Why Wiki Works:

Peer Production and Making Knowledge the Wiki Way

A Dissertation Presented

by

Michael Restivo

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Sociology

Stony Brook University

May 2014

Stony Brook University

The Graduate School

Michael Andrew Restivo

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Arnout van de Rijt, Ph.D. – Dissertation Advisor
Associate Professor, Department of Sociology

Michael Schwartz, Ph.D.
Distinguished Teaching Professor, Department of Sociology

Ian Roxborough, Ph.D., Chairperson of Defense
Professor, Department of Sociology

Patrick Grim, Ph.D.
Distinguished Teaching Professor,
Department of Philosophy, Stony Brook University

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Why Wiki Works:

Peer Production and Making Knowledge the Wiki Way

by

Michael Andrew Restivo

Doctor of Philosophy

in

Sociology

Stony Brook University

2014

With the widespread adoption of advanced information and communication technologies, the past decade has witnessed a proliferation of online organizations that aim to create public goods through voluntary collaboration among self-organized, networked individuals. This form of online collective action has been termed peer production. One of the most successful examples of peer production, and the most widely studied, is the free online encyclopedia Wikipedia, which is written in a collaborative fashion by hundreds of thousands of volunteers and used by millions of people daily. In this dissertation, I use Wikipedia as a strategic case site to critically reflect on the state of our knowledge about how peer production works and to develop a series of empirical studies that offer insights into several puzzles in the literature. These include questions of what motivates participation in online collectives, how emergent organizational structures shape patterns of participation, and how interactional aspects of participation relate to characteristics of the goods being produced. In each case, I argue that we must take into account – both theoretically and methodologically – the high degree of inequality in participation that we commonly see in such organizations. My findings reveal insights about the origins as well as consequences of participation inequality.

Table of Contents

List of Tables

List of Figures

Acknowledgments

This dissertation stems from intellectual collaborations with many people to whom I owe a debt of gratitude. The doctoral path is arduous and long, and much longer than just time spent in graduate school. Looking back, I recognize how I was started down this path by many professors who influenced me along the way since my first days in college. I would like to acknowledge in particular David Courtney and Joyce Poblete, who helped steer me to graduate school to begin with, and David Dickens who provided me with my foundational training in sociological theory and who encouraged me to continue onwards to become an independent scholar in my own right. I also extend my appreciation to my professors in graduate school, including John Shandra, Daniel Levy, Eileen Otis, Michael Kimmel, Diane Barthel, Said Arjomand, Ken Feldman, and Carrie Shandra.

I would also like to extend a heartfelt thank you to my fellow grad students who took classes alongside me, reviewed my papers, and supported me in many other ways over thes years. I also cherish the many friendships that have developed along the way, including Paul Bugy, Rachel Kalish, Misty Curreli, Gary Maynard, and Eric Shircliff. A special thanks to Ljuban Jaksic, Sean Michelsen, Eon Kotze, and Tony Young for keeping me sane throughout the process and for supporting this poor college student over more than a decade of friendship.

In particular, my project was shaped and developed in close collaboration and under supervision of my advisor Arnout van de Rijt. It is still not entirely clear to me how it was possible for Arnout to devote so much time and attention to my projects while advising other students as well as pursuing his own research agenda. I have not ruled out the possibility that he secretly has an identical twin. I also strive to emulate the other members of my committee m – Michael Schwartz, Ian Roxborough, and Patrick Grim – and thank them for their direct assistance and guidance as well as for serving as models for clear and expansive thought. It was a geninue pleasure and honor to work with such extraordinary scholars who were always eager to discuss my work and provide generous feedback with good humor.

Work from Chapter 2 has been published as articles in *PLoS ONE* and *Information, Communication & Society*. I am grateful to the reviewers and editors for their helpful comments and guidance. These articles were produced in true collaborative fashion with Arnout van de Rijt. I would also like to thank Damon Centola, Ori Heffetz, Georgi Kossinets, Michael Macy, Ian Roxborough, and Michael Schwartz for comments, and the Wikimedia Foundation and community for their advice and for facilitating access to the data. I would also like to acknowledge Travis Kriplean for sharing his Barnstar data with me, which was helpful when I was just getting this project off the ground. Work from Chapter 3 was reviewed at *Social Science Research*. I thank the anonymous reviewers and editor for their suggestions regarding how to frame the argument and enhance the analysis. Chapter 4 is based largely on research supported by a National Science Foundation grant. I thank the program officer and anonymous reviewers for their guidance during several rounds of unsuccessful applications, which contributed greatly to the construction of the proposal which was eventually funded.

My family continues to be a source of love and support. I would like thank my mother and father, Andrew and Linda, for their patience and encouragement as I found my path in life, for instilling in me a love of learning, and for teaching me the value of hard work. To my sister Liz, I am inspired by your example of how to live life with exuberance for what you are passionate about. To my sister Carolyn, I am forever grateful for your demonstrations of what can be achieved with determination and persistence – and no matter how hard I try, I will never be as skilled a pianist as you are! And to Andrew Dobson, sorry that you must tell your friends and colleagues that your brother-in-law is a sociologist.

Finally, I could not have made it to the end of this journey without the love, support, and devotion of Amy Braksmajer who has been with me, literally, from day one. Thank you more than words can say.

Chapter 1. Why Wiki Works

> A zeroth law is a codification of something so obvious or fundamental that it was not originally stated. On Wikipedia, the zeroth law is that editors are the most valuable resource. Some would say the articles – but it takes editors to write articles. Some would say reliable sources. Editors are the ones who finds reliable sources and incorporate them into articles. Some would say the neutral point of view. It takes editors to find additional resources to balance the POV resources, discuss balance on the article talk page, and craft the final product. Editors read copyrighted material and restate the concepts to prevent copyright violations. Editors take or find the pictures, upload them, CC release them or write fair-use justifications. Some would say it's the Wikipedia is an encyclopedia pillar. Editors are the folks who detect and remove original research and fancruft.[1]

Problematizing Wikipedia, the free encyclopedia that anyone can edit

Launched in 2001, Wikipedia is a project to create a free, online, open-collaboration encyclopedia. It has grown in a short amount of time from a small project to become a global phenomenon, with more than 250 native-language versions of the encyclopedia being written. The English language Wikipedia, which is the largest, contains more than four million articles as of 2014. Seven other language editions of the project have over one million articles, and 18 contain more than 200,000 articles.[2]

The novelty behind the success of Wikipedia is that anyone may contribute to it by editing its articles, as suggested by its informal motto, "the free encyclopedia that anyone can edit." Keep in mind that each language edition of Wikipedia is an independent project with its own volunteers who create, curate, and edit its selection of articles. Rarely are they mere translations of one another, but rather unique products made by volunteers from around the world. In this dissertation, I focus on the English Wikipedia, which is the first and largest edition – both in terms of the scope of its contents, as well as the size of its contributor base. The English Wikipedia has more than 80,000 unique contributors, also known as *Wikipedians* or simply editors, who are involved with the project on a month-to-month basis, with a stable core of about 10% who are highly engaged and very active.

One of the first major sources of publicity for the site came from an investigation by the journal Nature to discern the quality of Wikipedia's contents (Giles 2005). A sample of articles about topics in science *were* gathered from both Wikipedia and from Encyclopedia Britannica and sent out for blind review to Nature editors. The reviewers found that the error rate for the Wikipedia entries was only slightly higher in Wikipedia as it was in the corresponding Britannica entries. Furthermore, as part of the study *Nature* surveyed more than 1000 scientists who have had their work published in the journal; more than 70% had heard of Wikipedia and

[1] This introductory quote comes from an essay written by user:NE_Ent entitled the "Zeroth Law of Wikipedia"

[2] The editions with at least one million articles: German, Spanish, French, Italian, Dutch, Polish, Russian, and Sweedish. The data come from the Wikimedia Foundation.

nearly 20% consulted it on a weekly basis. This study lends support for the idea that Wikipedia's articles contain useful information and also that its use is widespread, including among experts in the field of science.

This study was not entirely uncontroversial, with some critics pointing out serious errors and flaws in Wikipedia's entries. However, since Wikipedia is more than just a static encyclopedia, its content were being continuously edited, revised, and augmented. These errors were often quickly corrected, once public attention was brought to them, demonstrating the quip, "given enough eyeballs, all bugs are shallow" (Raymond 1999). Wikipedia is also constantly being updated with new information. For example, in the wake of the campus shootings at Virginia Tech in April 2007, more than 2000 volunteers worked together to create a well-sourced article (including more than 140 footnotes) detailing information about the events and subsequent news coverage, as it unfolded. Nearly a million people visited the site within days of the shootings to learn more about the tragic event (Cohen 2007).[3]

Wikipedia has become the largest and most widely-used general reference source in the world. A Pew Study in 2007 reports that a third of U.S. Internet users have consulted Wikipedia at some point, with more than 10% relying on it for daily use (Rainie and Tancer 2007). Wikipedia remains one of the most frequently accessed Internet websites in spite of the fact that it is completely a non-profit that does not display advertisements on its site nor sells its product. Instead, it relies solely on charitable contributions to pay for their technical server costs and pure volunteer labor to perform all of the work on its encyclopedia articles.

Wikipedia is similar to other general-content encyclopedias, like the Encyclopedia Britannica, in that both comprehensively cover a wide range of topics. But the process by which each encyclopedia is produced is radically different. Leading encyclopedias, like Britannica, employ a small editorial board that exerts control and oversight of the entire project. The editors seek out leading figures in the fields of art, science and history to write the articles to be included in the compendium. At latest count, Britannica commissioned more than 6,000 experts to pen its articles.

The encyclopedia was not always constructed in such a fashion. The history and purpose of the encyclopedia stretches back to the Enlightenment. Diderot and d'Alembert saw their *Encyclopédie* as a summary of progressive thought and a place in which readers, by exploring cross-references, could participate in the ongoing conversation of expanding human knowledge. This stands in contrast to the more modern conception of the encyclopedia, one that symbolizes expert summary of universal or total knowledge about a field (Yeo 2001). Wikipedia encompasses both of these ideals: it encourages readers to participate in creating a compendium of as broad a range of human knowledge as possible.

Wikipedia is an example of peer production, which is a relatively new form of organization and collective action that is coordinated online, where people come together to produce a good or service that is generally made available to the public for free. In the case of Wikipedia, we should keep in mind that there are many ways to conceptualize what we mean by "Wikipedia" – are we referring to the organization, or the encyclopedia being created, or the culture of the community? In a similar fashion, since the new millennium, significant scholarly attention has been devoted to these three questions as well, with the body of literature containing more than 2000 peer-reviewed journal articles and conference proceedings.

[3] The fact that individuals may contribute to Wikipedia even while remaining anonymous has lead to some abuse, including a high-profile case where it was revealed that a frequent contributor claimed credentials that he did not have (Cohen 2007). Other common responses to Wikipedia is that it is an inherently unreliable source of information, systematically biased, or hopelessly incomplete and subject to frequent vandalism (Kirtley 2006).

The state of the field is one of fragmentation among many disciplinary approaches and substantive findings, although there is also agreement on many points. A large stream of research on peer production I term the technical literature, which includes things such as analysis of Wikipedia's technical properties, mapping its hyperlink network structure, testing informational retrieval and extraction properties, database structure, and so forth (Martin 2011; Medelyan et al. 2009; Voss 2005); additional line of research focused the design of human-computer interaction and computer-mediated work (Halavais and Lackaff 2008; Kittur, Pendleton, and Kraut 2009a). From a social science perspective, the most influential early works from economics and organizational studies include, in particular, Yochai Benkler, who offered a most cogent framing of the problem for social scientists: the purpose of research on peer production is to identify and understand "characteristics that make large-scale collaborations in many information production fields sustainable and productive in the digitally network environment with reliance either on markets or managerial hierarchy" (Benkler 2002:374). Benkler suggests that the "problem" that such organizations need to solve is how to get participants to determine what they should do, and actually do it. According to the framework of transaction cost economics introduced by Ronald Coase (1937) the question of how to organize individual agents is solved by markets through price signaling and within firms internally through hierarchical command (Williamson 1979, 1981, 1998). Economic sociologists challenged this typology by offering a third generic form of organization – networks – that does not focus strictly on economic factors and indeed saw markets and firms as two specific configurations of networks (Granovetter 1973; Podolny and Page 1998; Powell 2003). The generic properties of networks are that parties operate in relation to others, where reputation, interdependence, and reciprocity are integral parts of their relationships over and above economic considerations.

But this conceptualization of networks is not the same as Benkler envisions them. Benkler's focused on new forms of decentralized, digitally-enabled networks which consist primarily of weak ties of spontaneous collaboration between strangers or peers. Until the emergence of peer production on the Internet, the tripartite view – markets, firms, and networks – was capable of explaining the structure of most organizations. While bearing some resemblance to networks, decentralized digital networks are a problematic fit into this typology, hence their separate designation as peer production.

Like markets, firms, and networks, peer production contains an implicit contractual framework within which productive activities take place. Markets are governed by contract and property law while hierarchies rely on employment contracts and bureaucratic fiat. Networks, being the more generic conceptualization, emphasized trust and the non-legal bonds of mutual social obligation among parties. Nonetheless, even this remains a poor fit with peer production, where individuals were bound together by their shared purpose to create, as economists put it, "strong positive network externalities" (Demil and Lecocq 2006) – that is to say, creating a valuable good which they give away for free. Contributors are tied together by the contractual framework of the open source license or "copyleft." A play on the term "copyright," copyleft is a strategic use of existing property or copyright law that aims to accomplish the opposite of its usual intended purpose. Rather than being used to restrict ownership, the copyleft or open-source license becomes a contractual means of ensuring that what is produced is made available and accessible for all those who seek to use it. The fact that the product of economic activity is made freely available explains Benkler's emphasis on the "commons-based" characteristics of peer production, which challenges the economic assumption that controlling an asset is necessary for productive activity and extracting economic value from it.

Given this distinct governance framework, the actual structure of governance in the organization, as a result, must be different as well. For example, peer production organizations typically permit and encourage the widest possible pool of contributors to take part in the productive activity, since many hands make light work. By contrast, parties who fail to respond to price signals are “priced out” of markets, firms can only hiring a limited number of individuals while excluding others from employment, and networks can only be mobilized through social ties among participants, effectively limiting opportunities for parties who don't possess the prerequisite social capital. Peer production, by contrast, imposes few mechanisms for excluding parties from contributing to productive work if they so desire. Within this open environment, “a great babbling bazaar of differing agendas and approaches” (Raymond 1999:30), the responsibility for governance resides within the entire community of participants, constrained only by copyleft in absence of other legitimate organizational authority. Demil and Lecocq (2006) use the term “bazaar governance” to describe the responsibility where many parties must simultaneously coordinate their actions with each other in absence of legitimate organizational authority, and where any party (or peer) may join in, with equal standing, the production of the nominal goals of the organization.

Without strong incentives, like financial compensation, with few control mechanisms that come from reference to legitimate authorities, and without strong social ties to maintain exchange relationships, to understand why people contribute to peer production, scholars turned to theories of voluntarism (Wilson 2000). Aside from the generic individual level and structural factors that enable or constrain voluntarism more generally (Dolnicar and Randle 2007; Smith 1994), scholars wanted to understand the particular characteristics of participants in peer production. First were social-psychological motives of contributors (Hars and Ou 2002; Nov 2007; Schroer and Hertel 2009) who cited things such as fun and the hedonistic pleasure of solving complex intellectual challenges (Lakhani and von Hippel 2003; Schroer and Hertel 2009). Some forms of peer production, such as open source software development, also offer contributors the opportunity to increase their human capital by learning valuable new skills with the promise of being able to apply these to their careers or hobbies (Hemetsberger and Pieters 2001; Hemetsberger and Reinhardt 2006; Ye, Yamamoto, and Kishida 2004), although these conditions do not seem to be pertinent to Wikipedia per se (as it would be unusual to see someone list *Wikipedia contributor* on their resume). Nonetheless, participants also can accrue considerable peer-recognition and reputation, which confers status and prestige to contributors as a reward (Stewart 2005; Willer 2009). Bryant, Forte, and Bruckman (2005) even argued that Wikipedia contains an incentive system similar to the “cycle of credit” in the scientific community as described by Latour and Woolgar (1986), despite the limitation that individual recognition circulates only within the online community which consists of mostly of amateurs (Bryant et al. 2005; Nov 2007; Oreg and Nov 2008). Finally, for Wikipedia in particular, many contributors report feeling they call “wikiholism,” a portmanteau that captures the addictive nature of the volunteer work. In an essentially unbound system where individuals may apply their skills, energy, and time as they see fit, there may always be rewarding work to do that benefits the individual doing it. Indeed, Benkler identifies this as one of the most crucial components for peer production to succeed: that its work is modular (can be broken up into small pieces) which can be accomplished by individuals independently of one another.

In addition, the ideological component of volunteering should not be understated, since as I described, the idea of open source license underpins the permanence and universal accessibility of the product being created. Interview subjects in Kuznetsov (2006) and Shroer and Hertel (2009) suggest shared norms and a common outlook is one of the most powerful reasons for volunteering. Contributors often share the belief that it is desirable and admirable to share the information that one has, an ideal which originated in the computer sciences with the open-source software movement, which emphasized that access to information should be available to all who seek it (Stallman 1999). This open-source ethos or “hacker ethic” lies at the heart of peer production's copyleft license, which is the general framework within which participants can act in an altruistic manner for the betterment of humankind (Bitzer, Schrettl, and Schröder 2007; Glott, Ghosh, and Schmidt 2010; Hemetsberger and Pieters 2001; Ye and Kishida 2003; Zeitlyn 2003). In this regard, altruism is the a moral incentive that can substitute for material incentives found among other types of productive organizations

(Etzioni 1988). Contributors share their time and knowledge in order to provide a resource for all to use (Bryant et al. 2005; Hars and Ou 2002; Lakhani and von Hippel 2003).

To supplement the focus on individuals, scholars looked at aspects of the community and the process of community building. Konieczny (2009b) asked whether Wikipedia was a community or a social movement, and concluded that the expressed values, philosophies, and policies put in place by its contributors fostered a collective identity that sets itself apart as a community. But Konieczny also saw how Wikipedia was part of a wider, global social movement advocating free culture, FLOSS (free-libre open source software), and other copyleft ideas. Pentzold (2011) similarly interrogated what Wikipedian contributors meant when they identify themselves as part of a community, concluding that contributors understand their membership in the Wikipedia community as an "ethos-action" where membership is defined as "personal acceptance of a set of moral obligations and rules of conduct" (p. 13). Some saw Wikipedia's open membership policy, and emphasis on community discussion and consensus decision making, as a form of Habermas' ideal of rational discourse (Hansen, Berente, and Lyytinen 2009) while others saw Wikipedia as a model *adhocracy* (Konieczny 2010). Other scholars attempted to contrast these ideals with the actual practice of Wikipedians, finding some discrepancies; for example, the role of policies in compelling consensus (Kriplean et al. 2007) and the potential for control over policy making for oligarchical purposes (Konieczny 2009a). Still, it is interesting to note that Konieczny (2010) concludes that community governance emerges largely out of coordinated discussion and consensus rather than being predetermined by policy structure.

Such a review as I have offered of existing social science literature on Wikipedia inevitably remains incomplete, since it remains an active field of research and because there are so many interrelated topics being investigated. The approach that I take throughout this dissertation is to strategically poach from a distinct number of research traditions without embracing one as consistent framework throughout. I draw from social psychological work on motivations for cooperation, microeconomics, research on group processes, organizational theory, human-computer interaction, and theories of gender and social interaction. My rationale in this regard is that I believe it is self-limiting to attempt to impose one consistent frame across a multitude of different questions, which I believe would only yield the usual answers that emphasize the differences among these perspectives; I would rather recognize how each can contribute something of value to our understanding of Wikipedia as a whole, even if this means that my treatment of any individual area is less than comprehensive. Similarly, there are other fields that I do not draw on, despite ongoing research agendas on these areas on the same topic of this dissertation. Foremost of these is the growing literature on the broader political economy of peer production (Bauwens 2005, 2009; Benkler 2006; Weber, Latham, and Sassen 2005; Weber 2000), as well as the interest from scholars of organizational ecology (Healy and Schussman 2003; Krishnamurthy 2002; Lanzara and Morner 2003), social movements (Bonaccorsi and Rossi 2006; Konieczny 2009b; Morell 2012), and even military theorists (Arquilla and Ronfeldt 1996, 2001) about the potential of such peer-to-peer organizations to enact social change.

