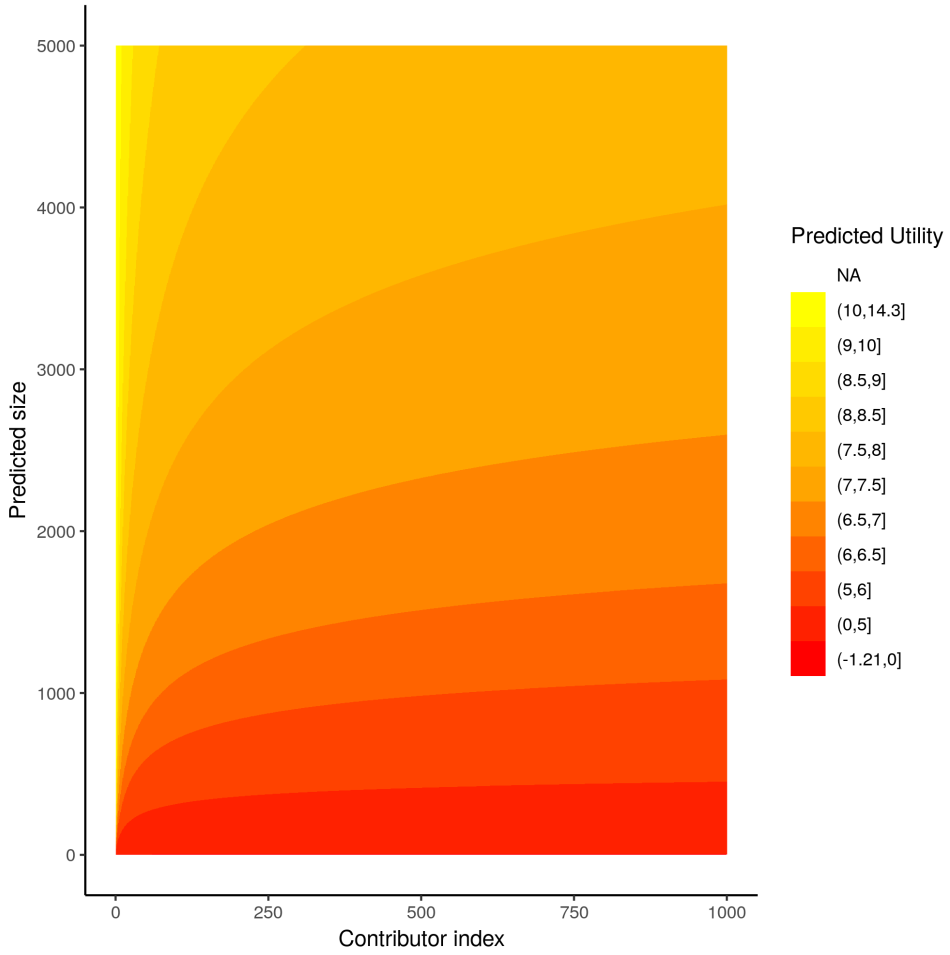


Figure 5.3. Plot of the utility that an agent would get if they were the Xth person to join a community that grew to size Y. Switching costs are set at 2.

are fixed and identical across communities, but only apply to communities an agent doesn't already belong to. The total utility for a given community is calculated as $PB + EA - SC$.

This leads to expected utility that has a reasonable congruence with the theoretical model (See Figure 5.3). Communities which are predicted to grow large have higher utility, especially those where the person has the opportunity to join while the community is still small.

There remains one important missing piece before we can simulate agents, however: we don't yet have a function for how people predict the size that a community will grow to. For our initial model, we assume that the current membership level is visible to people and that they make a linear projection from the time that the community started to six months in the future. They then use this estimate in the equations to estimate predicted utility.

5.3.4. Simulating word of mouth exposure

Unlike the model of participation decisions, we don't have a partially formalized model of word of mouth exposure to build from. There are many different ways to operationalize social exposure but we again begin with a naive approach that is consistent with the main ideas of the theory. In this model, we assume that people hear about new communities through people in their current communities. For each of the j communities that an agent belongs to, k fellow community members are chosen randomly. For each of these k members, l of their other communities are randomly shared. This leads to an exposure set with a maximum size of $j * k * l$.[4] The focal agent then randomly selects a subset of these communities to join with probability $p_j = .1$. If someone does not yet belong to any communities, then they are exposed to a random set of communities with probability $p_e = .1$.

We assume that people in larger communities are exposed to more others in the community. We therefore set $k = ceil(log(size + 1))$. We test different values of the number of projects each neighbor shares (l) from 1 to 4.

[4]The size of the exposure set will be smaller if the chosen fellow members have fewer than l other communities to which they belong or if the communities which they share overlap.

5.3.5. Combined model

For each of the standalone models (expected utility decision-making and word of mouth exposure), we simulate a naive model and then—based on the results—make reasonable modifications to attempt to create higher-level behavior that hews more closely to empirical data.

We take these modified models and combine them into a shared model that includes both social exposure and expected utility decision making. We compare these final results with the shape of the distributions found in populations of online communities.

5.4. Simulation results

We visualize a null model, showing the behavior of agents acting randomly, followed by three main sets of results. For each set of models, we show histograms of community sizes in blue followed by histograms of participation rates in gray.[5] We facet these plots to show how these distributions change at different levels of relevant parameters. In order to get a sense of the overall influence of a model we interpret the general shape of each set of histograms rather than focusing on the results from a single simulation.

Simulated agents using expected utility-based participation rules or word of mouth exposure rules alone resulted in skewed community sizes but failed to produce really large communities like we see empirically. A combined model, however, resulted in community sizes that are broadly similar to empirical communities, including the emergence of very large communities. The combined model also produced skewed participation rates, although they were not

[5]Our empirical results from reddit only include people and communities who were active; to create an accurate comparison we visualize only communities with at least one member and people belonging to at least one community.

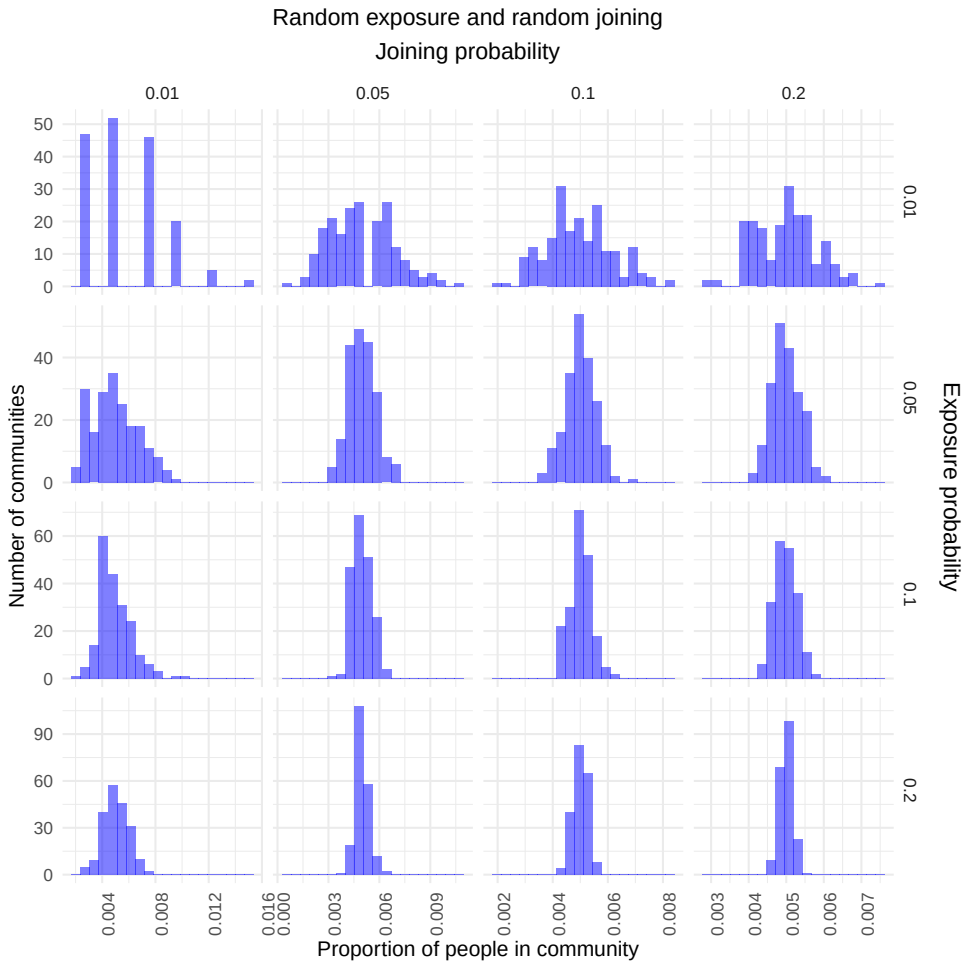


Figure 5.4. Distributions of contributors per community as the probability of exposure (vertical axis) and probability of joining a given community (horizontal axis) change. Within each plot, the X axis is the proportion of all contributors that are in a given community and the Y axis is the number of communities. Unsurprisingly, none of the distributions resemble the heavy-tailed community sizes seen empirically.

nearly as unequal as empirical data. We give a more detailed explanation of each set of results below, followed by a discussion of how to interpret them.

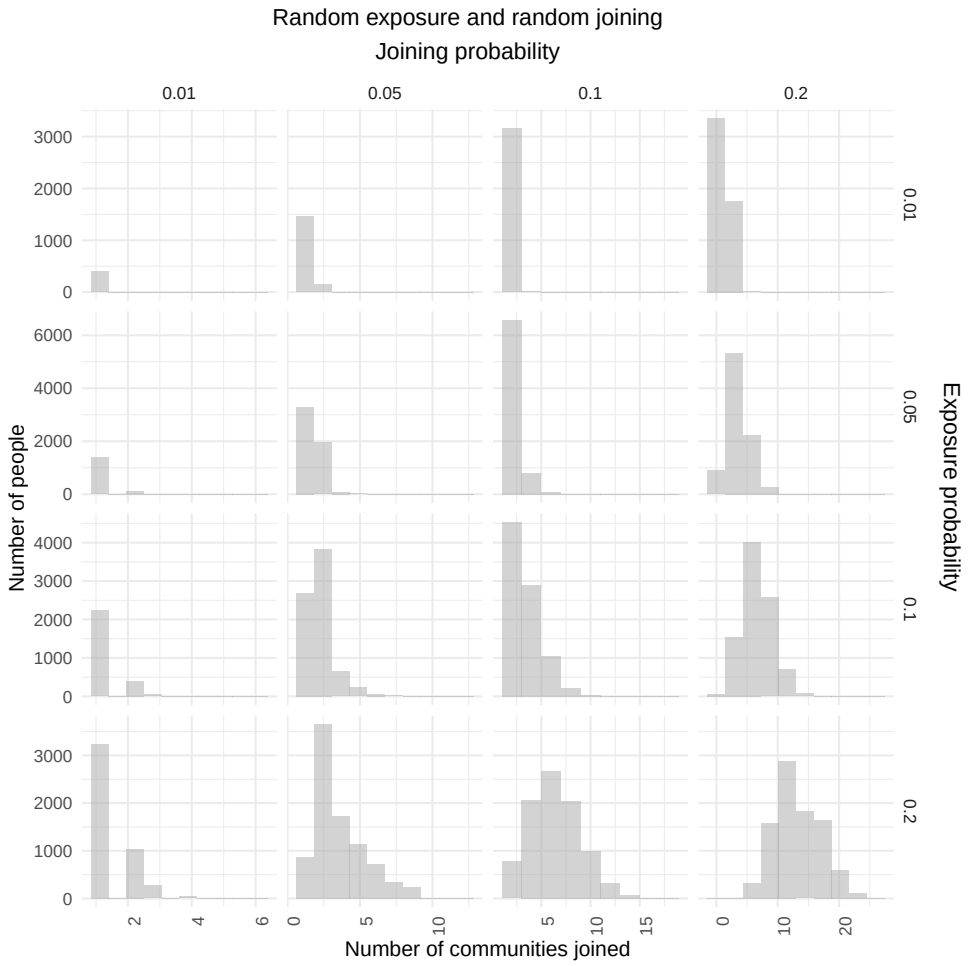


Figure 5.5. Distributions of communities per contributor as the probability of exposure (vertical axis) and probability of joining a given community (horizontal axis) change. Within each plot, the X axis is the number of communities joined and the Y axis is the number of people. None of the distributions resemble the heavy-tailed community sizes seen empirically.