Purpose of the dissertation

Having dispensed with my introductory remarks on the background of Wikipedia, as well as briefly reviewing the literature on peer production and delimiting what this dissertation is not about, I shall now discuss the purpose of this work. One of the recurring themes in the literature on peer production is the presence of persistent patterns of inequality. The overarching purpose of this dissertation, then, is a sociological reflection on the theories and methods we use to explain and understand the origin and consequences of these inequalities, as well as a series of empirical studies of different forms of participation inequality.

My focus on inequality is not because it reflects an inherent tendency of sociologists, nor for the reason of exposing an inconsistency between the inclusive, democratic ideals of the community and its actual practices. Instead, I put forth the suggestion that since inequality represents such a regularity of social life, in all the various instances where it occurs, that if our desire is to understand the social world, we should task ourselves with studying the social processes that generate such inequalities, how they are reproduced, and what such inequalities mean to the participants themselves.

I should be clear about what type of inequalities are of interest here. The ones that I focus on are related to participation – that is, the actual taking part in the process of creating and producing Wikipedia, in large or small ways – instead of merely use of Wikipedia. It is true that the divide between producers and consumers itself follows a predictable pattern of inequality. A common characteristic of online peer production is that many more people participate as consumers than as producers: one rule-of-thumb (Brothers et al. 1992; Whittaker et al. 1998) suggests that of the people who visit online peer production organizations, about 90% do so primarily as consumers while only around 10% produce any content for the site. A Pew Study (Zickuhr and Rainie 2010) found that nearly 1 in 2 Americans have consulted Wikipedia, yet since its inception, fewer than 5% of U.S. Internet users have registered an account with the site. However, patterns of use of Wikipedia is not my interest here.

Instead, I focus on multiple forms of inequality in participation. The first type of participation inequality I address is how work is distributed. Since this is a volunteer organization, work is not assigned to participants, but rather it is through their self chosen work that the whole organization, its structure, and its product are created and sustained. Among the volunteers who contribute to Wikipedia, research has shown that about 90% of all the work effort on Wikipedia is performed by the top 10% of all contributors (Ortega, Gonzalez-Barahona, and Robles 2008; Voss 2005). Understanding what creates and maintains this divide between very active contributors and the much larger group of infrequent contributors is a puzzle present in much research on peer production.

A second, closely related form of inequality is that of social status within the community. A small number of prodigious contributors have accumulated a great deal of recognition which boosts their social standing within the community, while most others receive little recognition for their efforts. The link between the distribution of work and earned status in the community has been taken as support for the notion that social standing is broadly meritocratic and that those who do the most work receive the most recognition. Status attainment may also drive inequality in the distribution of work, as receiving social recognition is thought to be a selective incentive that motivates further participation (Willer 2009), creating a positive feedback mechanism whereby work and status become concentrated at the top of the distribution. However, an alternate mechanism may be at play, since community members tend to evaluate a contributor's merit relative to existing social references such as that person's current status (Stewart 2005). Thus, cumulative advantage suggests that higher status actors will receive a disproportionate share of recognition for their efforts relative to lower status actors, in spite of performing less work. Establishing the cause and effect relationship between work and social recognition, and whether it is a consequence of participation inequality or generative of it, is the purpose of Chapter 2.

High degrees of participation inequality poses particular challenges for empirical research. One concern is the difficulty of generalizing from particular findings because the distribution of work efforts are so highly unequal that as a result, it is hard to characterize an "average" contributor in such a large and heterogeneous population. For example, in-depth studies from interviews with a small numbers of contributors may not be representative of the larger population (Kuznetsov 2006), whereas surveys of online contributors may suffer from certain types of response bias and non-representativeness (Hill and Shaw 2013). Many explanations either implicitly or explicitly focus on only parts of the population, without trying to bridge their explanations across the participation gap, or doing so in ways that are methodologically tenuous.

Another process thought to generate participation inequality focuses on increasingly formalized organizational structures that can shape pattern of participation for newcomers. As Wikipedia's contributor community grew in size during the first several years of its operation, it is believed that contributors were compelled to establish a set of informal bureaucratic procedures, guidelines, and policies whose purpose was to help manage the new recruits; yet paradoxically, this made it harder for such new contributors to "break in" to the core community. These bureaucratic structures were inimical to newcomers, who perceived rules as being burdensome or antagonistic to why they started contributing to the project in the first place, leading to very short volunteer tenures. On the other hand, the increasingly formalized bureaucratic structures were thought to turn existing community members in small-minded bureaucrats who were more intent on managing the efforts of others than on contributing new content on their own. This dynamic is believed to be behind the decline in contributor retention over time, and is also consistent with the large divide between contributors that we see on the population level. In Chapter 3, I empirically test propositions about how bureaucracy relates to participation inequality.

What is shared in common across these two empirical chapters is that I attempt to synthesize the micro and macro level scale, connecting the dynamics of contributors to the emergent organizational level structures that ultimately constrain their behavior to reproduce the forms of participation inequality that we see. However, in my final empirical study, I address a different type of inequality, over and above how work is distributed among contributors. In Chapter 4, I look at the stark contrast between men's and women's rates of participation in Wikipedia and their experiences doing while so (Cohen 2011; Cooper 2006). This issue of the "gender gap" in participation has received much attention as of late: less than 20% of Wikipedia contributors are women, and discussion of the gender gap has focused on how to boost that number as well as why it is so low to begin with (Gardner 2011). This has turned attention on the unequal treatment of women who are part of Wikipedia's community, which at time can be perceived as unwelcoming or outright hostile to women (Reagle 2012). The consequences of this gender gap has mainly been talked about in terms of differential coverage of topics along gender lines (i.e. fewer articles about important women scientists, musicians, and politicians). While I agree that this is an important issue for the community to address, particularly as it lays bare the contradiction between the *ideals* of a culture in which all are invited to participate, with the *reality* of that culture which can be inimical to diversity, there are other important unanswered questions about this type of participation inequality. In particular, I ask whether this lack of gender diversity in participation is internally detrimental to the accomplishment of Wikipedia's stated goals, that is, to create a quality encyclopedia. In Chapter 4, I address the puzzle of how gender diversity among contributors who author Wikipedia's articles affects the achievement of group goals.

There are of course other forms of participation inequality, such as the divide between regular contributors and those who seek elevated privileges or otherwise act as authorities in the community (Reagle 2007). But the three described here are sociologically relevant and have all been discussed in the existing literature, yet in not altogether convincing ways, as I detail in each chapter. Furthermore, the way in which these inequalities are understood, and more importantly, how they addressed methodologically during the design of research, can significantly alter the results of research and hence shape what we come to believe about how peer production operates. In this next section, I briefly reflect on some of the opportunities and challenges of conducting research on Wikipedia, and consider why existing approaches may be insufficient for us to conclude that we truly understand why Wikipedia works.

In this dissertation, I use Wikipedia as a strategic research site to examine characteristics and social dynamics of inequality in peer production. In Merton's usage (1987), a strategic research site is a social phenomenon that exists "to such an advantage and in such accessible form that it enables fruitful investigation of stubborn problems and discovery of new problems for further inquiry" (pp. 10-11). I strive to demonstrates both aspects of Merton's definition in this work: by using comprehensive micro-level data on contributions to Wikipedia, drawn from both experimental and observational studies, I aim to test theories about what compels individuals to make contributions to this public good, examine how the contributor community coordinates and governs their work and interactions, and discern how its group dynamics shape the product being created. In turn, my research points to several important theoretical and methodological implications for future scholarship on this topic.

The selection of Wikipedia as case, as well as its qualities as a strategic one, is not something that should merely be taken for granted. However, on the face of it, Wikipedia's phenomenal growth, global research, and near universal use do make it a compelling site for research. Each day, millions of people rely on the information in Wikipedia in order to better understand the world around them. Yet, what they are reading is the result of the collective efforts of hundreds of thousands of volunteer authors who have collaboratively written its content. Wikipedia represents one of the largest volunteer projects in the world whose aim is to produce a public good. There are also practical considerations that make Wikipedia strategic for sociological research. A full record exists of all the activities, behaviors, and interactions that people engage in while creating Wikipedia, stored in its publicly-accessible database. As such, it represents an opportunity to study with nearly perfect data the evolution of one of the most expansive projects of peer production ever conceived. Conducting large-scale analysis, in order to discern organizational structures that emerge from micro-level patterns of communication and interaction, is a primary goal of the emerging practice of computational social science (Lazer et al. 2009), which serves as a key reference for my work.

However, I also use Wikipedia to demonstrate some limitations and caveats of this approach to social science research. Weber (1922) proposed that causal explanations in the social sciences need to have several parts. The first is that explanations need to be "*adequate on the level of meaning*" in which the meaning that social actors invest in their own behavior is understandable to them as well as to an observer of the motives and reasons behind their behavior. This defines Weber's interpretive component of explanations, which he argues must be supplemented with a "*causally adequate*" component in which particular actions undertaken by social actors yield predictable consequences and effects. This second component has to do not only with the statistical probability of co-occurrences, but also comparisons between the actual consequences of social actions with what would (or might) have happened if social circumstances were different of other social factors were present or missing. The causal component of explanations should consider counterfactuals and case comparisons, and not simply the more formal assessment of "uniformities" in the social world through statistical analysis that have become predominant in social science research. Weber argued that the purpose of sociology is to present a united explanation and understanding of social phenomena.

The argument I put forth here is that there remains a tenuous link between our explanations and understanding of phenomena like Wikipedia. There have been several attempts to rectify this situation, most notably the unified account of online collective action offered by Shaw (2012). The crux of his argument is that "interactions and interactional dimensions of behavior play a central role in mobilizing, retaining, and organizing participants engaged in online collective action. Interactional motives and incentives not only mobilize participation in online collectives, they also contribute to the emergence and persistence of generic participation inequalities and collectives' organizational forms" (p. 9). Since there is no reason for me to tread over ground that has already been covered, my intention is not to offer a more definitive account that attempts to unite what we can explain about how Wikipedia works with a more interpretive understanding of why Wikipedia works the way it does.

On the contrary, I aim to offer a critical commentary on the empirical basis for the claims to causal adequacy that exist in much of the literature. In particular, each of the three empirical chapters addresses one dimension of my methodological critique, and takes aim at some of the various things we "know" about Wikipedia. Since Wikipedia exists as an online organization with voluminous digital-trace data, the methods of computer science, statistics, physics, and other "big data" techniques have been the primary forms of explanation of its structural features and organizational-level processes. On the other hand, since Wikipedia's most valuable resource (as the opening quotation in this chapter reminds us) is its editor community, the tools of the social sciences, such as interviews, surveys, and other forms of statistical analyses of sub-populations of contributors, remain the most frequently used means by which we understand why Wikipedia works the way it does. There remains a degree of incongruity between explanations and understanding across these two dimensions. Before turning to a more elaborate discussion of potential methodological limitations of past research and the approach I take to address these in this dissertation, I want to take a moment to reflect on Heisenberg's reminder that "since the measuring device has been constructed by the observer, we have to remember that what we observe is not nature itself but nature exposed to our method of questioning" (Heisenberg 1958:58). In the next section, I develop an extended criticism of the many difficulties we face when trying to bridge this divide.

Methodological considerations

One of my observations about the research literature on peer production is the divide between explanation and understanding, and in particular, whether the ways in which these two parts have been tied together have been appropriate or satisfying. I believe there are a number of reasons why we should still regard this link as being not fully developed.

Big and little data. The first point that I wish to discuss is one of the contemporary research frontiers in the social sciences that goes by the euphemism "big data," a term that gained widespread currency after Anderson's (2008) provocative article entitled "The End of Theory" in which he argues (in sound-bite form) that "correlations are enough" when the sheer volume of data "forces us to view data mathematically first and establish a context for it later." In response to this article, philosophers and historians of science found much to ridicule, with Pigliucci (2009) for example going as far as to say that Anderson "doesn't understand much about either science or the scientific method." Nonetheless, the underlying idea of big data – that the volume of data being collected and available for scientists to analyze has been increasing exponentially, while our models, theories, and techniques have struggled to keep up – has gained some traction; of particular interest here is support for this approach in the social sciences (Giles 2012; Halevi and Moed 2012; Snijders, Matzat, and Reips 2012), although one of the main concerns that continues to be expressed is who has the access to such big datasets and the ability to analyze them (Huberman 2012).

My stance on the big data question, and the accompanying computational approaches used to study them, is mild skepticism; but I will leave the ontological and philosophical questions to others who are more qualified, and instead I concern myself with the consequences of such approaches being used to study of phenomena like Wikipedia. Does the use of big data inherently mean an emphasis on explanation (that is to say, statistical correlations in the empirical data) over understanding? I do not believe this is the case. As I described, Shaw (2012) demonstrated a sustained and coherent link between explanation and understanding of online collective action, which is the model I believe sociologists should take.

In doing so, we must confront this large body of research, which I described earlier. Proponents of this approach to research claim that it offers the most cromulent link between large-scale explanation and more interpretative understanding. But, for example, what are we supposed to make of the finding from "On the Inequality of Contributions to Wikipedia" (Ortega et al. 2008), which shows a stable pattern of inequality on Wikipedia over time using a series of Gini coefficients? Can we combine this insight with those from an "online survey of 106 contributors" (Schroer & Hertel, 2009) that investigates the motivations of contributors? Are motivations related to inequality – as a consequence or a cause – or not related at all?

Even works that do attempt to bridge this divide are not altogether satisfying, although I laud them for their efforts. The difficulty here is that there is no commonly accepted way of connecting micro-level understandings to macro-level explanations; what I mean is, there remains a plethora of ways to conceptualize and operationalize key variables and an inconsistent set of methodological approaches to test theories. As I have described, this leads to the problematic situation where two sets of findings can contradict each other while both being plausibly true at the same time, as was the case in my own work compared to a similar study (Hill et al. 2012a; Restivo and van de Rijt 2012).

The situation I have just described is not a problem unique to peer production or big data. But big data has tipped the scales in favor of abstracted, statistical explanations that offer less in the way of meaningful adequacy. Is this not the same problem that Mills (1959) described more than half a century ago? The recurrence of this tendency in our discipline should give us pause. Rather than continuing to focus on this imbalance, I want to now turn to a number of potential methodological challenges or deficiencies that can be present in such large-scale statistical analyses.

Correlation and causation. As I mentioned, Anderson infamously contended that correlation was enough. But nonetheless, we cannot help but be interested in causation – or at the very least, proffering (and testing) likely causal mechanisms that are consistent with the correlations we observe. While in principle it may be impossible to establish causality, the best we can hope to do is establish time-ordering of events as well as ruling out spuriousness. With ever more voluminous data, I argue that even this more modest task becomes increasingly difficult, not simpler. This is because with more data, the number of correlations, which remain consistent with our theoretical understanding of the phenomena on the smaller or micro-level, tends to increase. Data measured over time offers the seduction of being able to tease out the sequence in which events occur, but as the phenomena under investigation becomes more complex (particularly in respect to the presence of feedback loops between "cause" and "effect" that characterizes most social phenomena), the techniques that we use to establish time ordering become so dependent on meeting an ever-increasing set of statistical assumptions that we can say at best that we are merely demonstrating the conditions under which these assumptions are met (or not met) rather than uncovering any actual patterns. If it seems like I am making sweeping generalizations about an entire class of statistical techniques, I am doing so only in respect to their appropriateness for studying phenomena like peer production. In other circumstances, these statistical assumptions may easily be met – for example, when contracting a disease is known to be a precursor to mortality, and not the other way around. Similarly, if working within a mature paradigm such as macroeconomics, we can adopt theoretical propositions to rule out the possibility of reverse causality, even if those possibilities might otherwise make intuitive sense. Given the "pre-paradigmatic" nature of our knowledge about peer production, regrettably these ideas offer no assurance despite their frequent appearance in the literature.

One innovation in my first empirical study (Chapter 2) is that I eschew the analysis of big data in favor of a carefully-planned field experiment that yields what I informally call “small data.” There continues to be both methodological interest in experimental research in sociology (Campbell, Russo, & Russo, 1999; Gomm, 2004; Guala, 2005) as well as substantive applications (Andreoni 1988; Frey and Meier 2004; Muchnik, Aral, and Taylor 2013; Raudenbush, Martinez, and Spybrook 2007; Salganik, Dodds, and Watts 2006; Salganik and Watts 2008; Shang and Croson 2009), although it still remains an unpopular approach to research, in part due to concerns about external validity (Lucas 2003). However, in this regard, Wikipedia demonstrates why it is a “strategic” site from a methodological standpoint, since it permits a controlled experiment in a naturalistic setting. Such “online field experiments” hold the promise of addressing one of the most challenging aspects of causality – that is to say, the time ordering of events – in social phenomena where everything can appear to be both a cause and effect of everything else.

Sampling and stratification. Given the high degree of inequality in participation, we arrive at another vexing problem when one characteristic (such as work or status accumulation) varies as a power of another, or follows a power law distribution. Because the properties of the mean and variance of such distributions are only well behaved under certain conditions, this makes application of standard statistical techniques such as regression analysis inappropriate if those conditions are not met in the population (Barabasi & Albert, 1999; Clauset, Shalizi, & Newman, 2009; Goldstein, Morris, & Yen, 2004; Newman, 2005).

Aside from the mathematical properties of power law distributions, we need to consider the substantive meaning of such a distribution. For instance, does it matter that the vast majority of contributors to Wikipedia have made only one edit to the project, never to be seen again? Anthony, Smith, and Williamson (2009) considered this question and found that in a number of instances, these one-off contributions to Wikipedia yielded a highly reliable and valuable contribution to the content of the article. They termed these contributors “Good Samaritans” who add significant value to the project despite not being regular community members.

However, many other contributors in the “long tail” of the distribution contribute nothing of any value – such as the inadvertent or overt vandalism to the encyclopedia’s contents.[4] The most common type of vandalism is often called “silly vandalism” because it is not intentionally malicious but represents curiosity about how peer production works (i.e. “Is this *really* an encyclopedia that *anyone* can edit?”). To counter such vandalism, experienced users – and even semi-automated computer programs “bots” designed to identify such vandalism – quickly restore the prior version of the article and wipe out these changes. Studies of vandalism (Kittur, Pendleton, and Kraut 2009b; Potthast, Stein, and Gerling 2008; Priedhorsky et al. 2007) show that mischievous vandals are discouraged by the immediacy of having their vandalism quickly reverted, rendering it ineffective. But seeing one’s first effort immediately reverted can be perceived in a negative light, serving to dissuade individuals from making future contributions. Experienced users are aware of the inhibitory power of such immediate and negative feedback, and the community has established a guideline (“Please do not bite the newcomers”) that suggests newcomers should be welcomed into the community in a polite manner, assuming good faith (of which curiosity is a form) on their part. As such, the community has developed ways to attempt to moderate the sting of being immediately sanctioned by turning it into a positive first interaction.

Such situations are methodological challenging for a number of reasons. If a person makes only one contribution, there is no “pattern” to statistically analyze; it can be nearly impossible to conduct a representative survey of such contributors to ascertain their motivations or their impressions of the negative

[4] Vandalism in Wikipedia is considered any change to articles that deliberately compromises the integrity of the project.

feedback they may have received; and given the frequency with which such behavior occurs, qualitative approaches to analyzing the interactional dynamics of such situations may be of limited generalizability.

One solution to this problem is to simply ignore it by cutting off the tail and only focusing on the head of the distribution, where most of the work is concentrated among a small number of contributors. The acceptability of this approach is contingent upon the question being asked: as long as we are focused on understanding what is going on among the most active contributors, there is no reason to consider whether the same social mechanisms at play are applicable to those contributors who rarely engage in any work whatsoever. In such a situation, it may even be desirable to stratify this small core of the distribution into smaller sub-groupings to attempt to discern differences among highly engaged contributors, and then sample from within these strata. This is the approach that I take in Chapter 2.

On the other hand, if we are interested in social interactions (such as counter-vandalism) that tie together contributors from different parts of the distribution, then a much larger part of the population must be accommodated in a way that is appropriate to the question at hand. In this regard, it is imperative to refer back to theory to try to determine an appropriate cut-off point for who should be included and who should be excluded from the study. However, this situation is inherently self-referential, since the theories we are working with are themselves partially the product of prior research decisions about which contributors to analyze. In this regard, the best we can do is follow a set of conventions established in previous research until sufficient gaps in our understanding have accumulated to suggest that we need to change the convention. This is the approach that I take in Chapter 3, where I am mainly interested in contributors whose "careers" span at least a month's time, although others have used weeks (Hill et al. 2012a) or even hours (Geiger and Halfaker 2013) as their metric.

Missing data, omitted variables, and specification error. As a final methodological point, I consider a situation that seems particularly relevant when analyzing voluminous digital-trace data. The potential problem of missing data is always present in social science research, and the various techniques and ways of managing (if not altogether overcoming) this problem are standard topics in the methodological litearture (Allison 2001; Enders 2010; Graham 2009; Kossinets 2006). The best we can hope is that data are missing truly at random, and we can apply ever more sophisticated corrections to the problem when we suspect that they are not, with varying degrees of success. The promise of big data is that we "capture it all" (to borrow a phrase from the biggest big data practitioners in the world, the United States' National Security Agency), rendering missing data a problem of the past. However, this is a faulty and naïve conceptualization of the problem.

The type of missing data that I am referring to here is the systematic failure to collect data on a theoretically-relevant variable of interest, simply because the system was not designed to measure or record such data. To keep this discussion from becoming too abstract, consider the correction that Shaw (2012) offers to my own research on the relationship between work and status-based social rewards (in Chapter 2): Shaw contends that not all contributors are equally predisposed to being motivated by status concerns within the community, and devises a clever way to measure this latent variable by distinguishing contributors who "signal" their status to others versus "non-signalers." This distinction is not altogether obvious in the data, yet it may represent the missing variable that helps make sense out of the variance in responsiveness to awards that we see in the community.

What other relevant information is not being captured in big datasets? This remains an important consideration, yet not one that I have seen treated with any seriousness in the literature. Digital trace data, such as the publicly available database of all actions taken by Wikipedia contributors, presents us with such an enormous opportunity to find correlations among things that are recorded that the field has not devoted sufficient attention to what is missing. Worse still is when the underlying construct is being measured, but only partially; for example, on Wikipedia, nearly 80% of contributors reveal no information about their gender, while about 20% do disclose this information, which of course is immediately captured in its large database. But contributors much choose their gender using the limited categories "male" and "female," and given these options, it is unsurprising that such a small percent choose to do so. With such a large portion of missing data, it stretches credulity that the data are missing truly at random. Nonetheless, researchers have been perfectly content to proceed with their analyses of gender on Wikipedia without giving so much as a second's thought to the potential pitfalls of this approach. In Chapter 4, I elaborate on the substance of this limitation in more detail, but in the remainder of this section, I want to explore some of the methodological limitations that can arise out of missing data and omitted variables. The general category of such problems is known as specification error.

Specification error can occur when data are missing not at random, when important variables are omitted, when the functional form is incorrectly specified, and in other circumstances revealed through analysis of variance and analysis of residuals. The first point is that that the discipline remains too wedded to tests of statistical significance, and less focused on the potential issue of statistically significant findings even in the presence of a poorly specified model. In the era of big data, even with 80% missing, we still have sufficient statistical power to find $p < .05$ in a multitude of circumstances. This is because the standard error decreases with the square root of N, so even mildly correlated variables appear statistically significant with a large enough sample size.

A second point is the common emphasis on proportion of variance explained in statistical models, or R-square. But these values can be equally misleading, since with longitudinal data, it is trivial to add a "lagged" dependent variable and see an increase in R-square if the data have even mild correlations over time and the time effects are not properly modeled. Similarly with a logistic regression, analysis of predicted classification is valuable during model building to assist in determining how much variance is able to be explained, but it is no guarantee that the model as specified can help us truly understand the social phenomenon being studied. This is especially problematic given the typically large variances present in social phenomena – and the large variances left unexplained in most analyses that get published in the top-tier journals in the discipline.[5] Even when we construct models from theory, we can accumulate a large amount of evidence in their favor without ever being troubled by the cruft of stubborn inconsistencies that can focus attention on a theory's deficiencies.

More fundamentally, analysis of variance is an important part of the process of model building, such as using Levene's test of equality of variances across groups. Homoskedasticity is an assumption behind many of the statistical techniques we rely on, such as the t-test that determine the statistical significance of regression coefficients. Given the aforementioned properties of power law distributions, it is likely that much of the statistical research on Wikipedia violates this core statistical assumption, if it does not pay close attention to distributional assumptions.

Other methods for assessing model fit may not be that much more helpful. For example, one of the earliest techniques for assessing model fit, called RESET[6], was proposed by Ramsey (1969). The idea behind the

[5] In one recent issue of the American Sociological Review, the R^2 values from analyses in different articles range from 0.16 to 0.92; and the change in R^2 across models within the same paper was about 25%. This suggests to me that the habit of focusing on values of R^2 needs to be retired.

[6] Although it is commonly referred to as the Ramsey RESET test, the acronym stands for "Regression Equation

Ramsey RESET test is that if any non-linear combinations of covariates are correlated with the dependent variable, then the model is misspecified. This approach was extended (Thursby and Schmidt 1977) to consider other sources of misspecification if powers of covariates (i.e., x^2, x^3, etc.) are correlated with the dependent variable more than at chance. But, for example, assume that on Wikipedia if we find that contributor productivity varies with the cube of the number of days since their last interaction with an administrator, how does this statistical explanation translate into a better understanding of the relationship between contributors with elevated administrative privileges in the community and those without?

Another test for misspecification analyzes the link function, which is what relates the dependent variable to the independent variables. The idea of the test was introduced by Tukey (1949) and refined by Pregibon (1980). The idea of the link test, as it is known, is to regress the dependent variable on the predicted values from the regression equation as well as on the square of the predicted values. If the term for the squared predicted values is statistically significant, it indicates that the link function chosen for the regression equation may be incorrect or that a predictor is missing. As I outlined above, omitting a theoretically-relevant predictor is a truly awful form of specification error that can dramatically change the results of a statistical analysis. While these procedures assist in assessing whether a model as specified passes the tests, there is no foolproof statistical technique to determine if the a variable has been left out of the regression model. So we are left with the situation of attempting to construct statistical models in a theoretically correct way which pass our statistical checks, even though the potential pitfall remains and we have no way of knowing whether we have crossed the chasm or fallen into it.

To demonstrate the potentially serious consequences of specification error due to omitted variables, I present a series of simplified models in Figure 1 that highlight the problem. If a theoretically-relevant predictor is omitted from the regression model, the regression results can nonetheless lead to seemingly favorable evidence for forming a conclusion that is in actually erroneous. To illustrate this point, I use a simple indicator (dummy) variable that contains information about the omitted variable. In this case, the omitted variable has two response categories, represented by the open and closed diamonds in the figure. The solid line represents the overall best fitting linear regression line, omitting information about whether a diamond is open or closed, while the broken lines are the linear trend within each group of the omitted variable.

Panel A represents a form Simpson's paradox (1951), where the slope of the best fitting regression line between x and y (solid line) is opposite in direction to the linear trend within each category of the omitted indicator variable. This is an example of how a coefficient can "flip" in sign by including an additional variable in the model. Simpson observed this surprising effect where the correlation between two variables became inverted when the population was partitioned into sub-populations, which in our case are represented by the omitted indicator variable. In this case, the two categories of the omitted variable differ both on their

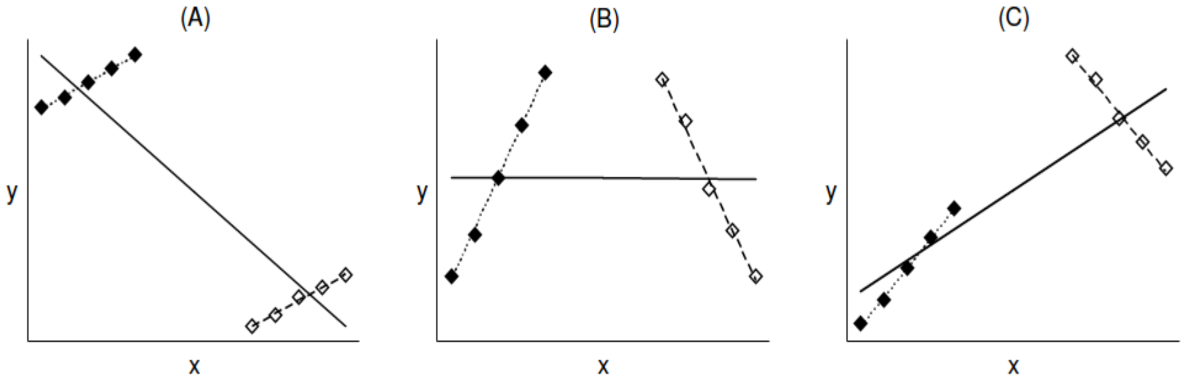


Figure 1: Three examples of specification error, which can lead to making erroneous inferences.

Despite this paradox being understood for more than half a century, when Simpson's paradox is identified in social sciences, it is commonly treated as a curiosity; as an example, consider the title of the article, "Simpson's Paradox in Real Life" (Wagner 1982). It is unclear to me where else Simpson's paradox may be observed. Psychologists have devoted some attention to this problem, likely because psychological research often entails partitioning the population into sub-groups (e.g. the "Big Five" personality types [McCrae & John, 1992]) which are then compared across other sub-groupings (e.g. males and females). Sociologists, on the other hand, have paid less attention to this potential problem, which may be the case because the discipline revels in finding a situation that appears false at first but is demonstrated to be true nonetheless, or what Quine (1966) called a *veridical* paradox. Murray Davis (1971) provided the most memorable articulation of this view when he suggested: "It has long been thought that a theorist is considered great because his theories are true, but this is false. A theorist is considered great, not because his theories are true, but because they are *interesting*" (p. 309).

One form of interesting finding is when no statistically significant relationship is found even though one is expected. Panel B shows a non-significant linear relationship (slope = 0) for the best fitting regression line between *x* and *y*. Even though there is clearly a linear trend within each category of the omitted indicator variable, because their means on the dependent variable are the same, there is no overall linear trend. In this scenario, it does not matter that there these categories have different means on the *x* variable; the slope would still be zero even if the within-group trend lines crossed to form an X shape. One suggestion for how to identify an omitted variable is through analysis of residuals, with heteroskedastic residuals serving as a clue that Simpson's paradox may be lurking. This is not completely satisfactory, though, since the results of a Breusch-Pagan test (1979) for homoskedasticity of residuals under such a scenario could still lead us to believe the residuals were homoskedastic even when omitting this key variable.

However, the fact that Panel B shows a difference in mean on the independent variable across these two categories yields an even more subtle problem. A standard solution is to include a quadratic term x^2 that would predict an inverted-U shaped relationship between *x* and *y.* (You may even have a hard time not seeing such a curve to fit these data, now that I've mentioned it.) But this is of no help and may only confuse the situation even more by incorrectly leading us to "find" a statistically significant non-linear relationship that does not really exist! This is an example of a *falsidical* paradox in Quine's classification, one that appears false at first and yet for other reasons is demonstrated to be false nonetheless. In the case of Panel B, the ostensible non-linearity between *x* and *y* is caused by an opposite linear trend between categories of the omitted variable, as well as a difference in their means on the independent variable as well.

Panel C shows a combination of both previous errors: the slope the best fitting regression line is less steep than the linear relationship within one category (closed diamonds) but of the opposite sign from the linear relationship within the other category (open diamonds). The slope of the best fitting line is a function of the difference in means between categories on both the independent and dependent variables. This example demonstrates how omitted variables can lead to significantly biased estimates, either in magnitude or direction (or both).

By this extended methodological discussion, it was my intention to place existing research findings as well as my own empirical studies on less solid footing than what we might otherwise be led to believe. Although this might at first blush seem a counter-productive thing to do before launching into a series of empirical studies involving statistical analyses, I hope that such reflexivity is welcome, as I believe that it should be part of all research. On that note, it is also appropriate for me to discuss ethical considerations regarding the conduct of the research in this dissertation.

Research ethics

This study's research protocol was approved by the Committees on Research Involving Human Subjects (IRB) at the State University of New York at Stony Brook (CORIHS #2011-1394). As the experiment presented only minimal risks to subjects and was designed and conducted so as to be non-disruptive to the community, following the IRB protocol, we maintain the confidentiality of individual records and identities of all research subjects. The research proposal was discussed on an online mailing list of Wikipedia researchers, who did not raise any serious objections to the study. However, I did incorporate one of their suggestions, which was to limit rewards to only the most qualified (i.e. highly active) users, in correspondence to the way Barnstars are used in the community.

All secondary data that was collected as part of this research was publicly available and did not contain confidential information that could identify or be used to disclose the identity of any individuals involved. These data come from Wikipedia's public databases. Any data that were collected were done so anonymously and the records in the dataset identify individuals using only arbitrary IDs. Although I do not believe any harm can come form these data, on the other hand, it is trivial for a malicious third party to exploit even seemingly anonymous records along with other metadata to ascertain that someone – for example – regularly participates in a discussion on Wikipedia related to the article on "Chinese democracy movement" from a computer registered to the campus of UC Berkeley between the hours of 1:00 and 3:00 am on Saturdays. Perhaps that does not matter, but it is not my place to decide that for other people, even if they are in principle "in public." I believe this shows why traditional research ethics guidelines need serious reappraisal; given the sheer volume of data that is available for new analytic techniques, the ability to truly deidentify data and protect the confidentiality of research subjects is becoming more challenging than ever.

Plan of dissertation

Chapter 1: In this opening chapter, I establish the overall structure of the dissertation. I introduce the social phenomena being studied, provide an intellectual justification for my original research plan, and situate all the empirical work within the sociological conversations which my scholarship is party to. The heart of the dissertation consists of three empirical chapters (Chapters 2-4). Each chapter addresses one of the three research questions, which I describe below. Although these questions are related in many ways, they each also stand alone as important contributions to the literatures that they are addressing. As such, it is my plan to situate each empirical chapter with its own relevant literature, describe the data and methods used to answer each research questions, and present and discuss each chapter's research findings. I conclude this chapter with a discussion of methodological considerations for studying inequality, and then present the plan of the dissertation as well as a chapter outline.

Chapter 2: Nominally, what is being made by Wikipedia contributors is a public good: an encyclopedia, with millions of articles in hundreds of language, that is free for people around the world to use. As an organization, Wikipedia relies solely on volunteer labor from hundreds of thousands of contributors around the world, who work for free to produce this content; it lacks traditional incentives such as monetary compensation, employment, or even strong ties among contributors. Absent these formal incentives, what permits Wikipedia and other organizations like it succeed in garnering time-consuming work and efforts of its many contributors over a long period of time?

A rich body of literature in the social sciences has demonstrated that purely social reasons can underpin contributions to public goods provision, without formal or material incentives being present. People are often initially drawn to such volunteer work because of altruistic motives or their desire to further the goals of the organization. But for the organization to be successful, it must avoid the situation where volunteer efforts taper off or stop. What form of social incentives, and under what circumstances, help to counter contributor decline?

Recent social science theories describe how concern for one's social standing in the community can act as a sufficiently strong incentive to motivate people to continue volunteering. As individuals begin to forge their identities more closely with the community, their initial reasons for contributing are channeled into considerations of their reputation and status within the community. One primary way this occurs is through peer-to-peer rewarding of informal tokens of appreciation and recognition of work performed. Such rewards signal that recipients have a sign of distinction and social standing in the community. In return, recipients give back to the community by performing additional volunteer work, fostering a virtuous cycle of work and rewards.

Although the contours of this theory are broadly supported by empirical research, the literature is not completely satisfying for two reasons: First, it is often difficult to operationalize and measure "fuzzy" social concepts such as informal rewards, despite contributors reporting in survey and qualitative research that they view receiving such social rewards in a positive light and that it motivates them to contribute more. Since Wikipedia is a computer-mediated organization that records all interactions between contributors in a publicly-viewable database, there is a concrete record of when contributors give and receive social rewards from one another, as opposed to in offline volunteer organizations where social rewards may be more ephemeral and not leave a permanent record. As a result, these data can be correlated with records of contributors' activity history, permitting rigorous statistical analysis of the relationship between work and reward.

However, this does entail a second methodological difficulty: in analysis of observational data (even data that is measured longitudinally), it is difficult to establish causal ordering of events. Do social incentives such as rewards actually increase subsequent effort, or do they simply serve as rewards for existing effort? For example, we observe a strong association between highly-active contributors being highly rewarded. But since work and reward histories co-evolve, it is methodologically difficult to disentangle them. Traditional statistical approaches are problematic because the assumption of statistical independence (between the receiver and giver of rewards, and for receiving multiple rewards over time) is not supported by the data. Instead, the practice of social rewarding in Wikipedia creates a dynamic network that ties together individuals in difficult to analyze ways. For example, the reward network contains cycles (A → B; B → C; C → A) and reciprocal exchanges (A → B; B → A). Existing solutions, such as using ERGM or "p*" models, are themselves sensitive to other sets of statistical assumptions. Quite frankly, given the lack of consensus regarding the appropriateness of different methodological approaches to this question and the variety of theories that could be used to justify different outcomes, we would find it unsurprising to find that rewards increase effort, decrease effort, or have no effect on effort whatsoever.

To overcome some these limitations, I conduct a series of randomized, controlled field experiments on Wikipedia to test the relationship between work and reward. The evidence from these experiments help us to disentangle the causal ordering, establish the effect size of social rewards on contributor behavior, and investigate many nuances in terms of when and why rewards are effective or ineffective. The heart of the experiment consisted of giving an informal reward in recognition of contributor effort to a randomly selected half of a sample of highly productive Wikipedia contributors. On Wikipedia, the most common type of recognition of contributor effort takes the form of a "Barnstar" or editing award. By allocating Barnstars orthogonal to merit, and having an equivalent control group for comparison, I could establish how rewarding affects work.

Conducting this randomized, controlled experiment yielded a second opportunity. By creating a small status differential between contributors where none existed before, any subsequent accumulation of status-based rewards could be attributed to the contributor's initial boost in status compared to his or her equally-deserving peers. This situation is highly beneficial to any attempts at understanding why rewards tend to be concentrated in a small group of peers. It is widely understood that biases in social evaluative processes can lead to a "rich get richer" phenomena when status grantors are more likely to grant additional status to those who already have higher standing in the community. On the other hand, empirically demonstrating the presence (and magnitude) of this form of decoupling of status from merit is difficult to establish because of their aforementioned entanglement in observational data.

Taken together, the results of these experiments yield considerable insights into the social dynamics that are both constitutive of and reflective of the high degree of participation and status inequality seen in Wikipedia. I find that the reward system in Wikipedia is broadly meritocratic, yet at the same time they can contribute to the high degree status inequality seen at the top of the distribution. Social rewards significantly boost efforts among the top contributors while being ineffective in this regard for less active contributors, suggesting that they are not the underlying cause of the participation inequality that we see, but they may make it more difficult for more peripheral contributors to close this gap since their efforts may remain under-recognized relative to their more prolific peers.

Chapter 3: In the previous chapter, I demonstrated how social rewards are one mechanism whereby participation in peer production is sustained. In this chapter, I consider a related puzzle: why do people stop contributing? The simple answer is that they fail to receive sufficient social recognition for their efforts, but a variety of alternative mechanisms have been proposed for contributor turnover.

One of the most prominent "decline" theories is that organizations such as Wikipedia become more bureaucratized as they grow larger out of necessity of coordinating the vast influx of new contributors. As community norms, procedures, and policies become more formalized, it is thought that newcomers become turned off by encounters with "small minded bureaucrats" who are more intent on promulgating rules than creating content. It is not hard to find anecdotal evidence (from *former* contributors) that supports this theory. However, in an organization with tens of thousands of contributors, a few stories of hostile bureaucratic encounters does not rise to the level of sufficient evidence that bureaucracy is the cause of decline.

Wikipedia underwent exponential growth during the first several years of its existence, and there were really no rules except to "ignore all rules" since there were so few contributors and so much work to be done. Five years later, there were many more contributors who had spent countless hours discussing and creating the many informal and formal structures and procedures to make their work more orderly and predictable. As time went on, new users encountered and were compelled to engage with these rules and policies; qualitative research documents that many new users did not enjoy having these rules foisted on them by their peers – after all, there was still the motto to "ignore all rules."