5.4.1. Null model

The results from the null model are shown in Figures 5.4 and 5.5. Figure 5.4 shows the distributions of community sizes as the joining probability and exposure probability changes. Figure 5.5 shows the distributions of communities per person.

Unsurprisingly, we find that the results of these simulations do not resemble the distributions that we see in real online communities. The results resemble binomial distributions: when the probability of exposure and joining are low the distribution is slightly skewed but at higher probabilities it is bell-shaped. To simplify our models, for those models which use random exposure or random joining, we report those where $p_e = p_j = .1$.[6]

5.4.2. Expected utility-based participation

Figure 5.6 shows the results when agents use the expected utility approach based on Resnick et al. (2012). Agents estimate that communities will grow linearly and use this estimate to calculate the utility they will receive from the community. When people only join a small proportion of the communities they are exposed to, this produces a more skewed distribution of community sizes. This makes sense, because when agents only keep a small proportion of communities they will prioritize the largest communities. However, there is a tradeoff: when people only join a small proportion of communities, then no one can join lots of communities, which is the behavior we see in real online populations. In short, this model produces a distribution of community sizes that is more skewed than the null model but does not have nearly the same heavy tail that we seen in reddit. Even the largest communities have only 9% of the membership total, while the largest reddit communities have over 20%. As expected, when it comes to the number of communities per person, this model offers no improvement over the null model.

It is difficult to think of any reasonable changes to the model that would strongly influence the participation rate to become more skewed. On the other hand, it is tempting to think that

[6]For more complicated models, we report $p_e = p_j = .05$ in the appendix as robustness checks.

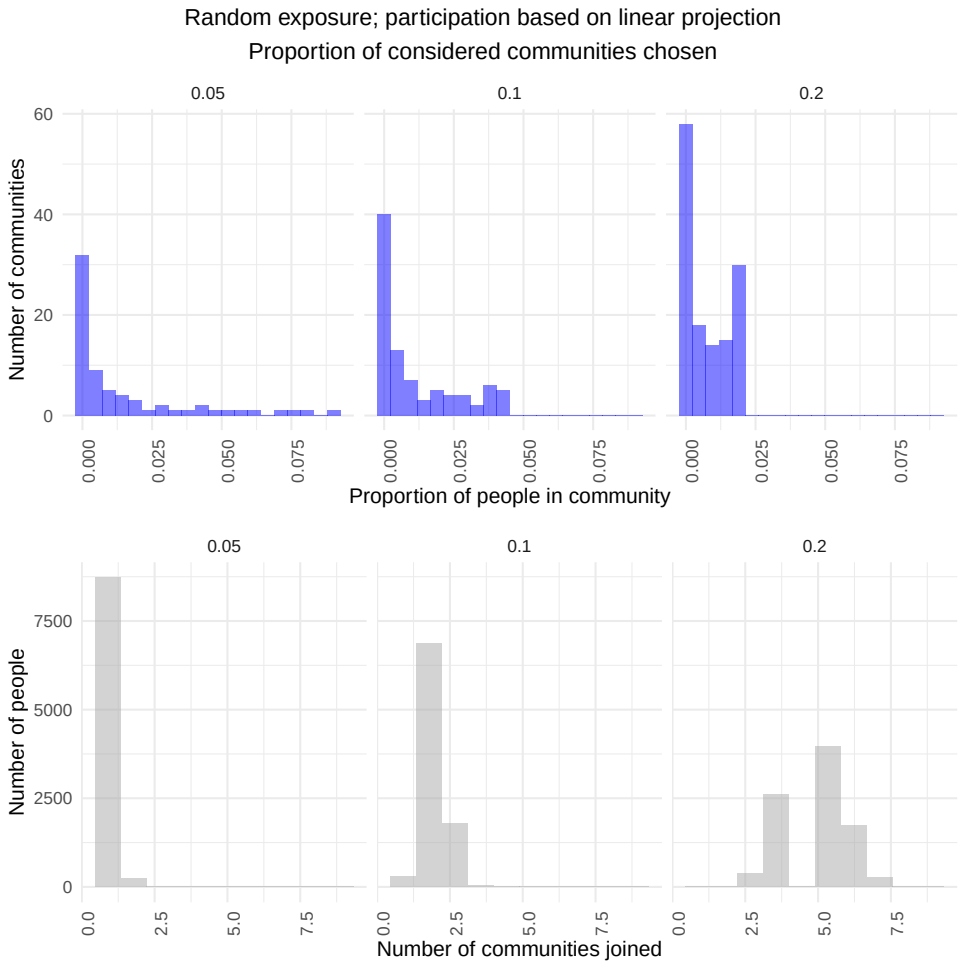


Figure 5.6. Distributions of community size (upper) and participation rates (lower) when agents are exposed to a random set of communities and choose based on expected utility. Moving from left to right, the proportion of communities that they join increases.

we might be able to produce more skewed community size distributions. In order to do that, people must be even more likely to join already large communities. There are a few ways that our model might push people in this way. The simplest is to tweak the way that they estimate future sizes. In Figure 5.7, agents fit a quadratic equation rather than a linear equation to a community's current size and use that to predict the size in six months. This has the effect

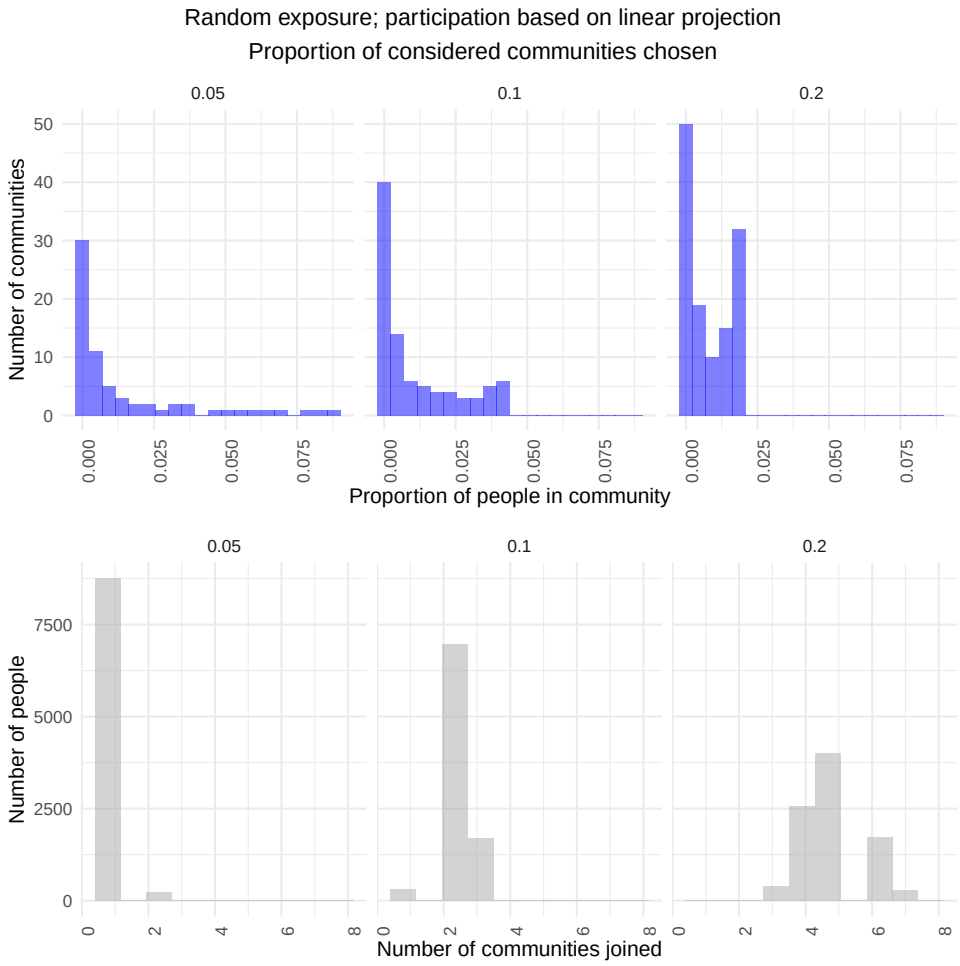


Figure 5.7. Results for random exposure and expected utility participation, when future size is estimated based on a quadratic equation. The results are nearly identical to Figure 5.6, suggesting that people were already joining the largest projects.

of increasing the predicted size of already large communities by much more than it increases the predicted size of small communities. However, Figure 5.7 is almost identical to Figure 5.6, suggesting that agents were already selecting the largest communities.

Indeed, no simple tweaks to our model of decision-making could produce the kind of superstar communities that we see empirically. In our model, people are exposed to a small,

random proportion of communities. Even if they always join the largest community or set of communities they are exposed to, no community will be seen often enough in 24 months to create a heavy-tailed distribution.[7]

5.4.3. Word of mouth exposure

We next look at simulations of word of mouth exposure. We predicted that word of mouth exposure would have cumulative advantage effects, and would lead to skewed community sizes and participation rates. We assumed that large communities should be more likely to be shared, solely because more people belong to them. However, this naive model where people share a random set of communities results in a bell-shaped distribution of community sizes. We discuss possible explanations in the discussion section.

Nor does this model produce clearly skewed participation rates, with the exception of when people share only one community. Even here, it does not capture the heavy tail seen on reddit and likely results from truncated participation rates rather than cumulative advantage mechanisms.

As before, we compare the results of the simulations with our empirical patterns and ask what changes we might make which seem consistent with common sense, prior theory, and empirics. The most obvious is to assume that instead of choosing randomly which communities to share, people will share the largest communities to which they belong.

Figure 5.9 shows the results of these simulations. Both community sizes and participation rates are much more skewed in this model, suggesting that our overall intuitions about the potential for cumulative advantage mechanisms were correct. However, these measures are

[7]Results from a simulation of this are shown in Figure 5.15 in the appendix.