I endeavored to test the bureaucratization theory of decline using data from a representative sample of contributors to Wikipedia over its history. This permitted me to assess whether working in an increasing formalized setting leads to less productivity for existing users as well as lower retention of new contributors. I find that while contributor retention has declined over time, new users who choose to engage with the increasingly bureaucratized Wikipedia community fare better than those who do not. This supports the notion that Wikipedia is asking more from new contributors, and retention is lower for those who choose not to, or are discouraged from, making this higher initial effort. However, I also do not find widespread support for the belief that established contributors focus more on bureaucratic work and less on productive effort. To the contrary, more formalized work environments yield greater productivity from these contributors, suggesting that the same organizational mechanisms that aid their work also promulgate bureaucratic structures on new contributors. These differential effects can create and exacerbate inequality among participants, but not in the precise way that critics warn. As the organization become more ossified, the retention rate of new contributors has decreased, yet we do not see an accompanying rise of a managerial class of contributors.

Chapter 4: This third empirical chapter addresses a consequence of another type of inequality in participation, that between men and women's participation in Wikipedia. Recent survey research of contributors to Wikipedia has found that less than 15 percent of its contributors are women. One cause for this high degree of gender inequality is self-selection, as we continue to imagine expertise in creating knowledge as a masculine pursuit, and women may buy into this belief as well, undermining their confidence in participating in online knowledge production. However, this gender gap has been shown to be more than simply self-selection, as other research suggests that aspects of online participatory culture more generally limit women's participation due to excessive conflict and contentiousness, devaluation of certain topics or perspectives, and in some instances, overt hostility or other forms of misogyny.

This raises several interesting questions about whether online realms are open to a diverse range of participants and whether they can ever truly represent "disembodied spaces" if participants' socially-learned and embodied gender, and others' perceptions thereof, accompany them into virtual spaces. One way that gender may continue to be salient is through presentation – in particular, whether contributors choose to publicly disclose their gender to others in an online community, which can affect how group members perceive one another and alter group interactions. A second way that gender may be salient is through socially-learned gender performances in terms of the roles people adopt and the types of work they perform.

In this chapter, I analyze whether gender diversity affects the quality of work produced by contributors to Wikipedia. I use the quality of the information in Wikipedia's encyclopedia articles as an indicator of the performance of the groups who author those articles. Organizational research has long known that diversity – in respect to group composition, the types of tasks being performed, and other group dynamics – affects a group's performance and shapes its collective output. However, the consequences of gender diversity on Wikipedia's articles not been explored in any detail.

Studying the work of Wikipedia contributors provides an invaluable opportunity to assess how gender diversity, and the different ways of doing gender, affects group performance and shapes the quality of knowledge being produced. In this chapter, I also highlight one of the persistent challenges for any analysis of online organizations. It is not entirely clear who contributors are, particularly on salient social and demographic characteristics such as gender, since contributors may participate online completely anonymously. As a result, existing analyses of the "gender gap" in participation have relied on publicly disclosed information from contributors, which is problematic since only about 20% of contributors choose to do so. Nonetheless, the results from my analysis suggest that gender diversity leads to favorable outcomes for article quality. My finding that diversity can be beneficial to groups is of practical relevance to the people who are building and working within these types of online organizations. Particularly, Wikipedia can take steps to shrink its gender gap by identifying and addressing barriers to inclusive participation from a diverse range of contributors.

Chapter 5: I conclude my dissertation by reiterating the findings from my series of studies and connecting them back to the overarching theme of participation inequality. I also reflect on the methodological limitations inherent in the choices that I made in my empirical studies, and consider what consequence those choices have on what we know (and continue to not know) about how Wikipedia, and peer production more generally, works. I also recognize that in principle there can never be a definitive account of such large, heterogeneous organizations, and consider new methodological challenges to answering substantive questions that arose while conducting the original research – an absurdly endless, recursive process.[7]

[7] While writing a dissertation on a topic related to computer science, it is impossible to avoid the concept of "recursion," which entails a form of self-reference. Recursion is often the form of humor such as in the example of the GNU Linux operating system, where GNU is an acronym for "GNU's Not Unix," where the G stands for GNU, creating an infinitely recursive semantic loop. Conducting research can similarly entail getting caught in such a recursive loop of answering new questions and testing theoretical propositions that emerge out of research findings.[7]

Chapter 2. Status, work, and rewards

The problem with Wikipedia is that it only works in practice. In theory it could never work.[8]

Introduction

One of the questions most frequently asked about Wikipedia is why people work for free. As an organization Wikipedia relies solely on volunteer labor from hundreds of thousands of contributors around the world who work for free to produce its content.[9] In turn, what is being produced by Wikipedia's contributors is a public good: an encyclopedia with millions of articles in hundreds of languages that is free for people around the world to use. Organizations that provide public goods through voluntary contributions of individuals face a constant challenge: how to retain existing members and recruit new ones. In the absence of material incentives such as monetary compensation, with no formal opportunities to build one's credentials (i.e. no one lists "Wikipedia author" on their resume), or any strong social ties that compel participation, what is it that permits Wikipedia and other peer production organizations like it to succeed in eliciting time-intensive work efforts from its contributors over a sustained period of time?

Many valuable public goods, such as open source software, question and answer forums, distributed proofreaders, and collaborative maps, are produced through voluntary contributions of individuals who work together on a peer-to-peer basis (Benkler 2002; Demil and Lecocq 2006; Markus 2007). Wikipedia happens to be one of the most prominent examples, but all pose a major puzzle to social science theories from organizational theory (Powell et al. 2005) to the literature on collective action (Olson 1965), as none of the classic mechanisms that prevent free-riding – coercion, enforceable contract, or network-embedded trust – are seen to be operative (Demil and Lecocq 2006). Posing this question is not to problematize volunteering more broadly. There is a growing body of literature about what motivates people to begin contributing to Wikipedia and similar organizations. Broadly, individuals are guided a number of considerations: internal or social-psychological motivations such as intrinsic enjoyment and "fun" (Hars and Ou 2002; Nov 2007), value-orientation (Kuznetsov 2006; Schroer and Hertel 2009) such as ideological commitment to contributing to free culture, and personal or social benefits such as the development of human capital (Mehra and Mookerjee 2012). This accords with the rich tradition in social theory which emphasizes the importance of social factors that underpin the production of public goods (e.g., Hardin, 1968; Heckathorn, 1993; Kollock, 1998; Ostrom, 1990).

[8] The earliest known variant of this saying is was in 2006 by user:Gareth_Owen who quipped, "The problem with Wikipedia is that it only works in practice. In theory, it's a total disaster" (https://en.wikipedia.org/w/index.php?title=User:Gareth_Owen&diff=35978744).

[9] The Wikimedia Foundation employs a small number of people to perform vital work managing the technical infrastructure which runs the sixth most popular website in the world. The Wikimedia Foundation also performs community outreach, educational coordination, and, increasingly, fundraising to support its mission. Nonetheless, their total employees have risen to "only" 142 employees worldwide in 2012, according to their 501(c)(3) filings.

Nonetheless, volunteer organizations face the challenge of retaining volunteers. In Olson's classic formulation of the problem (1965), individuals exhibit a tendency to "free ride," particularly ride as group size increases. This seems particularly relevant in a situation such as Wikipedia. In the absence of formal or material incentives, the selective incentives that motivate participation to public goods may be purely social (Willer 2009). Willer posited that concern for one's social standing in a volunteer community can act as a sufficiently strong social incentive to motivate people to continue their volunteer efforts. The idea that a contributor who is particularly generous in giving time, energy, and effort to the project (and being nice to others in the process) gets noticed and appreciated by the community. When the community recognize this contributor's efforts through an informal badge of honor, which indicates and signals this person's increased social standing in the community, that person in turn reorients his or her efforts toward further accomplishing the group's goals. Willer posits that such conditions foster a "virtuous cycle" where work leads to recognition which boosts one's status and in turn leads to more work.

In looking for support for this answer to the puzzle, the example of open source software development communities seems a relevant comparison. Stewart (2005) described how the drive for status attainment among software developers created relatively stable social orders. As more time elapses if actors do not establish a high status, the harder it is for them to do so since community members tend to evaluate a contributor's reputation according to existing social references. Status, in this community, meant prestige, respect, and honor. This form of status determined a person's status in the usual sense of an individual's position in the social system. Research on other open-source projects has found that highly-active contributors are often motivated to develop their reputation and forge their identities within the community (Ghosh & Prakash, 2000; Lakhani & von Hippel, 2003; Lerner & Tirole, 2002; Mockus, Fielding, & Herbsleb, 2002). This also appears to be present on Wikipedia, where a recent survey of contributors found that 18% of editors reported that they continue to edit Wikipedia to gain reputation within the community (Wikimedia Foundation (WMF) 2011). Recipients of such recognition repay their enhanced status within the community by making further contributions, fostering a virtuous cycle of effort and recognition.

The contours of this theory are broadly supported, but at the same time, there were some limitations. Prior work focused primarily on small group settings, either naturally occuring or in laboratory settings, but whether these were applicable in other contexts had not yet been shown. For example, Willer's evidence came from laboratory experiments but had not been field tested. On the other hand, one of the significant strengths of Stewart's work was that showed how connections and interactions at the actor-to-actor level led to the emergence of macro-level social order. This type of analysis was possible because these data were all recorded in the database, but this situation is atypical. It can be difficult to measure honor and status, particularly in a large heterogeneous community. But given that all social interactions are recorded on Wikipedia, this represents in principle a record that could similarly be analyzed. This comes with the tradeoff in that no measure of these concepts will be completely exhaustive of all the possible ways in which contributors can recognize one another's efforts and determine each other's standing in the community. Indeed, on Wikipedia, there is no centralized accounting of a person's status nor any official register of all the ways in which contributors can show appreciation for one another's efforts. However, the most common practice for contributors to recognize one another's efforts is through giving informal editing awards to one another. The most familiar of such awards in the community is known as a Barnstar (Kriplean, Beschastnikh, and McDonald 2008). While Barnstars are not the only indicator of status and recognition, they are the most frequently used method of rewarding hard work and due diligence. Data on Barnstars can be correlated with records of each contributors' activity history, permitting a rigorous statistical analysis of the relationship between work and reward.

Although having a detailed record of work and reward solves one challenge, it reveals yet another one. Even in longitudinal records, the problem of assessing the magnitude and causal ordering of events remains a challenge. Do social incentives such as Barnstars actually increase subsequent volunteer efforts, or do they simply act as rewards for past efforts? Willer's theory suggests these two factors mutually reinforce one another, and indeed there is a strong association between contributor activity and receiving such awards. But since work and reward histories co-evolve, they are difficult to disentangle. Traditional statistical approaches familiar to social scientists are not entirely appropriate because the assumption of statistical independence is not met by the data. The practice of giving and receiving Barnstars creates a dynamic network that ties together many contributors in a complex network. For example, the existing observational data contains instances of cycles (A → B; B → C; C → A) and many examples of reciprocal exchanges (A → B; B → A). See Figure 3 for an example of the type of network structure of just one month of Barnstar activity on Wikipedia.

As individuals continue to contribute to group goals, the group will continue to reward them with additional social recognition and higher status in the community. Hence, highly prolific contributors will begin to accumulate a large number of rewards. Combining the insights from Stewart and Willer would seem to suggests that contributors who are left out of this virtuous cycle will have a harder time getting their contributions recognized by the community, and hence will be locked into a cycle of low status and effort, while on the other hand, higher status actors in the community will tend to garner a disproportionate share of recognition. This remains consistent with the distribution of Barnstars in the community (see Figure 2) and is consistent with a social mechanism known to generate a similar distribution of social recognition: status-grantors often choose to reward higher-status actors than lower-status actors, all other things being equal, in a process of cumulative advantage. Higher-status actors therefore might continue to accumulate Barnstars, even though each one that they receive yields no increase in how much work they perform.

Effectively disentangling the directionality of the causal processes in co-evolving records of contributing and rewarding is a daunting task. Existing approaches to handling non-independence of data, such as exponential random graph models, are sensitive to their own sets of statistical assumptions, and for the reasons I described in Chapter 1, I decided not to apply them here. However, recent research has demonstrated that one way to overcome this problem of confounding is through a randomized experimental design (Muchnik et al. 2013; Salganik et al. 2006; Salganik and Watts 2008).

I use Wikipedia as a laboratory in a field setting. It contains extensive and detailed records of contributor work patterns as well as social recognition which could be reliably measured. It is also possible to employ a randomized controlled intervention, which combines the benefits of experimental control without the artificiality of traditional laboratory settings. This overcomes the limited external validity of traditional experiments because the treatment occurs under naturalistic settings. Thus, online experiments offer tremendous potential to complement existing research approaches to investigating the social world. The evidence from these experiments permits me to disentangle the causal ordering of work and reward, and to establish the magnitude of the effect size of social rewards on subsequent contributor behavior. The heart of the experiment consisted of giving a Barnstar in recognition of contributor effort to a randomly selected half of a sample of highly productive Wikipedia contributors. By allocating Barnstars orthogonal to merit, under controlled conditions yet in a naturalistic environment, with an equivalent control group for comparison, I could effectively provide novel evidence on how this social mechanism operates and what consequence it may have for the inequalities that we see.

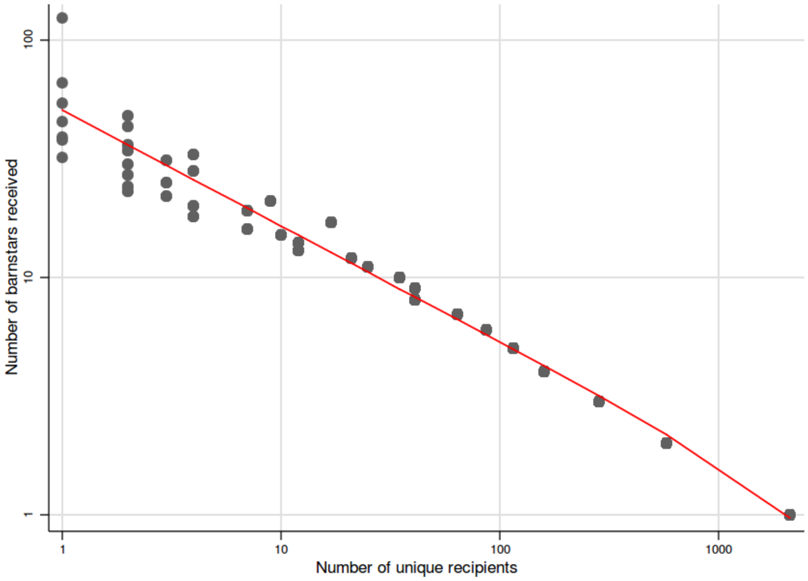


Figure 2: Distribution of total barnstars on Wikipedia through 2011, distributed among 3,700 unique recipients (N=10,649). The contributor with the most barnstars received a total of 123 barnstars, compared to 2,127 contributors who received only one barnstar. The distribution is plotted on a log-log scale, causing the polynomial fit line (red) to appear linear. This distribution shows a concentration of barnstars received by a small number of contributors: the ten most awarded contributors received 5% of all barnstars.

Preliminary work on this topic was was published in *PLoS ONE* (Restivo and van de Rijt 2012), in which we bestowed rewards upon a random half of a sample of highly-active Wikipedia editors. The results of the experiment were that subsequent productivity of rewarded editors was significantly higher than productivity in the control group, and recipients were also significantly more likely to receive another editing reward from third parties after having received the initial reward. This provides evidence that pure social rewards can be a mechanism to overcome the free rider problem, but it also raised a number of questions. If higher status actors are more likely to be rewarded, are they accumulating status at the expense of other deserving contributors? If this is the case, then such rewards also may inadvertently drive inequality as worthy editors are being deprived of their due share of recognition. However, it is unclear if Barnstars have a similar effect on other users. One reason this might be the case is because Barnstars don't necessarily mean the same thing to all contributors, as Shaw (2012) demonstrated. In other words, there may be a threshold below which contributors are non-responsive to community recognition which confers status, if those contributors are insufficiently involved with the community.

On the other hand, if less involved contributors – who still do perform a significant amount of work, just relatively less than their highly-rewarded counterparts – are similarly responsive to Barnstars, then targeted recognition of these contributors' efforts may be a key mechanism to increase participation and close the inequality gap. Research suggests that positive interactions among contributors can be vital to their continued involvement in the project (Panciera, Halfaker, and Terveen 2009). It would also mean that some of the inequality we see is due to the disproportionate accumulation of rewards by those with higher status in the

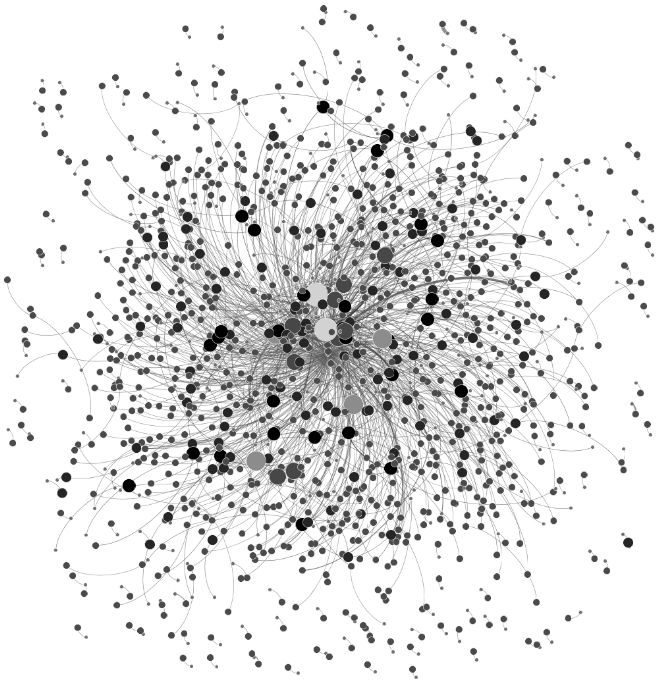


Figure 3: A directed graph of Wikipedia's Barnstar network, December 2010. Nodes are sized by in-degree (number of Barnstars received); the smallest nodes are individuals who gave a Barnstar to another contributor but did not receive any in return. Nodes are colored by out-degree, where lighter colors indicate a higher number of Barnstars given to other contributors. Notice at the center of the graph is a small cluster of contributors with a high in-degree (size) and out-degree (light color). Although this is a directed graph, the "arrowheads" are not visible at this resolution. The graph was created in Gephi using a force-directed layout. A closer analysis of sub-components of the graph indicate both reciprocal exchanges of Barnstars and network cycles.

community, at the detriment to others. Failure to be rewarded may lead to contributors decline in productivity or cessation of their activities altogether.

The way that I measure productivity is the rate of work over time, so a large *change* in productivity among less active users, relative to the most active users, requires less actual *expenditure* of effort in order to achieve the same percentage change in productivity that we saw in the most active users. A note of caution to the reader: the term "less active" contributors refers to their volume of work performed relative to the most prolific contributors. Thus, less active contributors are still quite active overall.

Methods

After receiving IRB approval to perform this research, I prepared and implemented the experimental design in the first part of 2011, and completed the research before the end of the year. During the formative stages of this research, I tried a few trial runs of the treatment using my personal account on Wikipedia (User:Mike_Restivo). Based on these experiences, and in consultation with other researchers and community members from the Wikimedia Foundation, I finalized the research plan described below.

Subject pool

The research procedure entails a randomized, controlled experimental treatment of three "tiers" within the 10% of contributors to Wikipedia (Voss 2005). Figure 4 shows the distribution of work to the English Wikipedia at the time this experiment was conducted, with a dashed line demarcating the top 10% (or "core") contributors. The experiment focused on the top 10% of Wikipedia's contributor community in order to maintain correspondence with Wikipedia's custom to reward barnstars for "hard work and due diligence." Barnstars can be given by any contributor to any other contributor by posting it on the fellow contributor's user-page for public display. Barnstars are given for numerous reasons (Kriplean et al. 2008), and are commonly used to recognize significant or tireless editing work such as general editing behavior, substantial textual additions to an article, or minor work such as copy editing.

I obtained a list of all active users (who performed at least one action in the past 30 days), excluding users with elevated or administrative privileges. I excluded from the sampling frame the bottom 90% of contributors by edit volume in the prior 30 days, then stratified the remaining top 10% of the population into three productivity tiers: top 1%, next 4%, and next 5% – which corresponds to the 100th percentile, 96th-99th percentiles, and 91st-95th percentiles in the population of core contributors. The rationale for this stratification is that although the top 10% of contributors on Wikipedia are very active, due to the way effort is distributed overall, there tends to be increasing concentration as we move toward the top of the distribution. By breaking this core group into three tiers, I am able to probe for potential differences among this already relatively small and stable community.

After stratifying the population, I conducted a random sample using equal probability of selection from each tier. The sampling procedure used a random number generator to rank users within each tier by lottery, after which I selected the first 200 from each tier who had never been awarded a Barnstar in the past. Through random assignment, 100 from each tier were entered into the control condition, while the remaining 100 were entered into the treatment condition.

During the month prior to the experiment, subjects in the top 1% performed on average 282.2 edits to Wikipedia's article pages, whereas the mean for contributors in the 96-99th percentiles was 62.5 edits, 21.6 edits for contributors in the 91-95th percentiles. (There were no statistically significant differences between treatment and control groups, which is consistent with the randomization procedure.)

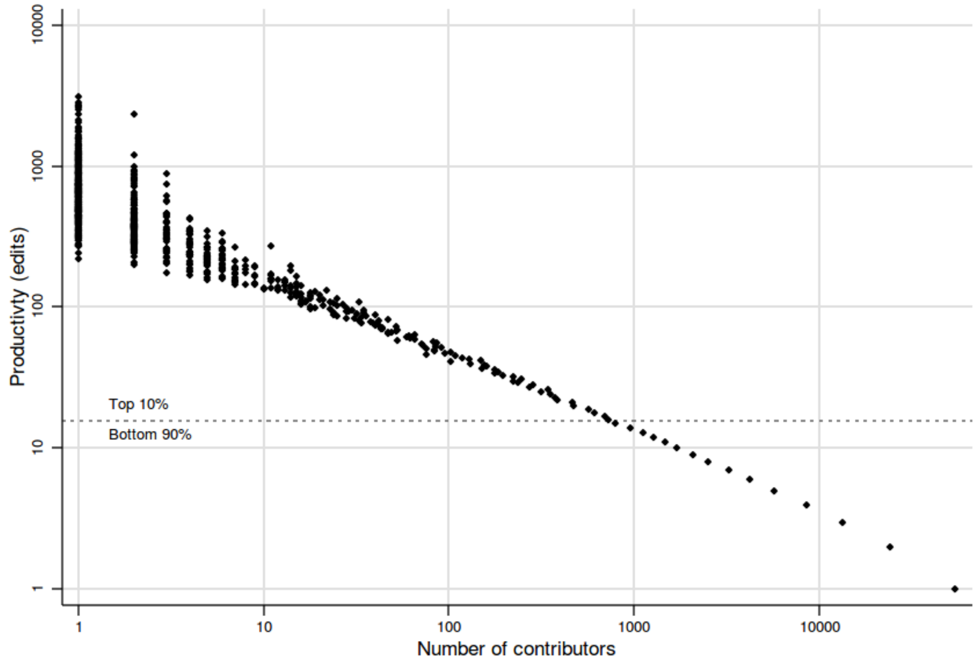


Figure 4: Distribution of contributor productivity to the English Wikipedia in the 30 days prior to the start of the experiment in 2011. The total population (N = 138,794) included all individual contributors with registered accounts, but excluded anonymous (IP) contributors and contributors with elevated administrative privileges. The graph shows the distribution of productivity on a log-log scale, with productivity (number of edits in the prior 30 days) on the y-axis. The graph plots the number of contributors at each level of productivity. The horizontal dashed line marks the top 10%, corresponding to approximately 17 edits in the prior 30 days. The graph display the characteristic long tail of peer production systems more generally, with a vast majority of editors whose contributions fall below this threshold.

Design of the treatment

The experimental treatment consisted of me assigning an informal, peer-to-peer editing award, or Barnstar, to contributors who had never previously received one from another member of the community. The treatment was given anonymously on the subjects' user_talk page, which is used for peer-to-peer communication. The Barnstar consisted of an image of a star combined with a generic positive text that expressed community appreciation for contributions, but it was not tailored to any recipient-specific activities or achievements. When users log in to their account on Wikipedia, a prominent message indicates that they have received a new message from another user. Through random assignment we allocated barnstars to half of the users in each tier (treatment condition) and withheld the reward from the remaining users in that tier (control condition).

Data collected

During the three months following the treatment, I tracked contributors' running number of edits as well as subsequent social recognition (count of additional barnstars received from third parties). I subsequently added to this information the last date of activity for all contributors in the sample, up through December 2013. I derive the dependent variable for productivity from each contributor's edit count. Existing research on peer production has used similar measures, including edits per day on Wikipedia (Kittur et al. 2009b; Panciera et al.

2009; Suh et al. 2009; Wilkinson and Huberman 2007; Zhu et al. 2013), number of code changes to an open source project (Apache) (Mockus, Fielding, and Herbsleb 2000), and number of work items attempted on Mechanical Turk (Mason and Watts 2010). As I am interested in assessing how rewards change contributor behavior over time, this measure was calculated as standardized within each user as a user's running total edits divided by the user's total pre-treatment edits.

To analyze the experimental results, I used a criterion of central tendency (median) and a non-parametric test (Mann-Whitney U) to test the hypotheses related to productivity. Both measures are robust to outliers and distributional skew (Mann and Whitney 1947). As a robustness check, I also construct regression models to investigate potential mechanisms that may explain the observed differences in treatment effect between tiers, and use an event history approach to assess whether rewards may affect contributor retention instead of directly affecting productivity.

Analysis of productivity

The experiment was designed to assess the overall effectiveness of rewards to elicit further contributions to Wikipedia. In addition, theory suggests that there may be a differential in effect of rewards across contributors of different productivity levels, where less active contributors could be more responsive to the treatment due to having greater room for their productive efforts to grow. However, the results point to a different effect across tiers.

Contrary to theory, rewarding less productive editors did not stimulate higher subsequent productivity. From Figure 5, a treatment effect can be observed as the shift in the median of the distribution of contributor productivity: among the most prodigious contributors (100th percentile, right panel), receiving a status-based reward led to post-treatment productivity that was 60% higher in the treatment condition compared to the control group (Mann-Whitney U test: $z = 3.222$, $p = .001$). By contrast, in both tiers of less active contributors (left and center panel), there was no treatment effect compared to their respective control groups (subjects in 96th-99th percentile range: $z = 0.934$, $p = .350$; subjects in 91st-95th percentile range: $z = -1.111$, $p = .267$).

I tested whether the lack of treatment effect was due to contributors possibly being unaware that they had received a reward. This could happen if they never logged in to Wikipedia after the treatment, in which case they would not have seen the prominently displayed message signaling their editing barnstar award. To explore this possibility, I excluded all subjects who made zero edits post-treatment. This rate of "drop out" varies across tiers: in the 91st-95th percentiles, 67 out of 200 subjects made zero post-treatment edits, while 37 out of 200 subjects made zero post-treatment edits in the 95-99th percentile, and only 19 out of 200 of the top-tier subjects made zero edits after treatment. This is consistent with our intuition that less active contributors are less likely to see their reward because their editing careers are marked by more frequent spells of temporary or permanent discontinuation. After excluding users who made zero post-treatment edits, I again tested productivity across conditions. The results remained consistent: the treatment effect is large and statistically significant only among the top 1% of contributors. The Mann-Whitney U test results were substantively unchanged (100th percentile: $z = 2.562$, $p = .010$; 96th-99th percentiles: $z = 0.783$, $p = .436$; 91st-95th percentiles: $z = -1.184$, $p = .236$).

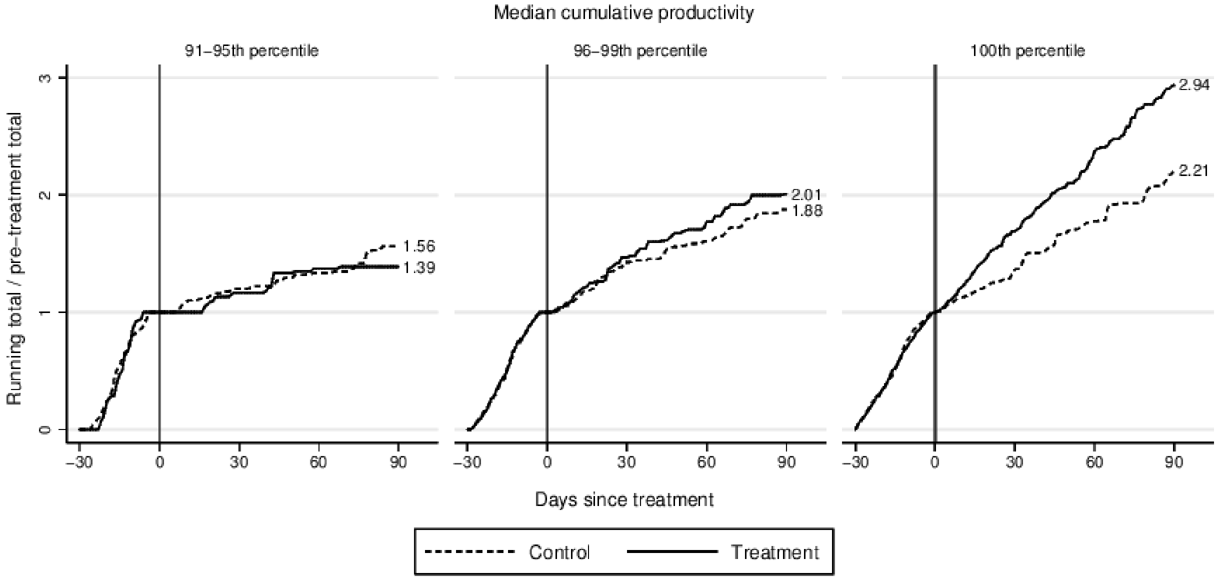


Figure 5: Median cumulative productivity, measured as the running total number of edits divided by the total edits made during the 30 days prior to treatment for each subject. The treatment was given on day 0 (vertical line), when each user had a baseline productivity of 1. The treatment groups (solid lines) are compared to their respective control group (dashed lines) in each productivity tier. By the end of the 90-day observation period, subjects in the top 1% exhibited a 60% higher median post-treatment productivity, whereas less productive subjects did not vary significantly by condition.

Because there was no treatment effect on productivity among less-active contributors, contrary to expectation, I considered whether this finding was a consequence of insufficient statistical power stemming from the greater variance in productivity among less-active contributors. To test whether this was the case, I pooled the data in order to perform a regression analysis which models the tier-specific treatment effect.[10] I also included several relevant control variables that may explain the differential treatment effect. If the regression analysis continues to show a treatment effect for the top tier that is significantly different than that for lower tiers, this would confirm that we have sufficient statistical power to discern a true difference in effect across tiers.

For this analysis, I use contributors' post-treatment edit count as the dependent variable. Because the distribution of edit count has high dispersion (variance far greater than the mean), I use a negative binomial model, reported in Table 1. For clarity of interpretation, I provide exponentiated coefficients which can be understood as an increase or decrease in the rate of edit count post-treatment. To aid comparisons in the table, I also list a rate of one for the omitted reference category (91st-95th percentile, control group). The models include a dummy variable for each tier, which indicates the mean edit count for unrewarded contributors in the 96th-99th and 100th percentile tiers compared to the 91st-95th percentile tier, which serves as the reference

[10] Since the samples were drawn from the same population, I appropriately weighted the data to correct for inter-tier differences in probability of case selection.

category. The treatment effect is represented by the product terms. All tests of statistical significance are two-tailed hypotheses.

Model 0 presents the null hypothesis of a tier-invariant treatment effect. Model 1 contains a product term for the interaction between treatment and tier, which represents the effect of the treatment on each tier relative to that tier's control group. Finally, in Model 2, I use a zero-inflated negative binomial model in an attempt to account for contributors who had a post-treatment edit count of zero. Because some individuals have dropped out of contributing to Wikipedia (either temporarily during the post-treatment observation period, or permanently), there is a group whose edit count is always constrained to zero. To account for these users in order to not bias the estimate of the treatment effect, I follow the procedures of a recent study (Zhu et al. 2013) which uses a zero-inflated negative binomial model to estimate edit counts of Wikipedia contributors. This model estimates the likelihood of a contributor making zero edits separately from the estimated positive edit count, providing a more conservative estimate. As predictors of an editor having a zero edit count, I use the number of days that have elapsed since a contributor's last activity before the treatment date (i.e. immediacy of feedback) and the amount of work performed on the day prior to receiving the treatment, following Zhu et al. (2013).

The results of the regression analysis remain consistent with the initial findings. We see that the coefficient for the tier-invariant treatment effect in Model 0 is not statistically significant. Recalling the various forms of specification error discussed in Chapter 1, this non-effect may be due to differences in the treatment effect between groups. To assess this possibility, Models 1 and 2 include product terms that represent the treatment effect in each tier. The effect continues to be positive and statistically significant in the top tier and statistically non-significant in the lower tiers. In other words, rewards do not elicit additional productivity among less productive contributors. Behind the overall non-significant treatment effect in Model 0 is hiding the positive effect in the top tier. The mean treatment effect for the top 1% of contributors is estimated to be between 37% (Model 2) and 50% (Model 1) higher edit count over this period, while Model 2 also shows that the chance of a user making zero edit increases by 12.8% for each additional day that has elapsed since the editor's last activity before the treatment date.[11] I also directly tested for a difference in magnitude of the coefficients of the treatment effects between the top and bottom tiers using a Wald test. The null hypothesis is that the coefficients are equal to each other in magnitude. This results of the Wald test (χ^2 = 4.84, df = 2, p = .03) suggest that we should reject the null hypothesis and conclude that there is a statistically significant difference in the magnitude of treatment effect across tiers. Thus, the analysis confirms that the experiment was sufficiently powered to discern that the treatment effect of receiving a reward on contributor productivity among less-active contributors was either much smaller than hypothesized or non-existent.

[11] The amount of work performed on the day immediately prior to treatment was not a statistically significant predictor that an editor would make zero edits post-treatment. We also tested a model where we included the variable for immediacy of feedback as an independent variable predicting edit count. However, there was no difference between control and treatment groups in this regard, which suggests that the treatment effect was not affected by how immediately the treatment came after activity; consequently, we retained this variable as a predictor in only the equation for the zero edit count.

Table 1: Negative binomial regression estimates of mean productivity rate after 90 days

	Model 0		Model 1		Model 2	
	exp(b)	**S.E.**	**exp(b)**	**S.E.**	**exp(b)**	**S.E.**
Tier (control)						
Reference (91-95th percentile)	1.000	-	1.000	-	1.000	-
96-99th percentile	2.665***	(0.534)	1.980*	(0.559)	1.747*	(0.485)
100th percentile	11.14***	(2.105)	7.228***	(1.995)	6.169***	(1.678)
Treatment effect						
Barnstar	0.91	(0.174)	-	-	-	-
Barnstar * 91-95th percentile	-	-	0.658	(0.224)	0.683	(0.231)
Barnstar * 96-99th percentile	-	-	1.201	(0.251)	1.202	(0.241)
Barnstar * 100th percentile	-	-	1.507*	(0.241)	1.372*	(0.211)
Inflate (logit)						
# of days since last edit	-	-	-	-	1.128***	(0.017)
# of edits on day before treatment	-	-	-	-	0.148	(1.270)
Overdispersion parameter						
ln(alpha)	1.248	(0.073)	1.241	(0.072)	0.599	(0.046)
alpha	3.484	(0.253)	3.459	(0.250)	1.822	(0.089)
Model summary						
N	600	-	600	-	600	-
zero count (*n*)	-	-	-	-	124	-

Note: exponentiated coefficients; standard errors in parentheses; * p < .05, ** p < .01, *** p < .001

Analysis of status accumulation

The treatment effect of a reward on productivity was only positive and statistically significant in the top tier. In addition to exhibiting greater productivity, subjects in this top 1% treatment condition were significantly more likely to receive additional rewards from other contributors, relative to their control group. Twelve experimental subjects were subsequently awarded one or more barnstars from other contributors, compared to two subjects in the control group (χ^2 = 7.681, df = 1, p = 0.006).

In the top tier, the twelve contributors who received additional barnstars were no more productive prior to receiving their second award, compared to others in the treatment condition (Mann-Whitney U test: z = .743; p = 0.458). For this test, productivity was calculated as the total number of edits up to day 8, when the first additional barnstar was awarded. The result of the test remains unchanged when productivity is calculated up to day 82, when the last additional barnstar was awarded (Mann-Whitney U-test: z = .796; p = 0.426). This suggests that there may be some decoupling of reward from work; as some editors begin to accumulate additional awards, their new-found higher status in the community further increases their chances of receiving subsequent awards. This can occur due to preferential attachment, which creates a rich-get-richer phenomenon.

To appreciate the significance of this finding, consider the following hypothetical situation: You are a contributor to Wikipedia, and someone has just sent you a message on your user page thanking you for your hard work in improving the article on *Social Psychology*. You appreciate that someone noticed your efforts and decide to pay forward this good feeling by giving someone else a Barnstar. You look through some of the recent changes to other articles that you are working on, and you notice a couple of contributors whose work might merit a Barnstar. One person has contributed more than a dozen new citations and references to primary source materials for the article on *Anthropology*. A second person rewrote a few sentences making them more clear and added a link to another article. Which person do you give the Barnstar to? Perhaps you are having a hard time deciding, so you check the first person's user page. Sure enough, that person has won two Barnstars already, from last year. You see this as being a good sign that this person is a lasting contributor to Wikipedia. You're inclined to give this first person another Barnstar, but out of a sense of fairness, you decide to check the second person's user page as well. No awards. Does that settle the matter in your mind about who to give the Barnstar to? While every day many contributors get their first Barnstar, the evidence from this experiment supports the notion that high status contributors accumulate a larger share of rewards than would be due to them based solely on merit. Prior status may influence your decision.

Less active contributors have much lower visibility in the community, where visibility is built bit by bit with each additional edit that leaves a trace in the public record. Every edit automatically creates links back to the editor's user page, where previous barnstars are displayed. Because of this high degree of transparency in peer production, the contribution histories of productive editors are easy to review. All other things being equal, visibility increases in proportion to effort expended.

Just there was no treatment effect on productivity among the lower two tiers (90-99th percentiles), there is also no status accumulation among these contributors. Indeed, there was no instance where any of these contributors – neither in the treatment nor control group – received a barnstar during the observation period. In the top 1% control group, only two editors received Barnstars. We can understand this to be the background or natural rate of Barnstars.

The findings so far from the experiment can be summarized in two points: the most productive contributors get a boost in productivity as well as gain an advantage in subsequent status attainment since there seems to be some decoupling of work and reward that occurs at the top of the distribution. Taken together, these two findings support the notion that the incentive structure in peer production is broadly meritocratic; but since Wikipedia is a highly competitive environment where less productive contributors are at a competitive disadvantage to gaining recognition, compared to their more productive peers, contributors must surpass a high threshold of accomplishment before their contributions are rewarded by the community. The process of cumulative advantage merely exaggerates the status disparities at the top of the distribution that originate due to unequal merit. After this experiment was conducted, the Wikipedia community activated a new feature called *WikiLove* which made it easier for users to award Barnstars and other types of positive messages to one another. In the first day alone, this feature was used 248 times (excluding users who gave themselves a Barnstar as a form of self love). My research was particularly timely in that it experimentally tested Barnstars right as there was a massive increase in their use on Wikipedia. We can tentatively conclude that such "kindness campaigns" (as they are sometimes called) are helpful in retaining and motivating the top contributors, but this may create some disadvantages for less active users as it becomes more difficult for them to achieve a higher status in the community, consistent with Stewart (2005). As a result, status inequalities may become more rigid on Wikipedia. In the next section, I consider another possible negative consequence.

Analysis of contributor retention

The experimental results show that rewards can be used to sustain productivity among highly-active contributors at the top of the distribution, yet are ineffective in this regard for less-active contributors. While theory suggests that a boost in social standing would initiate a virtuous cycle of work and reward, we did not see evidence of this dynamic for the majority of subjects in this study.