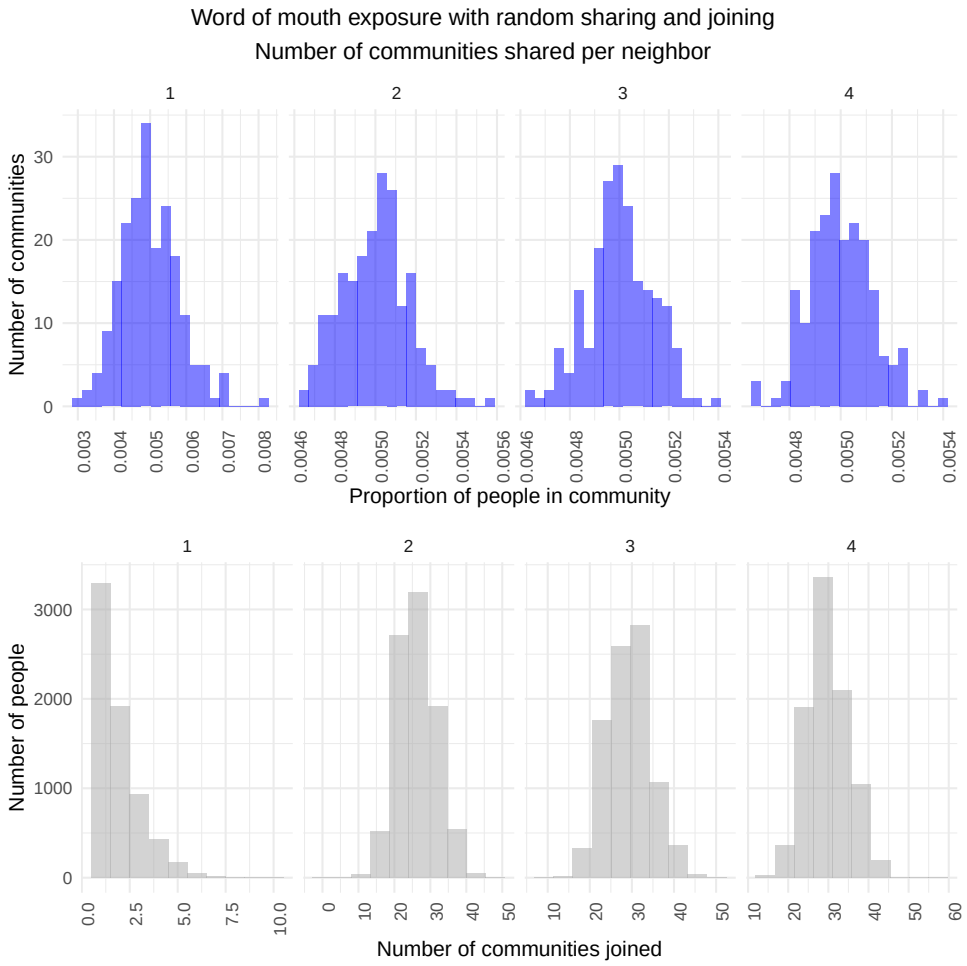


Figure 5.8. Community sizes (upper) and participation rates (lower) when people are exposed to random new communities via people in their current communities. Moving from left to right, the number of communities that each "neighbor" shares increases.

still much less skewed than empirical populations. In particular, as with earlier models, none of these effects were strong enough to produce "superstar" communities that were much larger than others. Nor did they produce "superstar" participants. In addition, even the general shape of these distributions appears to be fragile. In the appendix, we show the results from a robustness check with slightly lower joining probabilities where we find very different results.

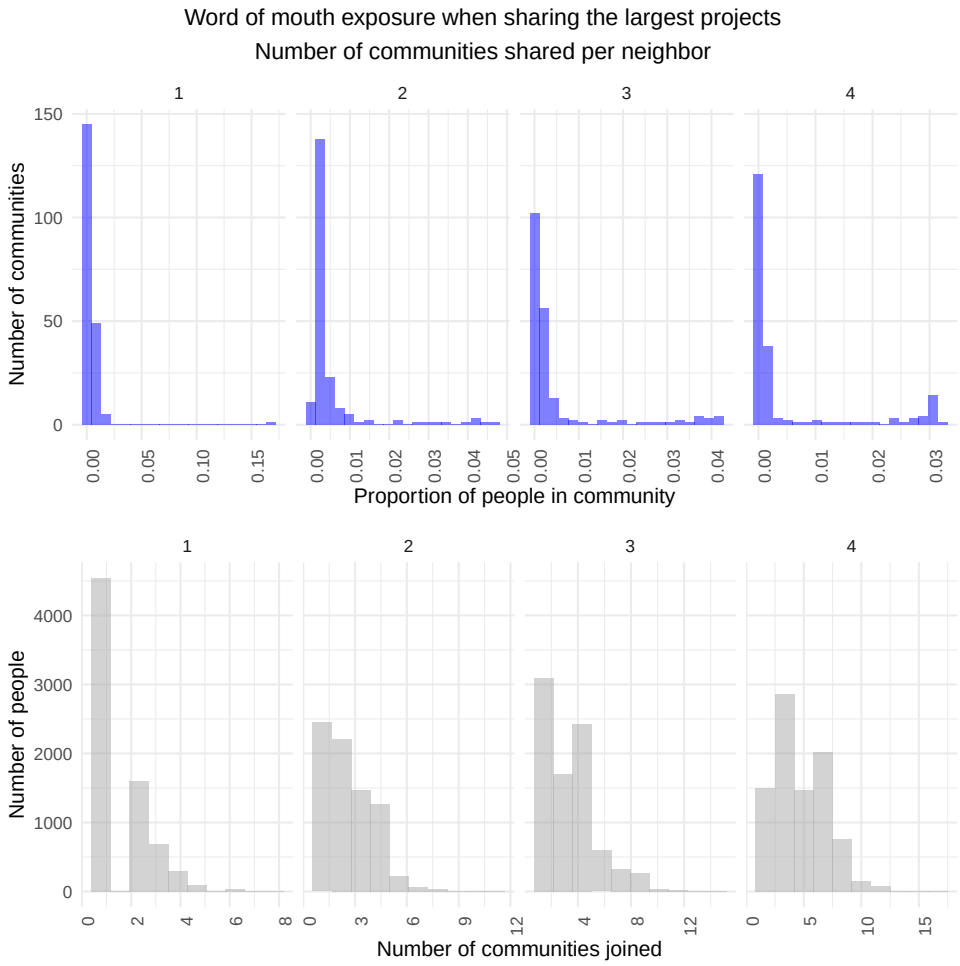


Figure 5.9. Community sizes (upper) and participation rates (lower) when people are exposed to new communities via people in their current communities sharing the largest communities to which they belong. Moving from left to right, the number of communities that each "neighbor" shares increases.

5.4.4. Combined model

Finally, we combine our models of exposure and joining behavior. In this combined model, people are exposed to new communities via people sharing the largest communities they already belong to, and they join or leave them based on the expected utility using the linear

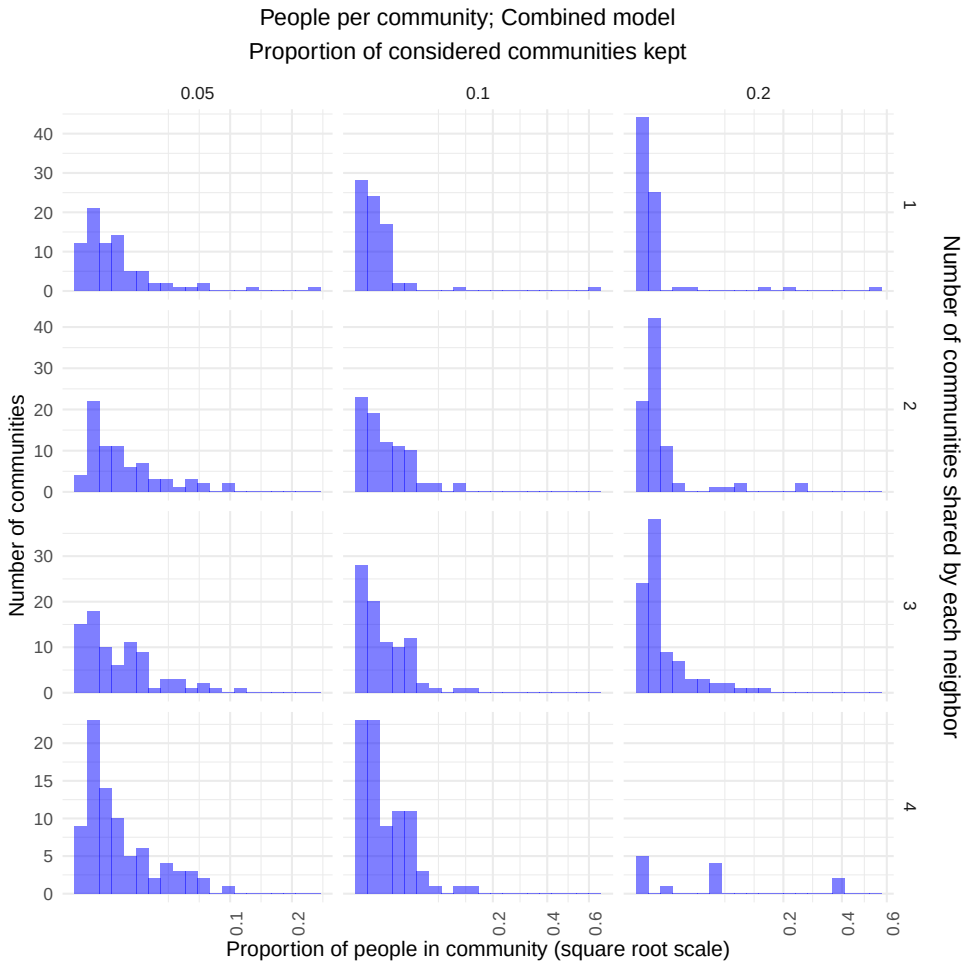


Figure 5.10. Community sizes when people are exposed to new communities via people in their current communities sharing the largest communities to which they belong. Moving from left to right, the proportion of communities kept increases; from top top bottom, the number of communities that each "neighbor" shares increases.

estimation technique. Figures 5.10 and 5.11 show the results of this combined model across a range of parameters.

When word of mouth exposure combines with expected utility decision-making, agents are exposed to and incentivized to join larger communities. As we expected, these effects are additive and for the first time the simulations produce both a skewed distribution and the

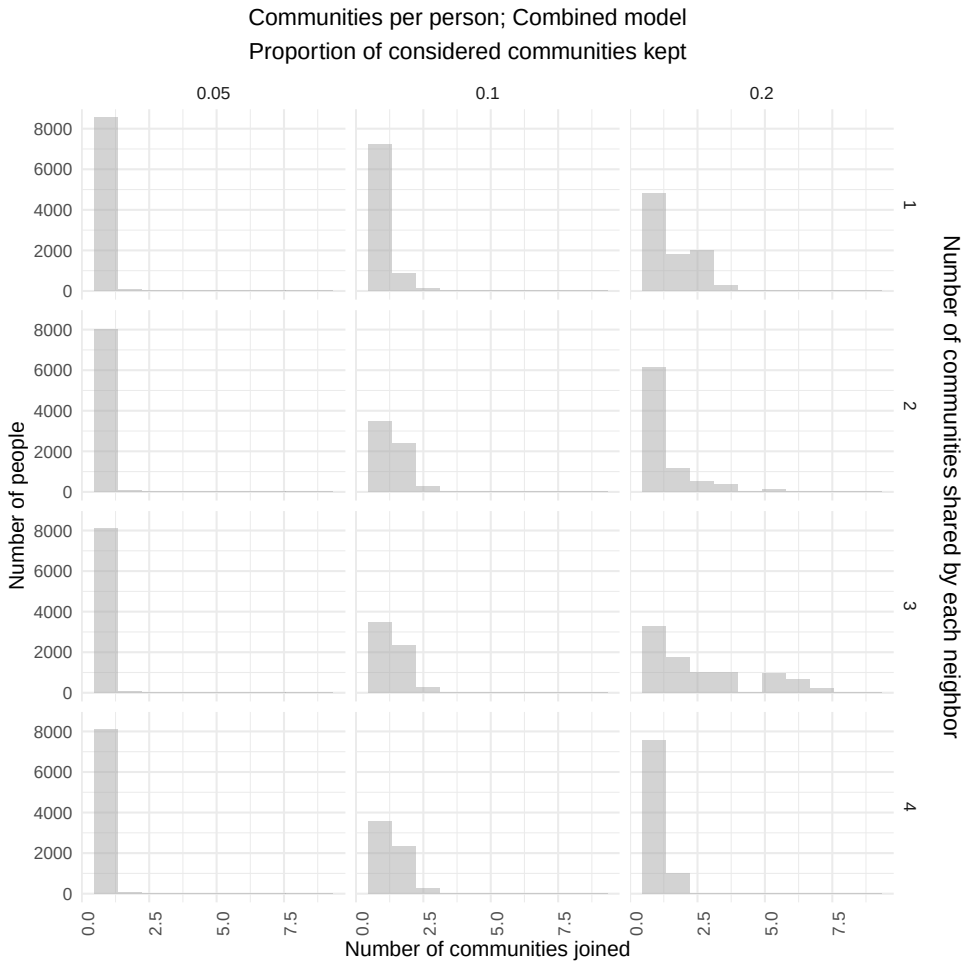


Figure 5.11. Communities per person when people are exposed to new communities via people in their current communities sharing the largest communities to which they belong. Moving from left to right, the proportion of communities kept increases; from top top bottom, the number of communities that each "neighbor" shares increases.