Because the results of the experiment were contrary to expectations, I also explored an alternate possibility, that rewards may affect contributor retention instead of directly affecting productivity itself. The rationale is that it can be wrongheaded to only look at productivity, as this research does. Indeed, the vast majority of contributors never receive any rewards or social recognition for their efforts, yet they still plug away at their keyboards and contribute countless hours to the project. More generally speaking, individuals contribute to public goods for a number altruistic reasons. Thus, it is important to recognize that individuals are broadly guided by moral incentives as well as whatever social rewards they may receive in exchange for their work (Etzioni 1988; Goodin 1980). Goodin notes that in many instances, "material incentives destroy rather than supplement moral incentives" (p. 140). This may be relevant to Wikipedia's contributor community, particularly for those contributors who have yet to exert significant effort, relative to the most heavily involved peers. Premature recognition of their work may convey a different meaning to these contributors; instead of signaling recognition and status in the eyes of the community, these individuals may perceive being rewarded as a signal that their contributions are sufficient, for the time being, or come to expect being rewarded for their contributions. This proposition is consistent with the theory of interdependence in giving (Andreoni and Scholz 1998; Andreoni 1988), often observed in charitable work and donations: continued giving may be conditional upon people seeing equivalent effort being put forth by their peers (Andreoni 1990; Frey and Meier 2004; Shang and Croson 2009). Rewards at this stage of their volunteer careers may "crowd out" other motives and yield a counter-intuitive opposite result than how they were intended. This may be true even though other research has found that gratitude and recognition for work is appreciated (WMF, 2011).

To assess whether rewards affect contributor retention, I consider contributors' activity (i.e. days in which they perform any work) as separate from how active they are on those days (i.e. how much work was performed). When contributors are inactive for a sufficiently long stretch of time, I consider this to indicate a break in their volunteer tenure. If we observe different rates of such breaks in volunteer tenure across treatment and control groups, then this suggests to us that being rewarded has an effect contributor retention, in addition to affecting productivity directly. To evaluate this possibility, we first must define what counts as a break in volunteer tenure. Since I initially observed contributors for 90 days after the treatment, I suggest that if a contributor becomes inactive at some time after the treatment date, and remains inactive for the remainder of the time under observation, that this satisfies a definition for a break in volunteer tenure. But, what is to say that such individuals will not become active again on day 91 or any time after that? In order to account for this possibility, I collected additional data on the last observed activity, up through the current calendar month (December 2013), which is more than two years after the original study ended. If a contributor became inactive within 90 days after the treatment date, and never again performed any activity on Wikipedia, I argue that this is a sufficiently convincing indication of contributors' true end to their volunteer tenure. Intuitively, contributor retention can be thought of as the proportion of contributors who do not experience such a failure event (i.e. those who do not "drop out" permanently from Wikipedia).

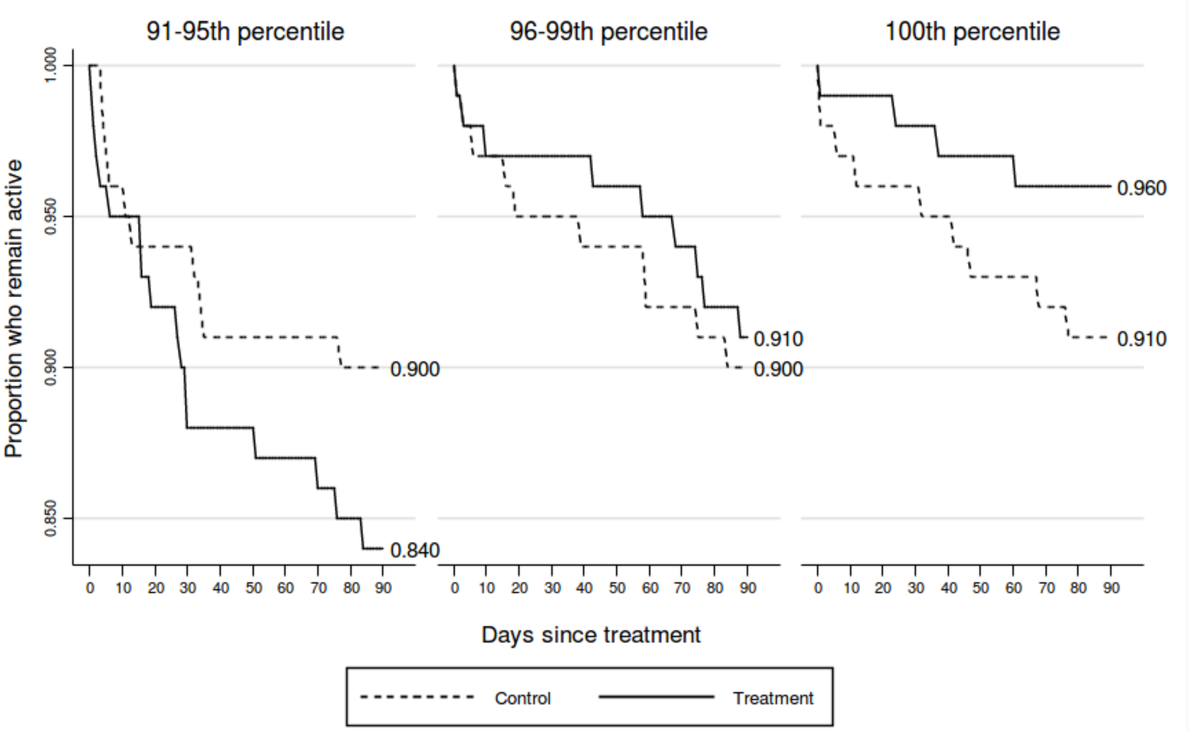


Figure 6: Retention of volunteer contributors over time by tier and treatment condition. This figure plots the survival function of editors not experiencing a failure event, which we define as becoming permanently inactive within 90 days of the treatment date. The lowest-tier treatment group has lower survival than its respective control group, whereas this is reversed in the top tier, providing evidence of differential effects of rewards on volunteer retention.

To formally assess whether there was a treatment effect on retention, I use survival analysis, which incorporates the notion of time at risk until a failure event occurs.[12] Since the failure event represents a permanent state of inactivity, contributor retention is the rate in which contributors remain active over time, conditional on their survival until such time; with each new failure, the pool of contributors at risk of becoming inactive shrinks. In Figure 6, I visually display contributor retention over time, which begins at 100% for each group and decreases with each contributor who experiences a failure event (i.e. permanent inactivity which begins during the 90-day observation period). From the graph, we see that retention tends to be lower for the less actively involved contributors overall; and retention is relatively lower for users in the lower-tier treatment group, compared to their respective control group, whereas this relationship is the reversed for the top contributors. Results of a Wald test of equality of survival curves between treatment and control conditions confirm this as well: for the 91-95th percentile ($\chi^2 = 4.24$, df = 1, p = 0.039), 96-99th percentile ($\chi^2 = 0.07$, df = 1, p = 0.795), and 100th percentile ($\chi^2 = 4.97$, df = 1, 0.026); these statistically significant differences for the lowest and highest tiers provides evidence that being rewarded can lead to either positive or negative consequences for contributor retention, as well as encouraging more work from the top contributors.

[12] Because we are interested in discontinuation, we excluded all editors who had no discernible activity after the treatment date. As discussed above, we cannot be certain that these editors have ever "received" (i.e. became aware of) their awards, so we exclude them from this analysis as well.

Conclusion

Overall, there was a surprising lack of consistency in experimental results for the top contributors compared to less active ones. Rewards sustain the efforts of high-level contributors, while not affecting the productive output of lower-level contributors. It is not surprising then that the most active contributors begin to accumulate additional rewards, while the lesser merit and visibility of contributors in the lower tiers seems to prevent cascades of recognition from emerging for them. This suggests to us that the incentive structure in peer production is broadly meritocratic.

Second, rewards also have differing effects on contributor retention. While rewards increase retention for the top contributors, I provide evidence that rewards may actually be counter-productive for some editors, particularly those who are not already sufficiently embedded within the core contributor community. Prematurely rewarding such contributors can lead to lower retention, a novel finding that warrants closer investigation in the future. I suggest that that scholars should explore this question in more detail using qualitative research designs, such as in-depth interviews, to more fully understand what rewards mean to contributors themselves. In doing so, researchers should make sure that they are sensitive to differences among contributors and include a representative sample of contributors from across the many different types and activity-levels of volunteers who make up the Wikipedia community, as I have done here. Given the highly-skewed distribution of work among contributors, it is insufficient to select a simple random sample from the entire population, which will vastly over-represent users with very low productivity, and may prevent researchers from having sufficient statistical power to discern clear differences that actually do exist among users in the small core contributor community. But simply performing a lot of work does not indicate that contributors have thoroughly aligned their identity and behaviors with the contributor community (Ghosh and Prakash 2000; Lakhani and von Hippel 2003; Lerner and Tirole 2002; Mockus et al. 2002). Rewards may not mean the same things to all contributors, despite survey evidence that suggests rewards are nearly universally appreciated. For example, recent research on this topic by Hill, Shaw, & Benkler (2012) suggests that some contributors are more inclined to “signal” their status to others in the community; such differences across contributors may account for some of the variability in response to rewards – for both productivity and retention – seen here.[13]

In sum, recipients of informal rewards who were not the most prolific contributors – the 91th-99th percentiles of contributors – did not exhibit the hypothesized increase to their productivity. The suggestion that failing to have one’s work recognized is what is holding back contributors from performing even more work does not seem to be supported by the evidence presented here. Contributors must already have a high level of involvement in the community before they can be motivated by use of informal rewards. In insufficiently qualified populations, recognition does not motivate recipients to increase effort, and may even do more harm than good by possibly undermining contributors’ intrinsic motivations to contribute to public goods online or by lowering contributor retention in other ways.

These findings provide us with significant insight into the incentive and reward structure of Wikipedia and peer production systems more generally. By comparing the effect of rewards across different tiers of users who make up the core of Wikipedia’s contributor community, we find that rewards have positive consequences for

[13] Hill, Shaw, & Benkler (2012) also note that users often receive barnstars at periods of local maxima in their contribution histories. This research suggests that it is also possible that the receipt of social recognition for one’s work may not merely occur at a local maxima, but may also create such local maxima if contributor activity declines after receiving a barnstar.

contributors at the uppermost part of the distribution. These consequences are largely positive (increased productivity), but initial arbitrary status differentiation can also exacerbate stratification when the involvement of already very productive contributors is reinforced through their accumulation of status, while more peripheral actors who are insufficiently involved and who do not typically receive recognition are not responsive to rewards; indeed, rewards may actually shorten their volunteer tenures. As a result, status and involvement can become even more concentrated among a small core of super-contributors, as is indeed the case in Wikipedia.

To conclude this chapter, it is prudent to take a step back and return to the bigger picture. The analysis in this chapter demonstrated the differential ways in which status awards affect contributor behavior. These largely do not support the idea that rewards are a mechanism that generate inequality in participation, even though they do have a large positive effect on the productivity of top contributors. Individuals rather "slot themselves in" to their self-chosen level of volunteering. However, status inequalities do seem to be driven by pursuit of status attainment and accumulation of recognition at the top of the distribution. Furthermore, we see a rather negative effect of receiving a Barnstar for some contributors, where immediately following the treatment, there was a higher rate of discontinuation in contributing. This suggests that rewards may supplant the intrinsic motivations for these contributors, which in the end may generate more participation inequality.

Understanding the incentive structure of participation is important because there is the concern that contributor turnover will undermine growth and sustainability of the project. But rewards only explain a small portion of the overall variation in participation. In addition to knowing what sustains participation, we need to understand what hinders and undermines participation. It is a starting point to suggest that "not being rewarded" will tend to erode a contributor's participation over time, and that under some circumstances, receiving a reward may be downright unhelpful.

One of the side-benefits of conducting this experimental research is having collected detailed, micro-level data on Wikipedia contributors. I reflected on the value of the dataset that I had put together for this line of research: in particular, the non-treated subjects in the control group constituted a representative sample of active contributors. Despite the limitations of having only observational data, it occurred to me that I could at least take a closer look at these contributors' behaviors to see if I could discern any common pattern that accompanied contributors who dropped out, which might become the basis for a more focused a line of new research. In the next chapter, I describe how this exporatory analysis of contributor retention led to the development of a full-fledged research project to test a theory proposed by Wikipedia observers about why contributors stop.

Chapter 3. Bureaucracy and decline

> In the beginning everyone knew each other. Once upon a time rules are unwritten, community rules. But it got to the point where more people were coming and have to write down these rules.[14]

Why people stop

To continue where the previous chapter left off, let us turn our attention to the puzzle of why volunteering stops. Of all the variation that exists in people's motivations and life considerations, can we identify a shared set of circumstances in the context of their volunteering under which people are more likely to stop or slow down their efforts? Parties interested in the success of peer production organizations would likely be interested in understanding what mechanisms decrease volunteer retention, so as to change or mitigate those circumstances to prolong volunteer tenures. A number of prominent Wikipedia critics have proposed the theory that "the main source of those problems is not mysterious. The loose collective running the site today, estimated to be 90 percent male, operates a crushing bureaucracy with an often abrasive atmosphere that deters newcomers who might increase participation in Wikipedia and broaden its coverage" (Simonite 2013).

When confronted by such assertions, two questions that a skeptical observer should ask is, "Is it true?" and "How do we know?" Surprisingly, the answer to both questions is both theoretically and empirically underdeveloped. By theoretically underdeveloped, I mean to suggest that critics who say that bureaucracy is harming contributor retention proceed from a rather simplistic conception of bureaucracy, both in the ideal type as well as how bureaucracy functions in Wikipedia in practice. By empirically underdeveloped, I simply mean that I do not believe they have proven their case with methodological rigor. (But this second point is a consequence stemming from the first inadequacy anyway.)

By leveling these two critiques, I do not mean to suggest that existing scholarship that attempts to understand bureaucracy in Wikipedia, and it's relation to contributor retention, is worthless. Quite the contrary. Rather, the point is that these questions – Is it true? How do we know? – are difficult to answer. This chapter is not the definitive statement of the problem, but the evidence from my analysis goes contrary to much of the existing "consensus" (among Wikipedia's most vocal critics, anyway) that bureaucracy is inimical to the project's success. Instead, I suggest that, for both individuals and the organization as a whole, the cause and consequences a growing bureaucracy are more nuanced – that bureaucracy serves as a stabilizing force for the organization but also makes it more difficult for newcomers to break in to the contributor community. As the previous chapter demonstrated, appreciation of these nuances can give us a more complete picture of how and why wiki works.

[14] Sydney Poore, quoted in Zhao (2012).

Decline theories

Widespread adoption of new information and communication technologies enables individuals to form relationships within organizations in ways that had previously been uncommon (Haythornthwaite and Kendall 2010). One example is the increasing role that consumers play in the production of digital goods (Gloor and Cooper 2011; Ritzer and Jurgenson 2010). Many successful organizations rely, in part or in whole, on users to create the content of their products in a type of mass mass participatory culture (Benkler 2006) facilitated by the Internet. Organizations like Wikipedia draw our attention to the troublesome analytic distinction between producers and consumers of goods (Humphreys and Grayson 2008; Tapscott and Williams 2008), since some of the consumers of the good are also responsible for its creation.

One of the more thoroughly asked and investigated questions regarding online peer production is why people volunteer, which I addressed in the previous chapter. But understanding why people volunteer focuses on only half of the puzzle. It is also imperative to discern why or under what circumstances individuals stop volunteering, since the success of these types of organizations hinges on the ability to attract new volunteers and retain existing ones. In this regard, contextual factors related to the volunteering experience may be more salient than individual-level factors related to motivation, yet the question of how context affects individuals' propensity to volunteer and their commitment to the volunteer work is one of the least understood aspects of the field (Smith 1994; Wuthnow 1998). Since peer production requires that at least some of consumers of a good also contribute to its production, the question of under what circumstances individuals discontinue their role as producers has obvious implications for the viability of this model of production. It also may point to a mechanism that generates the vast inequality that we see between producers, where the majority of work is accomplished by a small number of stable contributors, with a high degree of turn-over in a larger peripheral population.

Since peer production requires that contributors coordinate and collectively govern their work themselves, the structure of the organization itself becomes part of the good being produced (Demil and Lecocq 2006; Markus 2007). This evolving organizational structure in turn becomes the context within which subsequent volunteering takes place. The organization is a dynamic system that is formed and changed under the impulse of individual actions, just like the encyclopedia being created. Thus, our task is to consider how individual-level behaviors lead to the emergence of organizational-level structures which in turn may come to shape individuals' decisions whether to keep volunteering or in what ways they choose to do so.

Normative accounts of peer production describe its governance structure as "peer-to-peer" rather than bureaucratic. However, a number of observers have noted that as organizations employing peer production grow in size and number of participants, they display the tendency to develop increasingly formalized, bureaucratic structures. For example, Wikipedia saw a decline in the number of active participants after around 2006 (Angwin and Fowler 2009), which was also around the time that other research was showing that its organizational structure had become increasingly bureaucratized (Viégas, Wattenberg, and McKeon 2007). Prominent critics proposed "decline theories" that connected these two trends: to explain this sudden reversal of Wikipedia's growth, theories of the growth and formalization of bureaucratic structures, or *bureaucratization,* suggest that what was behind these macro level patterns was the growing bureaucracy that could be hostile to newcomers (Kittur, Chi, et al. 2007; Konieczny 2009a). For example, a survey of former contributors conducted by the Wikimedia Foundation found that while about half of former editors stopped contributing to Wikipedia for personal reasons, about 25% cited problematic interactions with other editors over the organization's rules as their reason for quitting (Wikimedia Foundation 2010). Similarly, Konieczny (2009) documented that some users believe that volunteering for the organization has becoming more

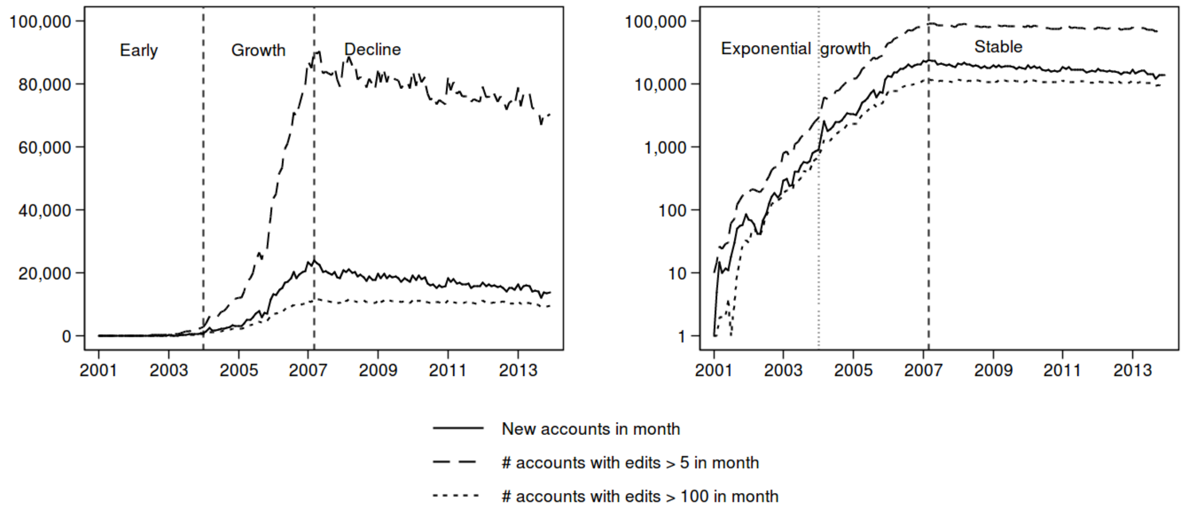


Figure 7: Two views of the growth of Wikipedia's editing community. Both graphs contain the same information but are plotted on different scales for the y-axis. The three trend lines are the number of new accounts in a month (solid line), number of accounts with 5 or more edits in a month (dashed line), and number of accounts with 100 or more edits in a month (dotted line). The left panel replicates the graph from Halfaker et al. (2012), while the right panel presents the same trend lines on a log scale, in which exponential growth appears as linear.

unpleasant due to a cabal of users “living in their own little bureaucratic world, creating useful content” (p. 169). The bureaucratization theory of decline suggests as the organization grew, volunteering entailed more frequent interactions with others in increasingly bureaucratized contexts, and needing to engage other users in bureaucratic context can turn people off to the experience and decrease their contributions or discontinue contributing altogether.

One significant articulation of the decline theory suggests that “the changes the Wikipedia community made to manage quality and consistency in the face of a massive growth in participation have ironically crippled the very growth they were designed to manage” (Halfaker et al. 2012). On Wikipedia, the cause of the growing bureaucracy was posited to be a rational response to the vastly expanding size of the contributor community (see Figure 7). Halfaker et al. (2012) posit that there are two interrelated causes: first, in an attempt to control vandalism of the project, semi-autonomous programs or “bots” may flag as suspicious behavior the contributions of newcomers, and some cases the contribution can be automatically reverted. Second is what the authors call “calcification of rules against newcomers.”

An alternative form of the decline theory focuses instead on the special roles and privileges within the Wikipedia community (Arazy, Nov, and Ortega 2014), focusing on how an informal organizational hierarchy has formed in Wikipedia despite its ostensible horizontal organizational structure. Of particular interest is the administrator role. See Figure 8 for growth in administrators, who are believed to shift their effort away from content production toward more managerial roles and actions. These two social dynamics are thought to perpetuate a cycle where growing numbers of managers oversee dwindling numbers of workers.

However, I suggest that these theories are wrongheaded and rely on an incorrect understanding of Wikipedia’s growth. In Figure 7, I show two ways of viewing the growth of Wikipedia's contributor

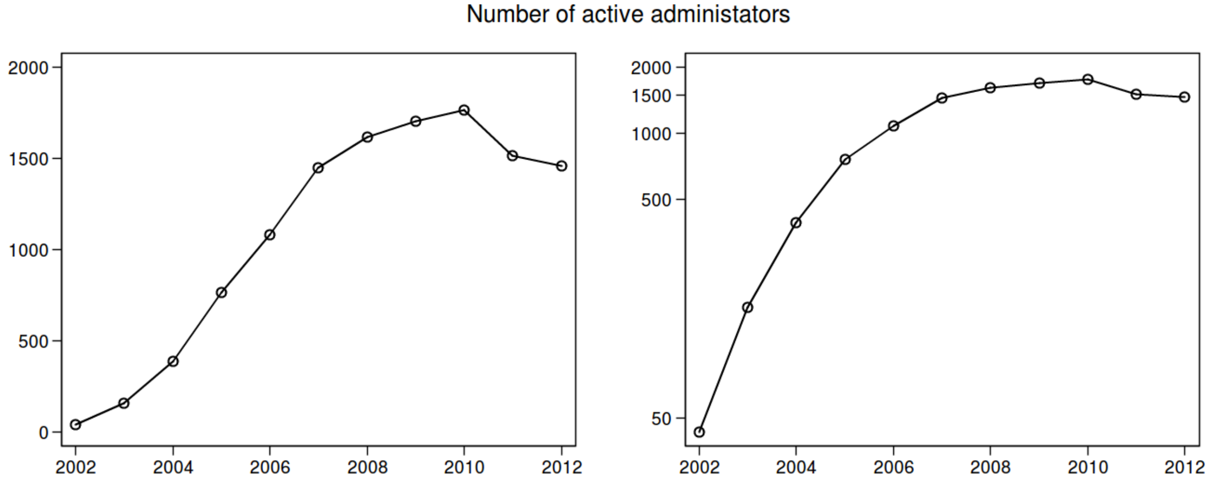


Figure 8: Number of active administrators over time. The left panel plots the count of active administrators on an untransformed scale, and the right plot uses a log scale on the y-axis. If this figure and the previous (Figure 7) were superimposed, the growth curves would appear very similar. Just as the size of the contributor community has stabilized, the number of administrators has also been relatively stable since 2007.

community. The left panel represents a a three-stage view (Halfaker et al. 2012), which partitions Wikipedia's evolution into three time periods: early (2001-2004), growth (2004-2007), and decline (2007-present). Critics in particular focus on the transition between the "growth" and "decline" periods, and attribute this dramatic halting of growth to the rise of Wikipedia's bureaucracy.

On the other hand, in the right panel, I plot these growth trends on a logarithmic scale, which turns exponential growth into a linear trend. In this scenario, it appears more accurate to characterize Wikipedia's editing community into only two periods: exponential growth (2001-2007) and stability (2007-present). In particular, the number of highly active contributors (with at least 100 edits in a month) has remained fairly constant since around 2007, even though there is has been a small decline over time in the number of new accounts and less active contributors. In this case, the critics may still be correct that the limits to Wikipedia's growth were due to the rise of Wikipedia's bureaucracy, but we must ask ourselves if any system can undergo unbound exponential growth? We can recall Boulding's infamous quip that "Anyone who believes exponential growth can go on forever in a finite world is either a madman or an economist" (Boulding 1973).

"Don't look now but we've created a bureaucracy"

The foundational statement for sociological thinking about bureaucracy comes from Weber's analysis of social and organizational management (1922). Weber identified six characteristics of the *ideal type* bureaucracy: professionalism, expertise, rules, impersonality, specialization, and hierarchy. Normative accounts of peer production, however, suggest that these types of organizations are inimical to bureaucracy as traditionally defined (Benkler 2002; Demil and Lecocq 2006). This is because fixed, hierarchical arrangements between contributors are absent. Instead, all individuals are responsible for coordination, planning, and governance of the collective activities. In such an environment (Raymond 1999), described as a "great babbling bazaar of differing agendas and approaches" (p. 30), governance of the organization occurs on a peer-to-peer basis with multiple foci of power and few top-down controls (Demil and Lecocq 2006). Accomplishing organizational goals requires participants to coordinate their actions with one another, but there are few formal institutional mechanisms to constrain their behavior. Instead, contributors must coordinate with one another to decide on what collective action to take, and to enact and complete it. As such, peer organizations have been compared to self-managed teams (Spek, Postma, and van den Herik 2006) that exhibit characteristics of "adhocracy" (the opposite of bureaucracy, as it were) where participation is focused around shared problem-solving rather than rule-following (Travica 1999). Organizations of this form are characterized by flexibility in contrast to what Weber described as the rigidity of bureaucracy's "hardening shell."

Nonetheless, it has been pointed out that in some cases in Wikipedia, a system of relations has evolved in such organizations that resembles several dimensions of Weber's bureaucracy. Scholars have explored this in some detail. For example, informal rules, policies, and guidelines – developed by contributors to aid in efficiently coordinating their multiplex collaborative work – have come to constrain behavior of other contributors during interactions between them (Butler, Joyce, and Pike 2008). While the development of these governing structures remains flexible and under constant negotiation, these communally-developed norms play an increasingly important role in the day-to-day operation of the organization. Enforcement of norms (Goldspink 2010) occurs on the micro-interactional level, which creates and reinforces power relations between individuals (Collins 1981). Such power relations play out on an interactional level (Shaw 2012) and over time, through these interactions, informal procedures solidify into rules and policies – in other words, participation becomes increasingly formalized (Forte, Larco, and Bruckman 2009). While this describes the broad contours of how the community is self-governed, disagreement remains about how is actually conducted.

On the one hand, some scholars argue the development of self-governance rules and policies remains decentralized (Forte and Bruckman 2008) or takes place through a "bureaucracy of peers" which develops out of a spontaneous division of labor (Aaltonon and Lanzara 2010). Others describe the situation as one where a small group of users come to specialize in distinct roles such as community manager. Even without reference to explicit, formal elevated privileges in the community (such as the administrator role), examples of authoritarian leadership can been found within the community (Reagle 2007) despite the ideal that all contributors are egalitarian peers. Still other research, to the contrary, suggests that Wikipedia has largely resisted Michel's "iron law of oligarchy" (1915), since no individuals dominate policy formation (Konieczny 2009a).

What is clear is that there has been some degree of specialization into management roles, as well as formalization of norms into policies, in Wikipedia's contributor community. Even in organizations that resist these characteristics of characteristics of bureaucracy, promulgation and attempts at enforcement of rules does indeed occur. The larger consequences of this growing bureaucracy, according to decline theory, is that participation may become more difficult to sustain as fewer new volunteers become long-lasting contributors and more experienced volunteers channel less of their work effort toward productive tasks. As such, the process of formalization within these organizations is thought to be one mechanism that produces the high degree of participation inequality that is commonly seen in peer production, which is thought to undermine the long-term sustainability of such organizations. However, these same practices are also thought to be one of the mechanisms that creates organizational stability by promoting a more regular, predictable, and efficient process of volunteer work. We need a way of assessing whether Wikipedia's growing bureaucracy tends to deter newcomers and create a managerial class, as some suggest, or to promote and enhance long-term engaged contributions, or both. In order to do so, I turn to the issue of how to define and measure this process of bureaucratization.

Measuring formalization in peer production

Butler, Joyce, and Pike (2008) use the terms rules, policies, and guidelines interchangeably to refer to aspects of Wikipedia's bureaucracy, and I adopt this practice as well. But, the way I define bureaucracy is in relationship to the type of work being performed and the organizational context it creates. I begin with the simple observation that, in peer production, all of the work necessary for creating and maintaining the organization must be performed by the community of volunteers. Since this occurs through repeated interactions among contributors in many different areas or parts of the organization, we must adopt a rather broad view of the ways in which bureaucracy can come about and affect contributor behavior. This is rather challenging since there is no way to simply view the "org chart" of Wikipedia's bureaucracy (Aaltonon and Lanzara 2010), nor are there a series of memos published by its (non-existent) Director of Human Resources that we can read to understand the nature of its rules – to which contributors may adhere or subvert, per their whims.

Nominally, the goal of Wikipedia is to write a compendium of human knowledge. Accomplishing these tasks involves combining the efforts of thousands of volunteers who coordinate their work with one another. Coordination has been defined as the impulse to "talk before you type" (Viégas, Wattenberg, and Dave 2004) although there is no imperative to do so. Nonetheless, most contributors find themselves in social interaction with one another in discussions that accompany every encyclopedia article or news story. These discussion pages are one of the primary sites where rules and policies are discussed and negotiated by contributors. For example, if an article is a biography of a living person, there are particular guidelines that the community broadly agrees should be followed, such as what counts as a reliable source of information about that person. These rules however are not set in stone, and contributors to such articles can spend an inordinate amount of time negotiating with one another what counts as "reliable" given the circumstances. Discussion pages are also a way for contributors to an article to talk amongst themselves about what needs improvement in the article's contents, and how to go about making such improvements. Disagreements are common (as would be expected) but one of the overriding principles of collaboration is to attempt to come to consensus. Edit conflicts, where one contributor changes the content of an article only to have another one override the change, can lead to a cycle of such back-and-forth edits (Sumi et al. 2011); to resolve such situations, contributors often refer to the three-revert rule (or "3RR"), which suggests that when such a pattern of editing goes beyond three rounds that the participants need to hash out their disagreement through discussion. This represents as serious of a "rule" as can be found on Wikipedia, with violators potentially being temporarily blocked from editing; but the Wikipedia policy page on this rule is full of caveats and exceptions that may suggest to contributors that the rule is somewhat flexible and open to negotiation. Each article's discussion page, then, becomes a site not just for coordinating work and talking about the substantive contents of the article, but also a space where policies and community guidelines are reconstituted through contributor interaction.

Another major coordination mechanism are "WikiProjects," which are self-managed groups of individuals working on similar subject matters within the larger organization (Kittur et al. 2009a). These groups are used to identify tasks that need to be performed (such as anti-vandalism or copy editing), provide resources pertaining to the subject matter at hand, and give contributors another forum to interact with one another and to share their experiences and facilitate collaboration. For example, there are WikiProjects within Wikipedia on topics that range from culture and the arts to science, technology and engineering. Research on users who join a WikiProject group (Kittur et al. 2009a) finds that these editors tend to focus more of their work effort on "pages relevant to the project and an increase in coordination and discussion work" (p. 7). However, research has shown that this type of coordination work can come at the expense of individual productivity, as editors expending more of their effort on coordination tasks. Group work in peer production, then, helps to combine multiple contributors' work on to common tasks, but does so by imposing some structural constraints on individual behavior in order to channel efforts toward accomplishing shared goals.

These are but two of many contexts within which coordination work on Wikipedia takes place. Coordination is the development, enactment, and reconstitution of structures meant to align one's behavior with others. Early in Wikipedia's history, the proportion of overall effort that went into such coordination work was much smaller than it is today; however, consistent with theory of bureaucratization, the necessity of coordinating one's actions with others tends to increase as overall participation grows.

I draw a distinction between coordination work and a more explicit type of bureaucracy. Efforts at collective self-governance is the most commonly cited mechanism for peer production's expanding bureaucracy. For example, Aaltonon & Lanzara (2010) cite "the management challenges of a complex distributed production system . . . must face a second major problem, that is dispute resolution" (p. 8). As I described above, contributors often get caught up in controversies and conflicts with other users such as "edit wars" (Brandes and Lerner 2008) where users dispute what content should be included within an article. This is part of the regular practice of coordinating work with others. Resolving such disputes often means referring to community guidelines and policies. But where do these originate from in the first place? Most policies emerge out of deliberations from coordination conflicts, with contributors establishing ways of avoiding or resolving such conflicts in the future. As the opening quotation in this chapter demonstrates, as more people started to contribute to Wikipedia, there was more of a need to write down these community deliberations and decisions so they could be referred to when those situations came up again. Because there are few means available to contributors to dictate policy by fiat, as in a traditional bureaucracy, attempts to enforce such policies can simply recreate discussions and deliberations of such policies. For example, actions taken in good faith can be perceived by others as intentionally disruptive, leading to situations where new contributors find themselves in the middle of arguments over policies that they were unaware they were violating, which were discussed and enacted before they even arrived on the scene. For example, one of the "pillars" of the community is the dictum to "ignore all rules" if they get in the way of improving or maintaining the organization's goals, a rule which is ironically often cited during disputes, which are thought to come to dominate so much of contributors' time that they can eventually get disgusted with the process and leave the project altogether. The ever increasing amount of self-governing work required to support Wikipedia's growing population supports the notion that decentralized consensus decision making is a worthy principle but cumbersome in practice, or in Polleta's wry quip, "freedom is an endless meeting" (Polletta 2012).

Kostakis (2010) explored another element of peer governance, particularly in Wikipedia, that has garnered considerable attention: the internal struggle between competing visions of the scope of the encyclopedia. One the one hand, "Inclusionists" argue that the project should aim for wide coverage that includes topics of questionable notability and retains articles that may yet to meet minimum quality standards; on the other hand, "Deletionists" argue that the encyclopedia should be more conservative in its coverage by excluding trivial topics and insufficiently-developed articles. Newly created articles often become battlegrounds for individuals from these competing camps to attempt to enforce their vision for what Wikipedia should encompass. Contributors can find themselves embroiled in these disputes for simply attempting to contribute to such articles.

These are two examples of the types of self-governance activity required in peer production. Taken together, effort spent on coordinating behavior with others and work done to enforce self-governance rules represent two facets of bureaucracy that may reduce contributors' productive activity or discourage them from continuing to volunteer altogether. Formalization is the extent to which individuals inevitably encounter coordinating and governing context during their volunteer work. For example, over time, the overall composition of work effort in Wikipedia has undergone a shift from content pages (encyclopedia articles) toward non-content pages (discussions, policy pages, dispute resolution) as a percentage of total work. This has been attributed to profound growth in the size and scope of the organization (F. B. Viégas et al. 2007). In other words, Wikipedia has experienced a shift toward a more formalized work environment, and the consequences of such a shift are theorized to lead to lower contributor retention and a shift from productive activity to bureaucratic activity. In the analysis that follows, I assess whether there is empirical support for these hypothesized effects of bureaucracy.

Because my research design relies on observational data, the analysis can only suggest what social forces are at play. The experimental control and careful attention to teasing out causal ordering in the Barnstar experiment (Chapter 2) cannot be replicated in this context, since it is not possible to increase the bureaucratic workload of contributors experimentally.[15] As an alternate approach, I analyze two datasets from Wikipedia to attempt to triangulate my findings. The first dataset consists of the control group from the experimental research in the preceding chapter. This control group is a representative sample of the most highly productive Wikipedia contributors from 2011. The second dataset is a representative sample of all contributors to Wikipedia from its inception through the first half of 2009, which contains both low and high activity users.

Much of our current theorizing about why individuals stop contributing to peer production derives from interview or survey data (Wikimedia Foundation 2010). One reason we should be cautious in over-generalizing from these findings to the broader population of contributors is selection bias. The methodological challenge lies in the ephemeral nature of online organizations: there is no completely reliable way to sample former contributors, leaving us with mainly anecdotal accounts to go by. The inability to contact former contributors in a systematic way limits what we can conclude about what former contributors tell us about how bureaucracy affected their choice to stop contributing. There are, without a doubt, countless anecdotal cases of users who egregiously promulgate their own little bureaucratic world on other users. However, given these organizations' scale in terms of overall number of contributors, we should not hastily attribute larger patterns of volunteer decline to these encounters; I argue that decline theorists are too quick to make this link. Given the limited generalizability of survey research, we require an alternate approach to the problem.

I address the question of how bureaucracy affects volunteering by linking patterns in individual contributors' behavior with the various types of work contexts in which their behavior takes place. If decline theories are correct, we should observe predictable correlations between bureaucracy, work patterns, and contributor retention. These are represented by the following hypotheses: (1) for new contributors, performing more bureaucratic work will increase the chance of discontinuing volunteering altogether; (2) as contributors perform more bureaucratic work, they will experience a decline in their overall productivity. I test these hypotheses derived from theory.

[15] Research ethics is the primary "obstacle" to replicating the experimental research design in this context. While some research, notably Zhu, Kraut, & Kittur (2013), used an experimental design to give "negative" feedback to contributors, I disagree strongly with this approach since the research design is intentionally meant to discourage participation. While their findings were not entirely consistent with theory, this nonetheless constitutes real harm to the research subjects and the organization they are volunteering for. The Wikimedia Foundation and other scholars have addressed this issue in their own internal research as well, where contributors report that they are turned off by negative feedback from peers.

Preliminary analysis

To begin, we must establish the existence of the problem. I performed an exploratory analysis using the subset of the data from the experiment described in the previous chapter. I retain data on only the control group, which constitutes a stratified, random sample of highly-active Wikipedia contributors (top 10%) in 2011. Using on the data from the control group, I explored the pattern of when contributors dropped out using survival analysis, to see if there was a correlation between the type of work contributors perform and their rate of discontinuation.

The original control group size was 100 per tier, for a total of 300. But a number (9) of editors deleted their accounts or were administratively banned either temporarily or permanently subsequent to the start of the experiment, leaving leaving a total of 291. Of these, a total of 67 contributors (23%) drop out of Wikipedia, never to be seen again. Were there any distinct characteristics in activity that corresponded with these disappearances?

I use survival analysis to explore this possibility. Survival analysis is used to describe and predict when an event will occur (Allison 1984). Another name for survival analysis is event history analysis or failure time analysis. In the case of Wikipedia, the "event" is when an editor makes his or her final contribution, which can be considered the "failure point" or when a contributor "drops out" of the community. Since we are talking about contributors who (during this period) were very active, such failures happen only to a fraction of the contributors during the observation period. However, if the failure occurred after the 120 day observation window, we say that observation is right-censored (Allison 1984). An important reason to use survival analysis in this instance is because of its ability to handle right censoring. But before using this technique, I first verified that the different survival curves between the tiers does not violate the proportional-hazards assumption. Figure 9 graphically compares the observed survival curves with the predicted curves, separately for each tier, which visually confirms that the statistical tests of the proportionality assumption within each tier is met. That is to say, there are proportional differences in this rate between tiers (which is consistent with the analysis in the previous chapter) but their form is otherwise the same. To account for this, I use a stratified Cox proportional hazards model (Andersen and Gill 1982; Cox 1972).

The parameter estimates are reported as hazard ratios. This permits an extremely intuitive interpretation, as they are similar to odds ratios in logistic regression.[16] Allison recommends that the data should contain information on at least 10 failure events for each covariate in the model. With 67 events, it is prudent to proceed with caution beyond six simultaneous covariates. However, like most regression models, misspecification error due to omitted variables can severely bias the parameter estimates. For this reason, I evaluate a series of alternative model specifications. This exploratory analysis, while not definitive, can at least establish the link between bureaucracy and retention, and give us some indication of how to move forward with a more robust analysis of this topic. I present the results of this analysis in Table 2.

[16] A simple mathematical transformation yields the percent change in risk for a one unit increase in the predictor: take the coefficient and subtract 1, then multiply by 100. A coefficient of .75 implies a 25% lower hazard: (.75 – 1) x 100 = – 25%. A coefficient of 2.5 represents a 150% higher hazard: (2.5 – 1) x 100 = 150%.