production of "superstar" communities to which a large proportion of the participants belong. While these distributions are still not as skewed as empirical community sizes, the general shape is fairly similar and is robust to different parameters for how exposure and community choice are made.[8] Participation rates under the combined model are also more skewed than

[8]In the appendix, we show the results of a robustness test with lower exposure probabilities and find similar results

either model when acting alone, although they are not as skewed as community sizes or as the participation rates we see on reddit.

5.5. Discussion

For the most part, our prior beliefs about the relationship between social computing theories of exposure and participation were borne out in the simulations. Expected utility-based participation rules and word of mouth exposure both resulted in right-skewed community sizes, although when considered separately they did not produce the extremely large communities seen in empirical data. A combined model produced moderately realistic distributions of community size—including very large communities—robust to different parameters. This combined model also produced the most realistic participation rate patterns, although these still were not nearly as unequal as the number of communities that online participants contribute to.

A few findings bear further comment. First, we observed that random exposure limited the opportunity for heavy-tailed community sizes. If we believe that people are only exposed to a small proportion of the total population of communities, then exposure processes cannot be random. Mechanisms that make it more likely that people will consider larger communities rather than smaller ones are required to produce extremely large communities.

Second, the effect of word of mouth exposure was highly dependent on which communities people shared. When sharing a random community, social exposure had very little impact on the distribution of community sizes or participation rates. A possible explanation is that when people only belong to a few communities (which is how our simulation begins) and share randomly, no community gets large enough to get the benefits of social exposure. Our

pattern-oriented approach led us to model agents choosing to share the largest communities to which they belonged, and only then did word of mouth exposure lead to skewed community sizes and participation rates. Our simulations suggest that people are more likely to share large communities, a theory which should be tested empirically.

Even when people shared the largest communities they belonged to, none of the communities became really, really large. One explanation is that through sharing, people start to naturally cluster so that they are mostly sharing the same set of communities with each other. Future research could explore directly the emergence of clusters in these models and compare these to participation clusters in online community platforms.

On the other hand, we found evidence for a positive feedback loop in the combined model. When agents were both more likely to join larger groups and more likely to be exposed to larger groups, we saw an increase in the skew of both community sizes and participation rates.

In sum, these simulations suggest that social exposure combined with expected utility decision-making provide a partial explanation for empirical participation patterns. Our results do suggest that there is more to the story, however. In particular, more research is needed to identify additional mechanisms of cumulative advantage in individual participation rates. For example, perhaps people's willingness to join a new project increases as they participate in more projects or their interests broaden and they increase their exposure to new communities.

5.5.1. Limitations

Our approach has a number of important limitations. One is common to all agent-based simulations: while we chose a set of models that we believe capture the most important aspects of the real-world system, other reasonable formulations could have different outcomes. Guided

by theory, we attempted to identify the aspects of the model most likely to be the source of variation and to paramaterize and test these. However, our findings suggest that other aspects of the environment are also important and influence patterns of participation behavior.

For example, we treated all of the agents and communities as homogeneous. While our model provided plenty of opportunity for heterogeneity to arise, it is possible that external heterogeneity in free time or interest level or skills could be magnified by participation in online community platforms and could help to explain additional skew in participation.

Finally, future work should look at how well these models predict additional behavioral patterns, such as clustering in participation networks or temporal patterns of contribution.

5.6. Implications and conclusion

Our simulations suggest that individual-level theory provides a partial explanation for how highly skewed distributions of group sizes arise: people learn about new communities from the communities they are already in and join the communities that are growing quickly. This simple model elides much of the complexity of how the real world works but is a plausible first approximation to explain the source of community size distributions. This suggests that much of the attention that large online communities get may be based on luck rather than being the result of any intrinsic value (c.f., Salganik, Dodds, & Watts, 2006).

On the other hand, we provide evidence that individual-level social computing theories do not provide powerful enough cumulative advantage mechanisms and do much less to explain why some people contribute to many different groups while most only contribute to one or two. We call for more research into understanding additional mechanisms at play.

Finally, we hope to have shown the value of simulation for social computing research. Theories like critical mass theory (Marwell & Oliver, 1993) and complex contagion (Centola & Macy, 2007), which were developed with the help of agent-based simulations, have been influential in social computing research (Raban, Moldovan, & Jones, 2010; Romero, Meeder, & Kleinberg, 2011). However, despite its influence, its fit with social computing questions, and calls for its use from prominent researchers (Ren & Kraut, 2014), agent-based simulations remain very rare in human computer interaction and social computing. We have given one example of how agent-based simulation can be used to link theories and observations from multiple levels. We believe that increased use of simulation methods would lead to better theories of human behavior and a better understanding of community dynamics.

5.7. Acknowledgements

This work relied on use of the University of Washington's Hyak computing cluster and was supported by NSF grants IIS-1617468 and IIS-1617129. Earlier versions of this paper received very helpful feedback from participants at the International Communication Association conference, the Organizational Communication Mini Conference, and the International Conference on Computational Social Science.

5.8. Appendix

5.8.1. Robustness check results

In the text, Figure 5.9 shows the results when in the initial round people are exposed to each community with $P_e = .1$ and in subsequent rounds are exposed to the k largest communities of randomly chosen neighbors. They then join each community randomly with probability

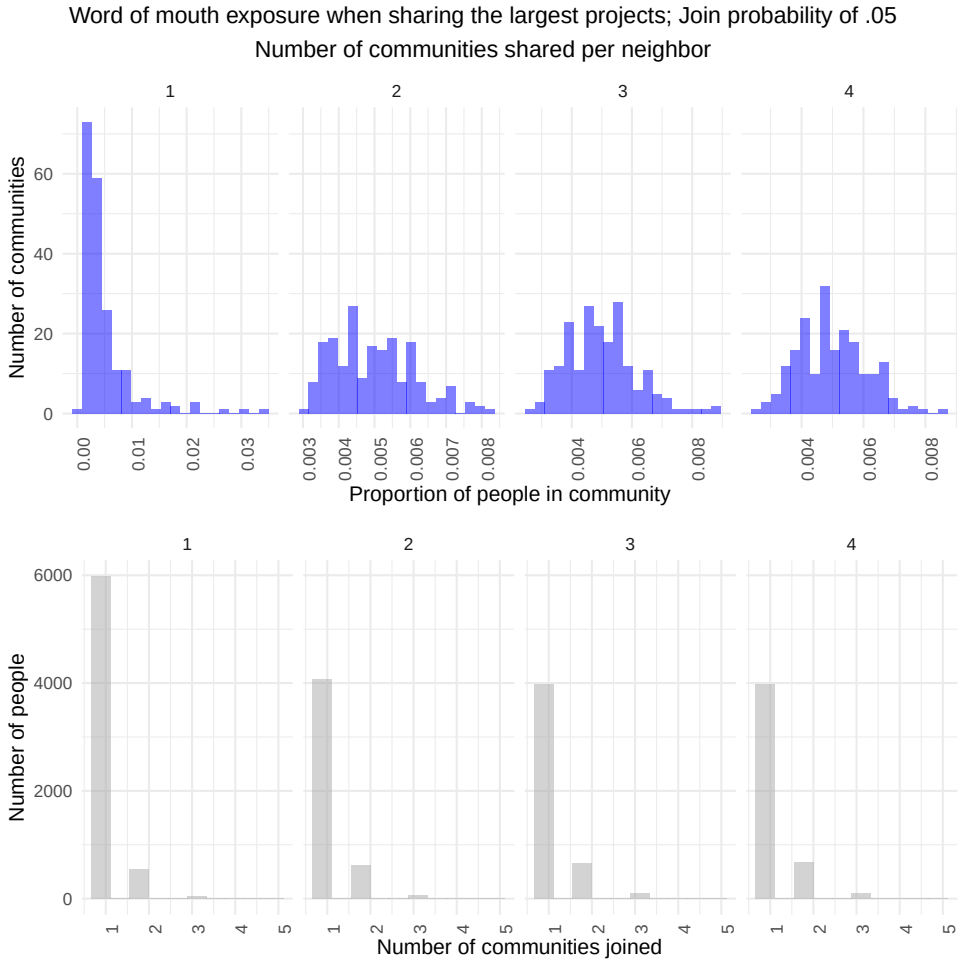


Figure 5.12. Community sizes (upper) and participation rates (lower) when people are exposed to new communities via people in their current communities sharing the largest communities to which they belong, and join those communities randomly with probability $P_j = .05$. Moving from left to right, the number of communities that each "neighbor" shares increases.

$P_j = .1$. Figure 5.12 shows the results of simulations when $P_e = P_j = .05$. These results show that when people start with a slightly smaller set of communities and are slightly less likely to join new communities, nearly all of the skew of community size disappears, suggesting that this effect is fragile.

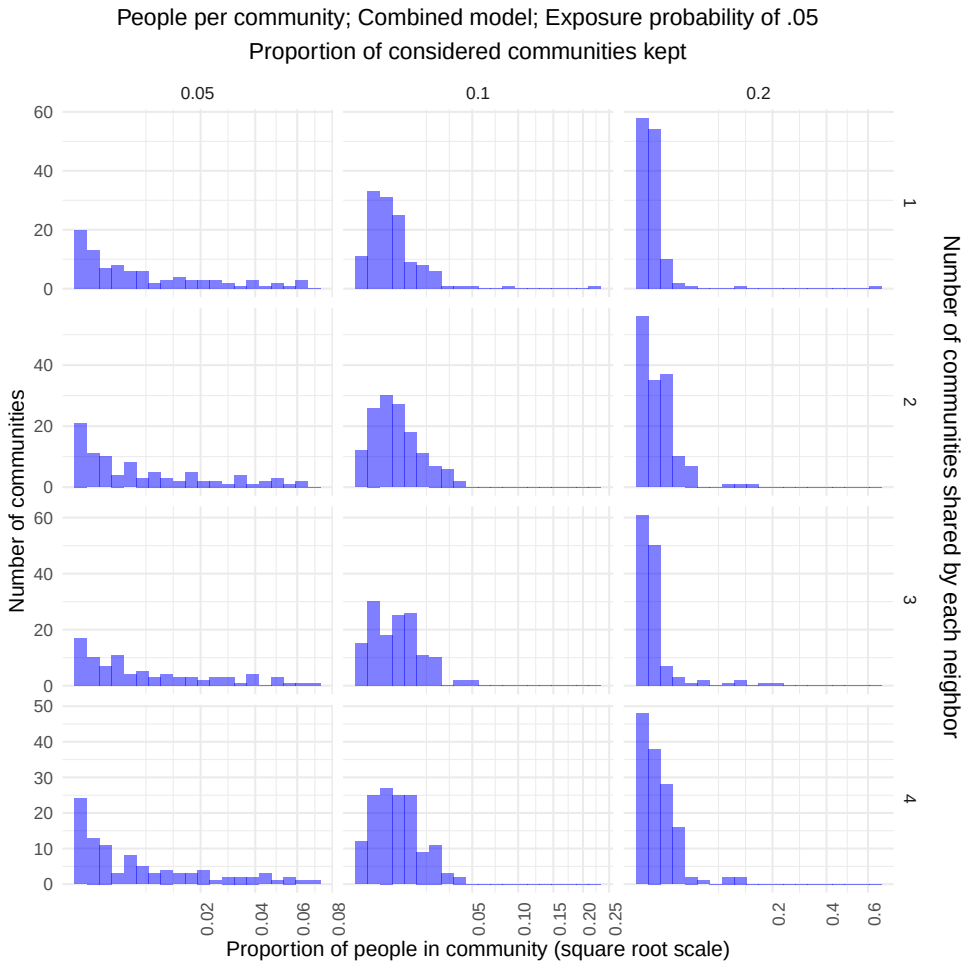


Figure 5.13. Community sizes when people are exposed to new communities via people in their current communities sharing the largest communities to which they belong, with initial exposure probability $P_e = .05$. Moving from left to right, the proportion of communities kept increases; from top top bottom, the number of communities that each "neighbor" shares increases.

On the other hand, Figures 5.13 and 5.14 show a robustness check for combined models. Here, when the initial set of communities people are exposed to, P_e, is changed from .1 to .05 there is very little effect on the overall shape of the eventual outcomes.

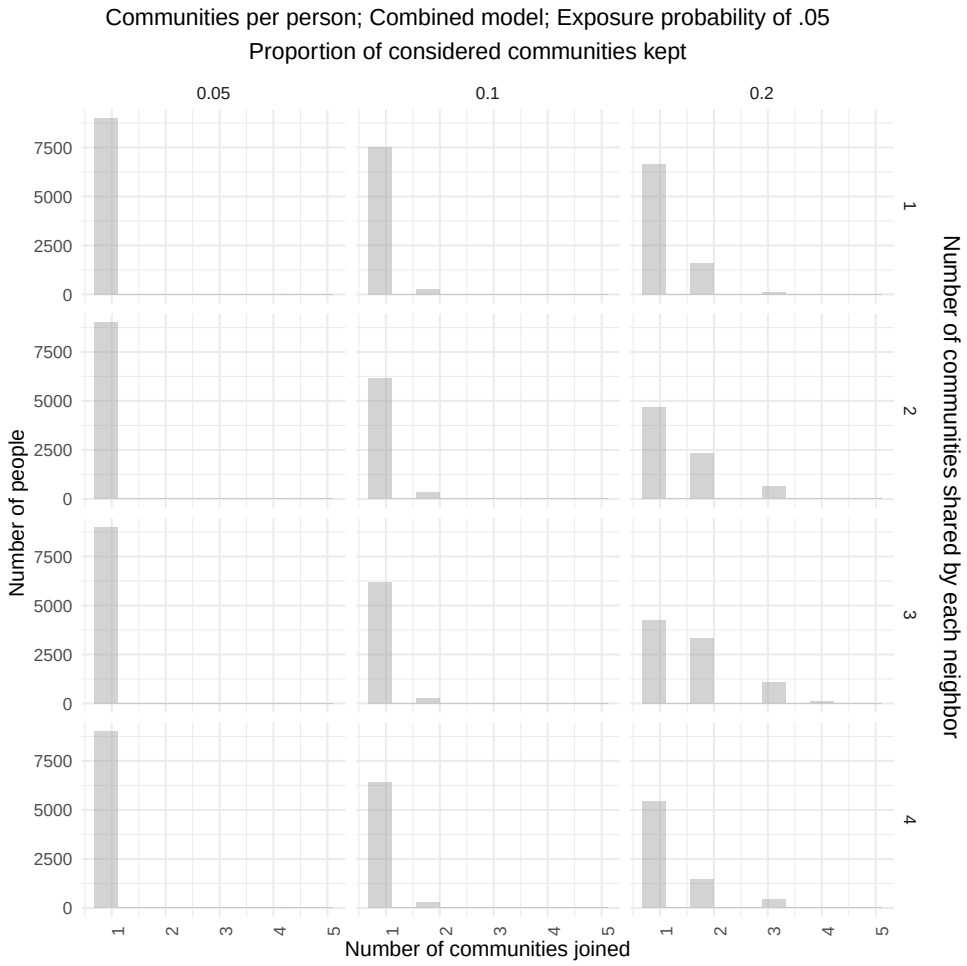


Figure 5.14. Communities per person when people are exposed to new communities via people in their current communities sharing the largest communities to which they belong, with initial exposure probability $P_e = .05$. Moving from left to right, the proportion of communities kept increases; from top top bottom, the number of communities that each "neighbor" shares increases.

5.8.2. Limits of random exposure

When people are exposed to only a fairly small subset of communities, then even if they join the largest community these communities still never grow as large as empirical communities.

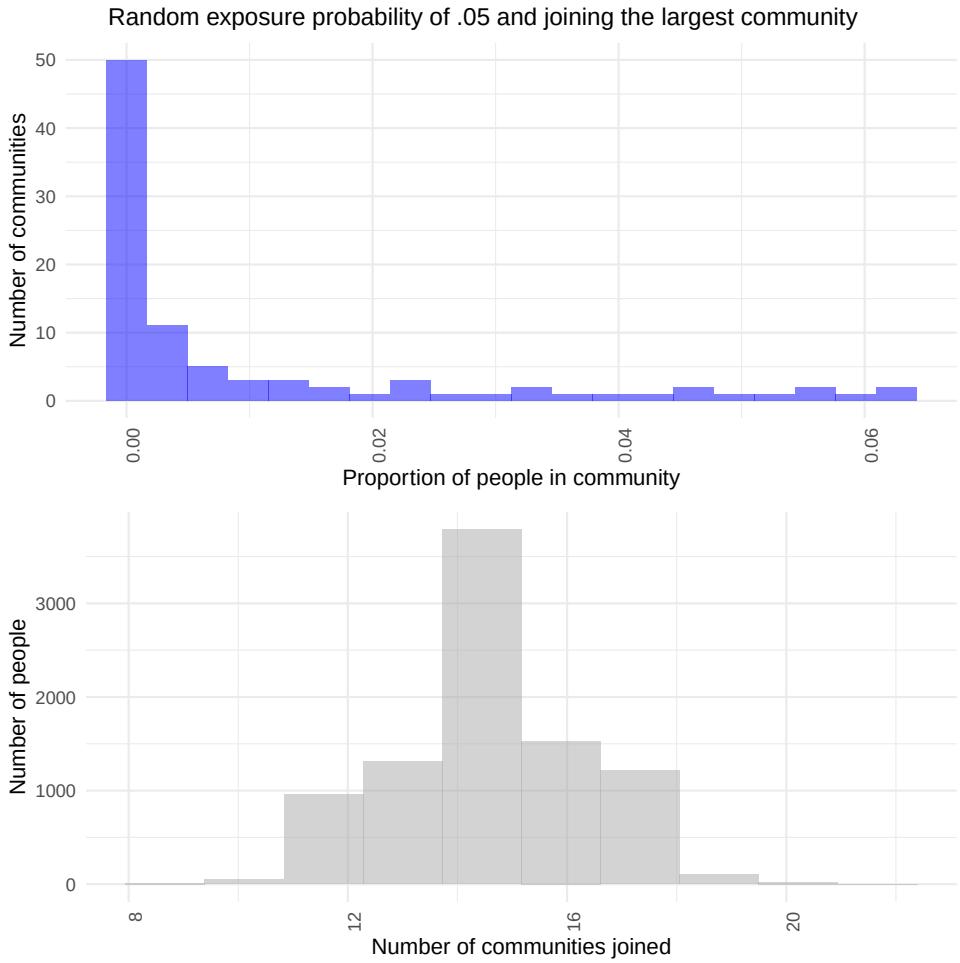


Figure 5.15. Community size (upper) and participation rates (lower) when people are exposed to communities randomly with $P_e = .1$ and join the largest community they are exposed to.

Figure 5.15 shows simulations where people are exposed to communities with $P_e = .1$ and simply join the largest.

CHAPTER 6

Conclusion

Each of the projects of this dissertation had a fairly small scope, focusing on a particular aspect of decision-making within COO. Each project had its own findings and insights which I have tried to enumerate both in the introduction and in each paper. In this chapter, I attempt to identify larger themes that appear when considering the projects as a body of work. I review what I consider to be the key insights and sketch out a brief roadmap of how to move this work forward in each area.

6.1. Ecosystems of collaborative online organizations

In the introduction chapter, I introduced a new approach to studying systems of COO, focused on understanding the relationship of systems and affordances to individual participation decisions. As discussed, the concept of treating online organizations as part of a larger ecosystem is not new: Butler (2001), H. Zhu, Kraut, and Kittur (2014), and H. Zhu, Chen, et al. (2014) are each built around the idea that COO are in relationships with each other. More recent research like TeBlunthuis et al. (2017) and Tan (2018) take the idea of community interdependence seriously and start to use the rich person-level data that we have to understand relationships between online organizations.

However, this dissertation shows the importance of identifying the influence of the larger ecosystem not only at the organizational level but also at the individual level. These projects repeatedly show that individual decisions are shaped by the larger ecosystem. For example, in

Chapter 4 we saw that being active in other COO was a good predictor that someone would become a founder. In Chapter 3 we showed that decisions about whether to start a new wiki and on what topic were directly influenced by founders' knowledge of what already existed. The founders that we surveyed reported that they intentionally started COO in niche topics with the expectation of gaining only a small number of contributors. Finally, in Chapter 5, we showed that even homogeneous agents who are randomly exposed to different portions of the space of communities can end up making dramatically different decisions in the future.

We showed some of the ways in which systems influence individual decisions but future work should do more to understand the mechanisms behind competition and mutualism. For example, researchers could study how individual behaviors differ in situations that are mutualistic versus competitive and could identify how individual decisions contribute to those dynamics. For example, under what conditions do people choose to create a very similar COO to an existing one, rather than finding a new niche?

Other explanations for relationships between COO could come from a deeper understanding of how people change as they move through a system of online communities. If people change in response to their participation in a community, then adjacent communities will be much more mutualistic that we would expect. For example, people who join a COO about the programming language Python may actually develop more interest in programming and may be more likely rather than less likely to join a COO about R.

Understanding how people are influenced by the COO they participate in has lots of other important implications. If online groups are merely reflective of their members then shutting down COO that are offensive or radicalizing will only move those people to darker corners of the Internet. If, on the other hand, people's interests can change based on the COO that

they see and join, then removing some COO may actually prevent people from becoming more offensive or radicalized. There is some important evidence that this may be the case (Chandrasekharan et al., 2017), but digital trace data affords more opportunities for studying the trajectories of users through COO systems.

Finally, there is room for more research about how organizations respond to ecological pressures. For example, Seattle has two competing subreddits whose leaders intentionally position their communities in opposition to each other. More work could be done to understand the decisions that online group leaders make when faced with direct opposition or how they work to establish niches and mutualistic relationships. A related line of research could explore the way that individual members adjust their behavior in response to competition. For example, does competitive pressure move communities into more narrow niches and if so, how does that happen? Do individual people change their posting behavior to a more narrow set of topics or is there a selection process where only those interested in the more narrow set of topics remain active?

6.2. Small, temporary organizations

The second insight is our evidence for the existence of small and temporary COO. Previous literature has focused almost exclusively on large online organizations with broad goals, such as Wikipedia. Even those researchers who study populations of organizations often assume that becoming a large, active community is the goal and treat community size and longevity as the key metrics defining success (e.g. Kraut et al., 2012; H. Zhu, Kraut, & Kittur, 2014). This dissertation provides evidence that these approaches have a blind spot. Rather than being homogeneous attempts to become large-scale organizations, the vast majority of

COO exist to meet narrow and often short-term goals. The strongest evidence for this comes in Chapter 3, where we interviewed founders of new wikis. The majority of these founders were very clear that their goal was not a large, broad community but the creation of a high-quality, niche information good.

The existence of these small communities is a direct consequence of the extremely low costs to create an organization on online platforms like Wikia or reddit. If we assume that people will only create a new community if they expect to get at least as much value out of it as the effort required to create it then lowering the costs of creation will have the effect of increasing the number and type of communities that are created. Low costs allow people to create organizations to fulfill short-term needs such as a group of students organizing work for a class project or online gamers planning their next raid. More work should be done to understand the prevalence and purpose of these small organizations.

The affordances that make it cheap to create new COO allow for more extensive exploration of the space of possible organizations. This process not only allows for the creation of intentionally niche organizations, it also allows for the discovery of surprisingly successful organizations. When he started the project that eventually became Linux, Linus Torvalds was famously modest about the scope and potential for his project, which he described as "a (free) operating system (just a hobby, won't be big and professional like gnu)".[1] Linux is now one of the key pieces of software behind the modern web and nearly all of the top supercomputers. This sort of niche-becomes-popular story may be common in online organization ecosystems and suggests a strategy that privileges exploration. Future researchers could explore the prevalence of these sorts of surprise successes.

[1]Post retrieved from https://groups.google.com/forum/#!msg/comp.os.minix/dlNtH7RRrGA/SwRavCzVE7gJ

Our findings also suggest that researchers should do more to identify alternative measures of success and that focusing on growth or longevity misses much of what's important. Researchers studying one type of online organization—open-source software communities—have been exemplary in this regard. Perhaps because software is more of a product than other COO outputs, researchers have long recognized the importance of trying to measure the quality and impact of the artifact and not just attributes of the community (e.g. Crowston et al., 2003; Schweik & English, 2012). Researchers of other online organizations such as information and entertainment communities should do more to take into account the artifacts that are created as measures of success.

6.3. Social motivations

Third, our results suggest that there is more work to be done to understand the role of social relationships as drivers of online cooperation. In our agent-based simulation project in Chapter 5, we found that some form of social exposure is key to explaining patterns of participation in COO. However, we found little evidence of the influence of social relationships on either founders or early-stage members. From the survey research discussed in Chapter 3 and in the previous section, we learned that wiki founders typically care much more about creating a high-quality and useful information artifact rather than about building relationships with other participants. In Chapter 4 we learned that founders could not or did not use their position in the social network to recruit others. Finally, in Chapter 2 we found that COO with structures of communication which reflect integrative, participatory networks performed no better than other COO.

It is worth discussing the possible mechanisms behind these findings and their implications. One possible explanation, which we discussed earlier, emerges from a systems view. Social exposure may help people to discover new COO, but when there are so many COO on so many topics then those who join a given COO do so because they are highly interested in the topic and are already motivated (Felin, Lakhani, & Tushman, 2017). Additional motivations, such as social motivations, simply are not necessary.

If the first explanation is for selection effects on who joins a community, another focuses on selection effects on who remains. In organizations like firms, where exit is costly, social relationships may help a group to reach consensus and avoid being mired in disagreement and dissent. In COO, it may be the case that the dissenters simply exit the group, creating consensus via exit rather than voice (Hirschman, 1970).

It is important to note that these explanations are slightly at odds with each other. If participants are highly motivated and dedicated to a topic then while the transactional costs to exit are low the personal costs may be quite high. Future research should work to tease these apart and to understand more explicitly the factors that influence these selection effects.

More broadly, there is room for much more research on the influence of relationships on individual decisions and COO outcomes. Conversations, collaborations, and relationships almost certainly carry over from one community to another and have a hidden influence. For example, researchers could study whether already-existing relationships help a COO to be more productive or how they constrain which COO people are exposed to and participate in.

6.4. Affordances and participation costs

Over and over again we have argued that many of the findings in this set of projects, and many of the higher level implications we have identified in this chapter, are caused in part by technological affordances which result in low costs to join, leave, and create new COO. One effect of these affordances is that the boundaries of an online organization are much more porous than traditional organizations like firms. The implications of porous boundaries have been explored from the perspective of organizational research, focusing on understanding how individual COO adapt to these conditions (e.g. Faraj et al., 2011; Felin et al., 2017).

We have tried to extend our understanding of the implications of low participation costs in two directions. First, we have worked to understand how individuals respond to a context where they can quickly move between COO. We have shown that low costs make people sensitive to the state of the system and their previous experiences when deciding to create, join, or leave COO.

We have also worked to extend these ideas up to the population level. In Chapter 5, we showed how low-cost dynamics can result in winner-take-most ecosystems with lots of niche communities. We have given evidence in Chapter 3 that these small COO are intentional, diverse, and exemplify a type of organizational form that directly results from the low costs of COO creation. Finally, we argued in Chapter 2 that low costs may result in communities of like-minded people who are internally motivated and have thin social ties.

These arguments stem from a theoretical understanding of COO but more empirical work should be done to study and validate the influence of transaction costs. For example, commenting on a subreddit is very simple while making an edit on an open-source project requires both background skills in programming and a fairly thorough understanding of the project.

Future work should explore how differences in costs influence individual decisions and group-level behavior.

6.5. Conclusion

The four papers of this dissertation help to illuminate the way that early-stage online communities work. In short, they are interdependent. Founders start new organizations with a knowledge of what already exists. People learn about new COO from their current set of communities. Past experience and relationships in one community predict the behavior that someone will have in another.

I've given specific ideas for future work above, but more broadly, work in this space should consider and embrace this interdependence. At the community level, we can understand the behavior of COO better if we understand the relationships and experiences that its members already have. At the individual level, we can understand better the decisions that people make by thinking about how they perceive and understand the larger system in which they are participating. The incredible individual-level and community-level data available allows for exciting opportunities to do this kind of work well.

References

Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the world wide web. *Science*, *287*(5461), 2115–2115. doi:10.1126/science.287.5461.2115a

Anderson, A. R., & Miller, C. J. (2003). ‘‘Class matters”: Human and social capital in the entrepreneurial process. *The Journal of Socio-Economics*, *32*(1), 17–36. doi:10.1016/S1053-5357(03)00009-X

Antin, J., Cheshire, C., & Nov, O. (2012). Technology-mediated contributions: Editing behaviors among new Wikipedians. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 373–382). CSCW ’12. doi:10.1145/2145204.2145264

Arazy, O., Liifshitz-Assaf, H., Nov, O., Daxenberger, J., Balestra, M., & Cheshire, C. (2017). On the "how" and "why" of emergent role behaviors in Wikipedia. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW ’17* (pp. 2039–2051). doi:10.1145/2998181.2998317

Artinger, S., & Powell, T. C. (2016). Entrepreneurial failure: Statistical and psychological explanations. *Strategic Management Journal*, *37*(6), 1047–1064. doi:10.1002/smj.2378

Backes-Gellner, U., & Moog, P. (2013). The disposition to become an entrepreneur and the jacks-of-all-trades in social and human capital. *The Journal of Socio-Economics*, *47*, 55–72. doi:10.1016/j.socec.2013.08.008

Balkundi, P., & Harrison, D. A. (2006). Ties, leaders, and time in teams: Strong inference about network structure's effects on team viability and performance. *Academy of Management Journal*, *49*(1), 49–68. doi:10.5465/amj.2006.20785500

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512. doi:10.1126/science.286.5439.509

Barberá, P., Wang, N., Bonneau, R., Jost, J. T., Nagler, J., Tucker, J., & González-Bailón, S. (2015). The Critical Periphery in the Growth of Social Protests. *PLOS ONE*, *10*(11), e0143611. doi:10.1371/journal.pone.0143611

Benkler, Y. (2002). Coase's Penguin, or, Linux and "The Nature of the Firm". *The Yale Law Journal*, *112*(3), 369. doi:10.2307/1562247

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. New Haven, CT: Yale University Press.

Benkler, Y. (2016). Peer production and cooperation. In J. M. Bauer & M. Latzer (Eds.), *Handbook on the Economics of the Internet* (pp. 91–119). Cheltenham, UK: Edward Elgar.

Benkler, Y., Shaw, A., & Hill, B. M. (2015). Peer production: A form of collective intelligence. In T. W. Malone & M. S. Bernstein (Eds.), *Handbook of Collective Intelligence* (pp. 175–204). Cambridge, MA: MIT Press.

Bolici, F., Howison, J., & Crowston, K. (2016). Stigmergic coordination in FLOSS development teams: Integrating explicit and implicit mechanisms. *Cognitive Systems Research*. Special Issue of Cognitive Systems Research – Human-Human Stigmergy, *38*, 14–22. doi:10.1016/j.cogsys.2015.12.003

Bryant, S. L., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 1–10). GROUP '05. doi:10.1145/1099203.1099205

Butler, B. S. (2001). Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information Systems Research*, *12*(4), 346–362. doi:10.1287/isre.12.4.346.9703

Butler, B. S., Joyce, E., & Pike, J. (2008). Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1101–1110). CHI '08. doi:10.1145/1357054.1357227

Cassar, G. (2014). Industry and startup experience on entrepreneur forecast performance in new firms. *Journal of Business Venturing*, *29*(1), 137–151. doi:10.1016/j.jbusvent.2012.10.002

Centola, D., & Macy, M. (2007). Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, *113*(3), 702–734. doi:10.1086/521848

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact. 1*(CSCW), 31:1–31:22. doi:10.1145/3134666

Cheung, C. M. K., & Thadani, D. R. (2012). The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems*, *54*(1), 461–470. doi:10.1016/j.dss.2012.06.008

Coleman, J. S. (1990). *Foundations of social theory*. OCLC: 781975612. Cambridge, Mass.: The Belknap Press of Harvard University Press.

Cooper, A. C., Gimeno-Gascon, F. J., & Woo, C. Y. (1994). Initial human and financial capital as predictors of new venture performance. *Journal of Business Venturing*, *9*(5), 371–395. doi:10.1016/0883-9026(94)90013-2

Cress, D. M., McPherson, J. M., & Rotolo, T. (1997). Competition and Commitment in Voluntary Memberships: The Paradox of Persistence and Participation. *Sociological Perspectives*, *40*(1), 61–79. doi:10.2307/1389493

Crowston, K., Annabi, H., & Howison, J. (2003). Defining Open Source Software Project Success. In *ICIS 2003 Proceedings* (pp. 327–340). Seattle, Washington, USA.

Crowston, K., & Howison, J. (2005). The social structure of free and open source software development. *First Monday*, *10*(2). doi:10.5210/fm.v10i2.1207

Crowston, K., & Howison, J. (2006). Hierarchy and centralization in free and open source software team communications. *Knowledge, Technology & Policy*, *18*(4), 65–85. doi:10.1007/s12130-006-1004-8

Crowston, K., Wei, K., Li, Q., & Howison, J. (2006). Core and periphery in free/libre and open source software team communications. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006. HICSS '06* (Vol. 6, 118a). doi:10.1109/HICSS.2006.101

Cummings, J. N., & Cross, R. (2003). Structural properties of work groups and their consequences for performance. *Social Networks*, *25*(3), 197–210. doi:10.1016/S0378-8733(02)00049-7

Cunha, T., Jurgens, D., Tan, C., & Romero, D. (2019). Are all successful communities alike? Characterizing and predicting the success of online communities. *arXiv:1903.07724 [cs]*. doi:10.1145/3308558.3313689. arXiv: 1903.07724 [cs]

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13* (pp. 307–318). doi:10.1145/2488388.2488416

DeSanctis, G., & Monge, P. R. (1998). Communication processes for virtual organizations. *Journal of Computer-Mediated Communication*, *3*(4). doi:10.1111/j.1083-6101.1998.tb00083.x

Dobrev, S. D., & Barnett, W. P. (2005). Organizational Roles and Transition to Entrepreneurship. *Academy of Management Journal*, *48*(3), 433–449. doi:10.5465/AMJ.2005.17407910

Eisenhardt, K. M., & Schoonhoven, C. B. (1990). Organizational Growth: Linking Founding Team, Strategy, Environment, and Growth Among U.S. Semiconductor Ventures, 1978-1988. *Administrative Science Quarterly*, *35*(3), 504–529. doi:10.2307/2393315

Farace, R. V., Monge, P. R., & Russell, H. M. (1977). *Communicating and organizing*. OCLC: 12271833.

Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization Science*, *22*(5), 1224–1239. doi:10.1287/orsc.1100.0614

Felin, T., Lakhani, K. R., & Tushman, M. L. (2017). Firms, crowds, and innovation. *Strategic Organization*, *15*(2), 119–140. doi:10.1177/1476127017706610

Foote, J., & Contractor, N. (2018). The behavior and network position of peer production founders. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds.), *iConference 2018:*

Transforming Digital Worlds (pp. 99–106). Lecture Notes in Computer Science. doi:10.1007/978-3-319-78105-1_12

Foote, J., Gergle, D., & Shaw, A. (2017). Starting online communities: Motivations and goals of wiki founders. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)* (pp. 6376–6380). doi:10.1145/3025453.3025639

Freeman, L. C., Roeder, D., & Mulholland, R. R. (1980). Centrality in social networks: II. Experimental results. *Social networks*, *2*(2), 119–141.

Geiger, R. S., & Halfaker, A. (2013). When the levee breaks: Without bots, what happens to Wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration (OpenSym '13)* (6:1–6:6). doi:10.1145/2491055.2491061

Gibbs, J. L., Kim, H., & Ki, S. (2016). Investigating the role of control and support mechanisms in members' sense of virtual community. *Communication Research*. doi:10.1177/0093650216644023

Gibson, C. B., & Gibbs, J. L. (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), 451–495. doi:10.2189/asqu.51.3.451

Gorbatâi, A. D. (2014). *The paradox of novice contributions to collective production: Evidence from Wikipedia* (SSRN Scholarly Paper No. ID 1949327). Social Science Research Network. Rochester, NY.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., … DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, *310*(5750), 987–991. doi:10.1126/science.1116681

Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2013). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, *57*(5), 664–688. doi:10.1177/0002764212469365

Halfaker, A., Geiger, R. S., & Terveen, L. G. (2014). Snuggle: Designing for Efficient Socialization and Ideological Critique. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 311–320). CHI '14. doi:10.1145/2556288.2557313

Hannan, M. T., & Freeman, J. (1977). The population ecology of organizations. *American Journal of Sociology*, *82*(5), 929–964. doi:10.2307/2777807

Hargittai, E., & Hsieh, Y. P. (2012). Succinct Survey Measures of Web-Use Skills. *Social Science Computer Review*, *30*(1), 95–107. doi:10.1177/0894439310397146

Hertel, G., Niedner, S., & Herrmann, S. (2003). Motivation of software developers in Open Source projects: An Internet-based survey of contributors to the Linux kernel. *Research policy*, *32*(7), 1159–1177.

Hill, B. M. (2013). Almost Wikipedia: What eight early online collaborative encyclopedia projects reveal about the mechanisms of collective action. In *Essays on volunteer mobilization in peer production*. PhD Dissertation. Cambridge, Massachusetts: Massachusetts Institute of Technology.

Hinds, D., & Lee, R. M. (2009). Communication Network Characteristics of Open Source Communities. *International Journal of Open Source Software and Processes*, *1*(4), 26–48. doi:10.4018/jossp.2009100102

Hinds, P. J., & Bailey, D. E. (2003). Out of Sight, Out of Sync: Understanding Conflict in Distributed Teams. *Organization Science*, *14*(6), 615–632. doi:10.1287/orsc.14.6.615.24872

Hinds, P. J., & Kiesler, S. (Eds.). (2002). *Distributed work*. Cambridge, MA: MIT Press.

Hinds, P. J., & Mortensen, M. (2005). Understanding Conflict in Geographically Distributed Teams: The Moderating Effects of Shared Identity, Shared Context, and Spontaneous Communication. *Organization Science*, *16*(3), 290–307. doi:10.1287/orsc.1050.0122

Hirschman, A. O. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press.

Howison, J., Inoue, K., & Crowston, K. (2006). Social dynamics of free and open source team communications. In *Open Source Systems* (pp. 319–330). Springer.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in Organizations: From Input-Process-Output Models to IMOI Models. *Annual Review of Psychology*, *56*(1), 517–543. doi:10.1146/annurev.psych.56.091103.070250

Jo, H., & Lee, J. (1996). The relationship between an entrepreneur's background and performance in a new venture. *Technovation*, *16*(4), 161–211. doi:10.1016/0166-4972(96)89124-3

Johnson, S. L., Faraj, S., & Kudaravalli, S. (2014). Emergence of power laws in online communities: The role of social mechanisms and preferential attachment. *Management Information Systems Quarterly*, *38*(3), 795–808.

Kairam, S. R., Wang, D. J., & Leskovec, J. (2012). The Life and Death of Online Groups: Predicting Group Growth and Longevity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (pp. 673–682). WSDM '12. doi:10.1145/2124295.2124374

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. doi:10.1177/001316447403400115

Kane, G. C., & Ransbotham, S. (2016). Content and Collaboration: An Affiliation Network Approach to Information Quality in Online Peer Production Communities. *Information Systems Research*, *27*(2), 424–439. doi:10.1287/isre.2016.0622

Katz, D., & Kahn, R. L. (1966). *The social psychology of organizations*. Wiley New York.

Katz, N., Lazer, D., Arrow, H., & Contractor, N. (2005). The network perspective on small groups. *Theories of small groups: Interdisciplinary perspectives*, 277–312. doi:10.4135/9781483328935.n8

Keegan, B. (2015). Emergent Social Roles in Wikipedia's Breaking News Collaborations. In *Roles, Trust, and Reputation in Social Media Knowledge Markets* (pp. 57–79). Springer.

Keegan, B., Gergle, D., & Contractor, N. (2013). Hot Off the Wiki Structures and Dynamics of Wikipedia's Coverage of Breaking News Events. *American Behavioral Scientist*, *57*(5), 595–622. doi:10.1177/0002764212469367

Kittur, A., & Kraut, R. E. (2010). Beyond Wikipedia: Coordination and conflict in online production groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)* (pp. 215–224). doi:10.1145/1718918.1718959

Kittur, A., Lee, B., & Kraut, R. E. (2009). Coordination in collective intelligence: The role of team structure and task interdependence. In *Proceedings of the 27th international conference on Human factors in computing systems* (pp. 1495–1504).

Kittur, A., Pendleton, B., & Kraut, R. E. (2009). Herding the cats: The influence of groups in coordinating peer production. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09)* (7:1–7:9). doi:10.1145/1641309.1641321

Klandermans, B. (2004). Why Social Movements Come into Being and Why People Join Them. In J. R. Blau (Ed.), *The Blackwell Companion to Sociology* (pp. 268–281). doi:10.1002/9780470693452.ch19

Kotlarsky, J., van den Hooff, B., & Houtman, L. (2015). Are We on the Same Page? Knowledge Boundaries and Transactive Memory System Development in Cross-Functional Teams. *Communication Research*, *42*(3), 319–344. doi:10.1177/0093650212469402

Krackhardt, D. (1994). Graph theoretical dimensions of informal organizations. In K. M. Carley & M. J. Prietula (Eds.), *Computational organization theory* (pp. 89–111).

Kraut, R. E., & Fiore, A. T. (2014). The Role of Founders in Building Online Groups. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 722–732). CSCW '14. doi:10.1145/2531602.2531648

Kraut, R. E., Resnick, P., & Kiesler, S. (2012). *Building successful online communities: Evidence-based social design*. Cambridge, MA: MIT Press.

Lakhani, K. R., & von Hippel, E. (2003). How open source software works: "Free" user-to-user assistance. *Research Policy*, *32*(6), 923–943. doi:10.1016/S0048-7333(02)00095-1

Lakhani, K. R., & Wolf, B. (2005). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. In J. Feller, B. Fitzgerald, S. A. Hissam, & K. R. Lakhani (Eds.), *Perspectives on Free and Open Source Software* (pp. 3–22). MIT Press.

Lampe, C., Wash, R., Velasquez, A., & Ozkaya, E. (2010). Motivations to participate in online communities. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 1927–1936). doi:10.1145/1753326.1753616

Lazear, E. P. (2004). Balanced Skills and Entrepreneurship. *The American Economic Review*, *94*(2), 208–211.

Lee, C. P. (2007). Boundary Negotiating Artifacts: Unbinding the Routine of Boundary Objects and Embracing Chaos in Collaborative Work. *Computer Supported Cooperative Work (CSCW)*, *16*(3), 307–339. doi:10.1007/s10606-007-9044-5

Levine, S. S., & Prietula, M. J. (2013). Open Collaboration for Innovation: Principles and Performance. *Organization Science*, *25*(5), 1414–1433. doi:10.1287/orsc.2013.0872

Marwell, G., & Oliver, P. (1993). *The critical mass in collective action: A micro-social theory*. Cambridge, UK: Cambridge University Press.

Matei, S. A., & Britt, B. C. (2017). *Structural differentiation in social media: Adhocracy, entropy, and the "1 % effect"*. Lecture Notes in Social Networks. Springer.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*(2), 273–283. doi:10.1037/0021-9010.85.2.273

Mathieu, J. E., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team Effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future. *Journal of Management*, *34*(3), 410–476. doi:10.1177/0149206308316061

McPherson, J. M., & Ranger-Moore, J. R. (1991). Evolution on a Dancing Landscape: Organizations and Networks in Dynamic Blau Space. *Social Forces*, *70*(1), 19–43. doi:10.2307/2580060

McPherson, J. M., & Rotolo, T. (1996). Testing a Dynamic Model of Social Composition: Diversity and Change in Voluntary Groups. *American Sociological Review*, *61*(2), 179–202. doi:10.2307/2096330

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.

Monge, P. R., & Contractor, N. S. (2003). *Theories of communication networks*. Oxford, UK: Oxford University Press.

Monge, P. R., Farace, R. V., Eisenberg, E. M., Miller, K. I., & White, L. L. (1984). The process of studying process in organizational communication. *Journal of Communication*, *34*(1), 22–43.

Monge, P. R., Fulk, J., Kalman, M. E., Flanagin, A. J., Parnassa, C., & Rumsey, S. (1998). Production of collective action in alliance-based interorganizational communication and information systems. *Organization Science*, *9*(3), 411–433. doi:10.1287/orsc.9.3.411

Nanda, R., & Sørensen, J. B. (2010). Workplace peers and entrepreneurship. *Management Science*, *56*(7), 1116–1126. doi:10.1287/mnsc.1100.1179

Narayan, S., Orlowitz, J., Morgan, J., Hill, B. M., & Shaw, A. (2017). The Wikipedia Adventure: Field evaluation of an interactive tutorial for new users. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1785–1799). CSCW '17. doi:10.1145/2998181.2998307

Nov, O. (2007). What motivates Wikipedians? *Communications of the ACM*, *50*(11), 60–64. doi:10.1145/1297797.1297798

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Opp, K.-D. (2011). Modeling Micro-Macro Relationships: Problems and Solutions. *The Journal of Mathematical Sociology*, *35*(1-3), 209–234. doi:10.1080/0022250X.2010.532257

Ostgaard, T. A., & Birley, S. (1996). New venture growth and personal networks. *Journal of Business Research*. Entrepreneurship and New Firm Development, *36*(1), 37–50. doi:10.1016/0148-2963(95)00161-1

Panciera, K., Halfaker, A., & Terveen, L. (2009). Wikipedians are born, not made: A study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work* (pp. 51–60). GROUP '09. doi:10.1145/1531674.1531682

Preece, J., Nonnecke, B., & Andrews, D. (2004). The top five reasons for lurking: Improving community experiences for everyone. *Computers in human behavior*, *20*(2), 201–223.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction*, *1*(1), 13–32.

Qin, X., Cunningham, P., & Salter-Townshend, M. (2015). The influence of network structures of Wikipedia discussion pages on the efficiency of WikiProjects. *Social Networks*, *43*, 1–15. doi:10.1016/j.socnet.2015.04.002

Raban, D. R., Moldovan, M., & Jones, Q. (2010). An Empirical Study of Critical Mass and Online Community Survival. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (pp. 71–80). CSCW '10. doi:10.1145/1718918.1718932

Ransbotham, S., & Kane, G. C. (2011). Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *MIS Quarterly*, *35*(3), 613.

Raymond, E. S. (1999). *The cathedral and the bazaar: Musings on Linux and open source by an accidental revolutionary* (T. O'Reilly, Ed.). Sebastopol, CA: O'Reilly and Associates.

Reagans, R., & Zuckerman, E. W. (2001). Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams. *Organization Science*, *12*(4), 502–517. doi:10.1287/orsc.12.4.502.10637

Reagans, R., Zuckerman, E., & McEvily, B. (2004). How to Make the Team: Social Networks vs. Demography as Criteria for Designing Effective Teams. *Administrative Science Quarterly*, *49*(1), 101–133. doi:10.2307/4131457

Ren, Y., Harper, F., Drenner, S., Terveen, L., Kiesler, S., Riedl, J., & Kraut, R. (2012). Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *Management Information Systems Quarterly*, *36*(3), 841–864.

Ren, Y., & Kraut, R. E. (2014). Agent Based Modeling to Inform the Design of Multiuser Systems. In J. S. Olson & W. A. Kellogg (Eds.), *Ways of Knowing in HCI* (pp. 395–419). doi:10.1007/978-1-4939-0378-8_16

Resnick, P., Konstan, J., Chen, Y., & Kraut, R. E. (2012). Starting new online communities. In *Building successful online communities: Evidence-based social design* (pp. 231–280). MIT Press.

Rogers, E. M. (1962). *Diffusion of Innovations*. New York, NY: The Free Press of Glencoe.

Rogers, E. M., & Agarwala-Rogers, R. (1976). *Communication in organizations*.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 695–704). WWW '11. doi:10.1145/1963405.1963503

Roth, C., Taraborelli, D., & Gilbert, N. (2008). Measuring wiki viability: An empirical assessment of the social dynamics of a large sample of wikis. In *Proceedings of the 4th International Symposium on Wikis* (27:1–27:5). WikiSym '08. doi:10.1145/1822258.1822294

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, *311*(5762), 854–856. doi:10.1126/science.1121066

Schroer, J., & Hertel, G. (2009). Voluntary Engagement in an Open Web-Based Encyclopedia: Wikipedians and Why They Do It. *Media Psychology*, *12*, 96–120(25). doi:10.1080/15213260802669466

Schweik, C. M., & English, R. C. (2012). *Internet success: A study of open-source software commons*. Cambridge, MA: MIT Press.

Scott, C. R. (2007). Communication and Social Identity Theory: Existing and Potential Connections in Organizational Identification Research. *Communication Studies*, *58*(2), 123–138. doi:10.1080/10510970701341063

Seidman, S. B. (1983). Network structure and minimum degree. *Social networks*, *5*(3), 269–287. doi:10.1016/0378-8733(83)90028-X

Shaw, A., & Hargittai, E. (2018). The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing. *Journal of Communication*, *68*(1), 143–168. doi:10.1093/joc/jqx003

Shaw, A., & Hill, B. M. (2014). Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication*, *64*(2), 215–238. doi:10.1111/jcom.12082

Shen, C., Monge, P. R., & Williams, D. (2014). Virtual brokerage and closure: Network structure and social capital in a massively multiplayer online game. *Communication Research*, *41*(4), 459–480. doi:10.1177/0093650212455197

Shi, Y., Dokshin, F. A., Genkin, M., & Brashears, M. E. (2017). A Member Saved Is a Member Earned? The Recruitment-Retention Trade-Off and Organizational Strategies for Membership Growth. *American Sociological Review*, *82*(2), 407–434. doi:10.1177/0003122417693616

Soule, S. A., & King, B. G. (2008). Competition and resource partitioning in three social movement industries. *The American Journal of Sociology*, *113*(6), 1568–1610. doi:10.1086/587152

Stam, W., Arzlanian, S., & Elfring, T. (2014). Social capital of entrepreneurs and small firm performance: A meta-analysis of contextual and methodological moderators. *Journal of Business Venturing*, *29*(1), 152–173. doi:10.1016/j.jbusvent.2013.01.002

Tan, C. (2018). Tracing community genealogy: How new communities emerge from the old. In *Proceedings of the Twelfth International Conference on Web and Social Media (ICWSM '18)* (pp. 395–404). Palo Alto, California: AAAI.

TeBlunthuis, N., Shaw, A., & Hill, B. M. (2017). Density dependence without resource partitioning: Population ecology on Change.org. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 323–326). CSCW '17 Companion. doi:10.1145/3022198.3026358

TeBlunthuis, N., Shaw, A., & Hill, B. M. (2018). Revisiting "The rise and decline" in a population of peer production projects. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)* (355:1–355:7). doi:10.1145/3173574.3173929

Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin*, *63*(6), 384–399. doi:10.1037/h0022100

Tuckman, B. W., & Jensen, M. A. C. (1977). Stages of Small-Group Development Revisited. *Group & Organization Studies*, *2*(4), 419–427. doi:10.1177/105960117700200404

Tushman, M. L. (1979). Work Characteristics and Subunit Communication Structure: A Contingency Analysis. *Administrative Science Quarterly*, *24*(1), 82–98. doi:10.2307/2989877

Van De Ven, A. H., Delbecq, A. L., & Koenig, R. (1976). Determinants of Coordination Modes within Organizations. *American Sociological Review*, *41*(2), 322–338. doi:10.2307/2094477

Van Knippenberg, D. (2000). Work Motivation and Performance: A Social Identity Perspective. *Applied Psychology*, *49*(3), 357–371. doi:10.1111/1464-0597.00020

von Hippel, E., & von Krogh, G. (2003). Open source software and the 'private-collective' innovation model: Issues for organization science. *Organization Science*, *14*(2), 209–223. doi:Article

von Krogh, G., Haefliger, S., Spaeth, S., & Wallin, M. W. (2012). Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development. *MIS Quarterly*, *36*(2), 649–676.

von Krogh, G., & von Hippel, E. (2006). The promise of research on open source software. *Management Science*, *52*(7), 975–983. doi:10.1287/mnsc.1060.0560

Wagner, J. (2003). Testing Lazear's jack-of-all-trades view of entrepreneurship with German micro data. *Applied Economics Letters*, *10*(11), 687–689. doi:10.1080/1350485032000133273

Wang, X., Butler, B. S., & Ren, Y. (2013). The impact of membership overlap on growth: An ecological competition view of online groups. *Organization Science*, *24*(2), 414–431. doi:10.1287/orsc.1120.0756

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods And Applications*. Cambridge University Press.

Wegner, D. M. (1987). Transactive Memory: A Contemporary Analysis of the Group Mind. In *Theories of Group Behavior* (pp. 185–208). Springer Series in Social Psychology. doi:10.1007/978-1-4612-4634-3_9

Welles, B. F., & Contractor, N. (2015). Individual Motivations and Network Effects A Multilevel Analysis of the Structure of Online Social Relationships. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 180–190. doi:10.1177/0002716214565755

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference* (pp. 122–129). iConference '11. doi:10.1145/1940761.1940778

Wiesenfeld, B. M., Raghuram, S., & Garud, R. (1998). Communication Patterns as Determinants of Organizational Identification in a Virtual Organization. *Journal of Computer-Mediated Communication*, *3*(4). doi:10.1111/j.1083-6101.1998.tb00081.x

Wilensky, U., & Rand, W. (2015). *An introduction to agent-based modeling: Modeling natural, social, and engineered complex systems with NetLogo*. Cambridge, Massachusetts: MIT Press.

Wise, S. (2014). Can a team have too much cohesion? The dark side to network density. *European Management Journal*, *32*(5), 703–711. doi:10.1016/j.emj.2013.12.005

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence in teams and organizations. *Handbook of Collective Intelligence*, 143–168.

Zhang, X., & Wang, C. (2012). Network positions and contributions to online public goods: The case of Chinese Wikipedia. *Journal of Management Information Systems*, *29*(2), 11–40. doi:10.2753/MIS0742-1222290202

Zhang, X., & Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at chinese Wikipedia. *The American Economic Review*, *101*(4), 1601–1615. doi:10.2307/23045913

Zhao, H., Seibert, S. E., & Lumpkin, G. (2010). The relationship of personality to entrepreneurial intentions and performance: A meta-analytic review. *Journal of Management*, *36*(2), 381–404. doi:10.1177/0149206309335187

Zhu, H., Chen, J., Matthews, T., Pal, A., Badenes, H., & Kraut, R. E. (2014). Selecting an effective niche: An ecological view of the success of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 301–310). CHI '14. doi:10.1145/2556288.2557348

Zhu, H., Kraut, R. E., & Kittur, A. (2012a). Effectiveness of Shared Leadership in Online Communities. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 407–416). CSCW '12. doi:10.1145/2145204.2145269

Zhu, H., Kraut, R. E., & Kittur, A. (2012b). Organizing without formal organization: Group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 935–944). CSCW '12. doi:10.1145/2145204.2145344

Zhu, H., Kraut, R. E., & Kittur, A. (2013). Effectiveness of Shared Leadership in Wikipedia. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*(6), 1021–1043. doi:10.1177/0018720813515704

Zhu, H., Kraut, R. E., & Kittur, A. (2014). The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)* (pp. 281–290). doi:10.1145/2556288.2557213

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x