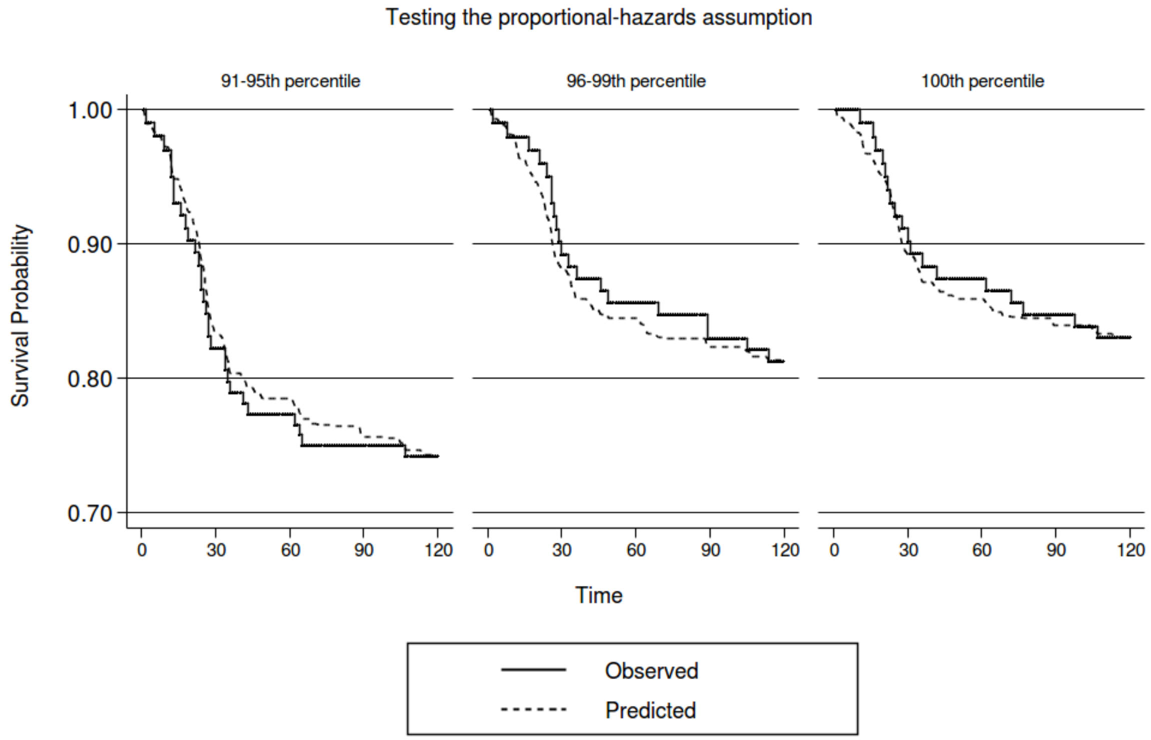


Figure 9: Testing the proportional-hazards assumption. This figure plots the Kaplan-Meier observed survival curves with the Cox predicted curves. The closer the observed values are to the predicted, the less likely it is that the proportional-hazards assumption has been violated.

Discussion of preliminary findings

In all models, the coefficient for total activity is is less than 1 and statistically significant. This indicates a negative relationship between total activity and experiencing a failure event. This should make intuitive sense: if you are highly active in the period under observation, it is unlikely that you will cease contributing altogether forever.

In Model 2, the coefficient for user-to-user activity is above one and statistically significant. Contributors who dedicate a larger proportion of their overall efforts to communicating with other users have a higher chance of discontinuing their volunteer activity. The magnitude of this coefficient suggests that a contributor whose proportion of user activity is one standard deviation above the mean has a 20% higher chance of experiencing a failure event. However, it is important to not attribute causality to this correlation. It is akin to saying that a person who is observed talking to more people at a party is at greater risk of leaving the party; but talking to people is not causing the person to leave, since it may be the case that such a person is saying goodbye to everyone before leaving. Nonetheless, after controlling for contributors' total activity, engaging in much social activity can decrease retention. Dedicated volunteers are here to work, not talk for the sake of talk; in the community, this is informally referred to by the phrase, "Wikipedia is not a social network."

Moving on, we see no association between proportion of coordination activity (Model 3) or governance activity (Model 4) and hazard of discontinuation. Retaining user activity in the model, after controlling for coordination (Model 5), governance (Model 6), and both (Model 7), we see that the finding for user activity is robust: there remains a positive and statistically significant relationship between user communication activity and discontinuation.

Finally, I consider potential non-linearity in the risk of discontinuation. This tests a "Goldilocks" proposition of contributor retention: when a contributor engages in too much or too little of these types of bureaucratic or social activities, it could indicate that the contributor is insufficiently engaged with the community or conversely too engaged in only one aspect of the community to the detriment of other essential aspects. In either case, these contributors may be at a heightened risk of discontinuing their volunteer tenures. Model 8 retains the fully specified linear predictors and adds a quadratic term for user activity. Neither is statistically significant in this model, which is likely due to the quadratic and linear terms being highly correlated in the absence of there being a true curvilinear relationship with the failure event.

However, the linear term for user activity is statistically significant in Model 9, which includes a quadratic term for coordination activity. The quadratic term is also positive and statistically significant. This suggests that the relationship between proportion of coordination activity and hazard of experiencing a failure event is U-shaped: very low and very high proportions of coordination activity are correlated with higher rates of failure. Intriguingly, in both Models 8 and 9, the linear term for governance activity becomes positive and statistically significant. However, Model 10 suggests that there is no non-linear effect of governance activity on retention. I believe that we should not over-interpret these findings, but instead use the results of this survival analysis to guide the construction of a more robust analysis. This suggests that contributors who are embedded in the community are most likely to be retained. Dedicating a high proportion of one's effort to user communication is associated with lower retention. Controlling for other covariates, contributors whose governance activity is one standard deviation above the mean have a higher risk of discontinuation. The relationship between coordination activity and retention appears to be curvilinear such that too high or too low levels of this type of activity may be detrimental. Overall, these findings support the notion that too much non-productive activity, such as engaging in community discussion of rules, policies, and other types of governing behavior, can be harmful to contributor retention, as can too high a proportion of social activity.

Based on this preliminary evidence, it seems incumbent upon me to develop, in a theoretically informed way, a more methodologically appropriate analysis of these correlations, which would take into account the time-ordering of contributors' activities across their entire volunteer tenures. I do so in the remainder of this chapter. My initial findings speak to an important proposition that has been suggested by prominent observers of Wikipedia's growth and success; they note that greater non-productive activities on Wikipedia are correlated with lower contributor retention. The strongest form of this argument asserts that as Wikipedia has become more bureaucratized – requiring more coordinating and governing activities from its contributors – fewer contributors will stick around for the long haul.

Table 2: Cox stratified proportional hazard estimates of contributor failure

	Model 1	Model 2	Model 3	Model 4	Model 5
Productivity (ln)	0.474 ***	0.469 ***	0.473 ***	0.469 ***	0.467 ***
	(-4.89)	(-5.17)	(-4.94)	(-4.96)	(-5.26)
User talk		1.202 **			1.208 **
		(2.79)			(2.80)
Coordination			1.028		1.048
			(0.39)		(0.65)
Governance				1.095	
				(1.85)	
N obs.	34920	34920	34920	34920	34920
pseudo-ll	-328.700	-326.700	-328.700	-328.100	-326.600

	Model 6	Model 7	Model 8	Model 9	Model 10
Productivity (ln)	0.464 ***	0.462 ***	0.468 ***	0.492 ***	0.461 ***
	(-5.24)	(-5.34)	(-5.19)	(-4.94)	(-5.35)
User talk	1.205 **	1.212 **	0.889	1.199 **	1.211 **
	(2.80)	(2.81)	(-0.43)	(2.83)	(2.79)
Coordination		1.052	1.062	0.562	1.051
		(0.68)	(0.90)	(-1.78)	(0.65)
Governance	1.099	1.102	1.117 *	1.108 *	1.143
	(1.90)	(1.92)	(2.31)	(2.13)	(0.83)
User talk^2			1.107		
			(1.23)		
Coordination^2				1.121 *	
				(2.31)	
Governance^2					0.995
					(-0.23)
N obs.	34920	34920	34920	34920	34920
pseudo-ll	-326.100	-325.900	-325.300	-322.500	-325.900

Notes: hazard ratios with t statistics in parentheses; * p<0.05, ** p<0.01, *** p<0.001

I believe it is clear why the evidence presented thus far is not entirely sufficient to form any conclusive answers regarding this proposition. For one, these contributors are all observed during only a short slice of time in 2011. We would like to see if and how these correlations change over the course of Wikipedia's decade-long history. Second, this analysis may suffer from *informative* censoring because I do not take into account other characteristics of contributors such as when they started contributing to Wikipedia or how far contributors were into their current volunteer "tenure" at the time of observation. To overcome these limitations, I analyze a representative sample of users over the entire course of their contribution histories. I now turn to a discussion of the methodological approach that I use to address this pressing question in a more satisfactory way.

Methodology for full analysis

Data

To overcome the limitations of the preliminary analysis (presented above) required me to acquire a representative sample of contributors to Wikipedia. I obtained these data from a data dump of the full longitudinal record of the English-language Wikipedia from its inception in 2001 through 2008. Every user contribution to this organization is recorded in a time-stamped database, which can be downloaded from the Wikimedia Foundation (the nonprofit that provides the technological infrastructure for these projects). After retrieving these databases, I processed them with a Perl script written by Georgi Kossinets and imported them into a local database for analysis. Because of the large size of the full database, I obtained a random sample of 10,000 users using equal probability of selection; I cleaned the data to remove anonymous users[17], known spam accounts, bots (automated software programs designed to perform routine maintenance work), and blocked users, to yield my final sample: N = 4748.

A few notes on the sample: The data represent the complete edit-histories of work performed by each user. This offers a significant improvement over the data presented at the beginning of this chapter, since those data represent only a limited time-slice (121 days) from a sample of highly-active users during the period under observation. In this case, the sample is representative of the population of all Wikipedia contributors (through 2008), and including the complete history of all activities undertaken by users in the sample. However, because there were only a small number of contributors in the sample in years 2001 and 2002, these editors were not retained in the analysis; the analysis starts in 2003. The dataset also contains data through the end of 2008, but the longitudinal analysis of contributor productivity only includes contributors who began prior to June 2008, in order to allow me to have at least 6 full months of data on their contribution histories. (For the logistic regression analysis of contributor retention, which replicates the preliminary analysis presented above, I included all contributors who began prior to October 2008, in order to have at least two months of records of their activity.)

[17] Anonymous users are only identifiable by their generic IP address, which may not yield a one-to-one correspondence to an individual contributor. For example, several anonymous contributors using a university computers may have share identical IP addresses. Because the analysis in this paper models changes to individuals' contributions over time, it is imperative to reliably identify what work was performed by what individual, hence the necessity to eliminate anonymous users from the analysis.

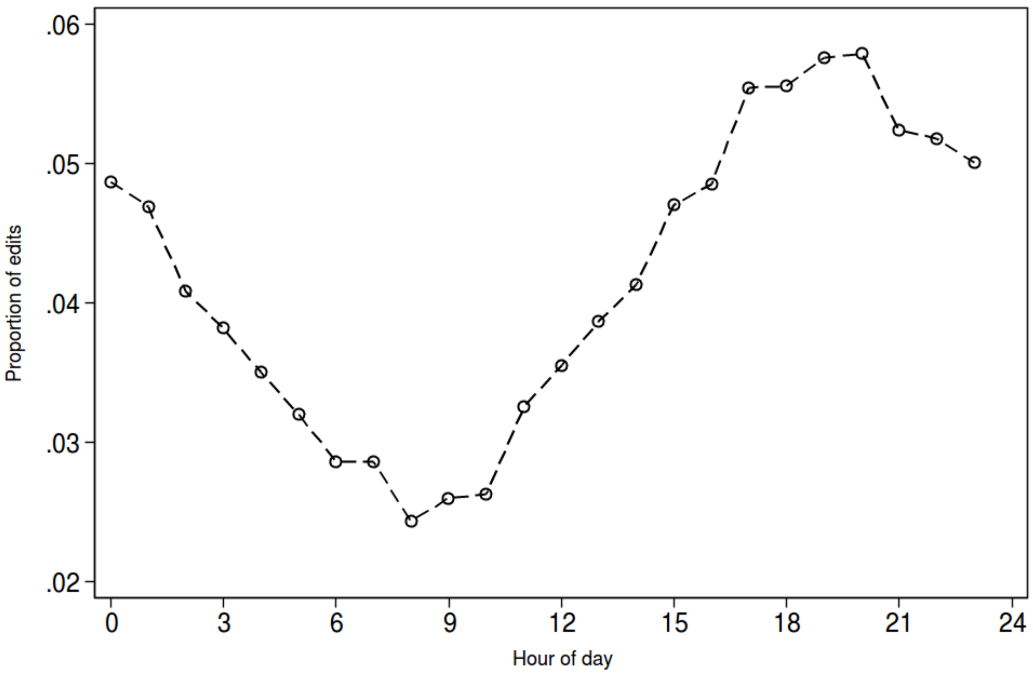


Figure 10: Circadian rhythm of Wikipedia contributors.

The reason the data is not more current is because the Wikipedia database from which this sample was extracted simply became too large to process and manage. The Wikimedia Foundation regularly makes available "database dumps" of their projects; actually, two versions of the database were made available: one containing metadata on user contributions and a second containing data on the actual contribution itself. For example, the metadata might contain a record which conveys information that *on May 1, 2008, user:Mike_Restivo edited the article "Sociology" which at the time was 32kb, and the user left the following comment: "Added information about the famous sociologist Michael Schwartz."* By contrast, the second dataset included all the previous information, as well as the actual, substantive addition, modification, or removal of content: *"Michael Schwartz is a famous sociologists who received his Ph.D. in Sociology from Harvard University and taught at Stony Brook University until his retirement in 2013."* The size of the metadata dataset was larger than 4 gigabytes, with the full dataset being nearly an order of magnitude larger. By 2009, this was indeed already a very large dataset. The Wikimedia Foundation continued to make more of these full dataset snapshots available for later years, but they were simply too large to process on a standard desktop computer. Thus, I was limited to using the most recent data that I could practically manage. (More recent developments in the field of computational social sciences, since this research was conducted, could be useful in overcoming this limitation.)

These data trace data contain extremely detailed records about contributor behavior. For example, Figure 10 shows that the contributions follow a predictable circadian rhythm (Yasseri, Sumi, and Kertész 2012). Contributor frequency peaked (perhaps not surprisingly) in the late evening and dipped to its lowest rate after dawn.[18] For the analysis of contributor retention, I analyzed the first month (30 days) of recorded activity for each user, and correlated this with contributor's probability of continuing activity into a second month. For the analysis of contributor productivity, I chose to aggregate the data into discrete time periods yielding user-months as the unit of analysis. I recognize that choosing to temporally group these data does cause a loss of information regarding the specific ordering of individual events. However, the modeling strategy appropriate to the research question requires using such discrete time periods. Exploratory analysis made it clear that more frequent temporal frames (daily or weekly) yielded too sparse of records with many empty periods, while periods longer than monthly would drastically limit the ability to make inferences about the time-ordering of events.

Research design

The overall plan of the research is to better understand how bureaucracy affects volunteering by linking patterns of contributor behaviors with the types of work contexts in which their behavior takes place. If the bureaucratization theory is correct, we should observe correlations between bureaucracy, contributor retention, and the overall shape of work patterns. These are represented by the following hypotheses: (1) for new contributors, performing more bureaucratic work will increase the chance of discontinuing volunteering altogether; (2) as contributors perform more bureaucratic work, they will experience a decline in their overall productivity. I test these hypotheses derived from theory.

Figure 11 shows the proportion of contributors who began in a particular year who continued contributing to Wikipedia beyond their first month, from the sample described above. As the figure shows, this initial turn-over rate for new contributors has increased for Wikipedia (i.e. the proportion of new editors who endure past a first month has declined). The large confidence interval for year 2003 is due to the relatively small number of contributors in the sample from this year. This visually confirms descriptive accounts about Wikipedia's contributor community that while many people try their hand at contributing to the project, over time a greater proportion of those curious individuals who started contributing also stop within their first month. Decline theories suggest that bureaucracy is the driving factor that leads to higher rates of discontinuation.

To test the hypothesis that bureaucracy affects the retention of new contributors, I use logistic regression to predict whether an editor continues to contribute to Wikipedia beyond one month. After controlling for the overall downward trend in contributor retention over time, can we understand more of the variation in contributor retention – and hence conversely, when people stop – from the context of early volunteer experiences? The initial experiences of contributors is thought to be highly predictive of their continued involvement in the project, and the long-term trajectories of contributors' volunteer tenures can be discerned through analysis of early activity (Panciera et al. 2009). In this first step of the analysis, I analyze how bureaucratic work contexts correlate with retention.

[18] The times are adjusted so that the new day begins at 12:00 am UTC, based on the local timezone setting for each contributor. Where this information is missing, I used UTC-06:00 (which corresponds to U.S. Central Time Zone). These data are not entirely reliable since contributors to the English Wikipedia can be located anywhere in the world. Nonetheless, on the aggregate, these data are illustrative of the overall pattern of when people contribute to the project.

In the second stage, I build upon the same indicators from the first stage to consider how bureaucracy affects contribution patterns over time for those users who survive past one month. With data measured at the user-month level of analysis, I use longitudinal panel models to estimate how a contributor's monthly productivity changes over time as well as how contextual work patterns shape future productivity.

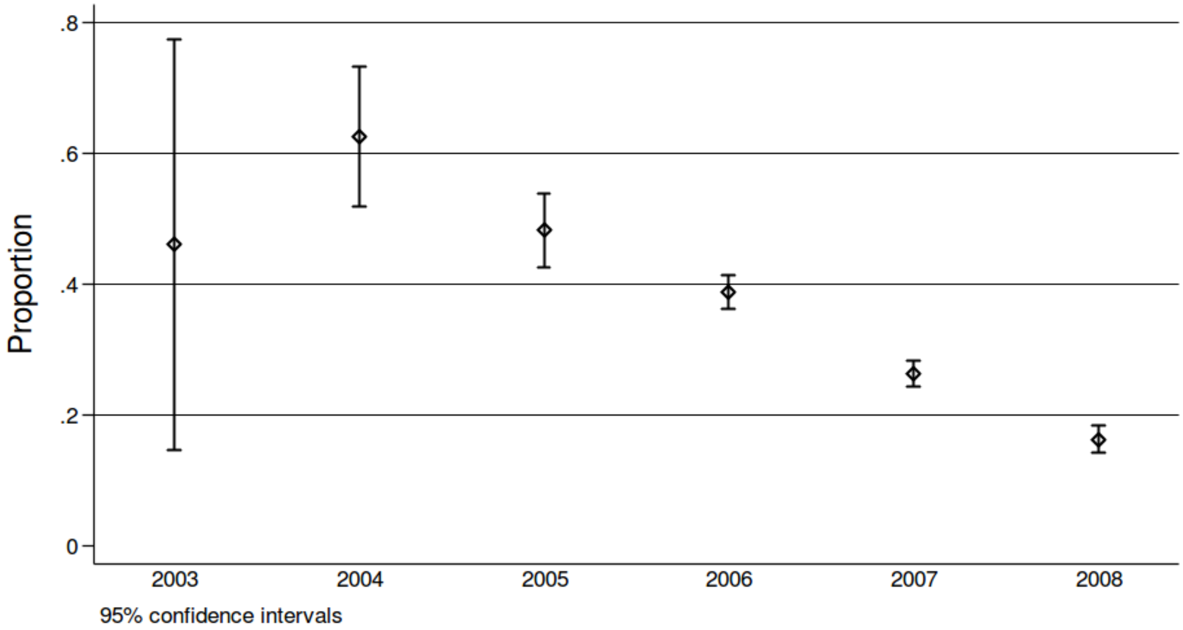


Figure 11: Proportion of contributors who continue beyond one month, by year, with 95% confidence intervals

Analysis of retention

Dependent variable

Fail: The dependent variable for the logistic regression analysis in the first stage is a dichotomous measure of whether an individual continued to contribute to the project past one month, where a positive value (fail = 1) indicates a failure to continue to a second month. Thus, a positive relationship in this context is a higher risk of failure.

Control variables

Time: As shown in Figure 11, there is a declining rate of contributor retention over time. Therefore, in all models, I control for the the calendar month when a user begins contributing to the project. I use month instead of year because it provides more precision, and permits me to also consider a non-linear effect of time. However, in results not presented here, there was not support for a non-linear relationship between time and contributor retention. Therefore, I use only the linear measure of time. The calendar begins in January 2003, so a contributor who begins in February 2003 will have a one-unit higher value of time. Therefore, the coefficient can be understood as the change in odds of failure for each later month when an editor first begins contributing.

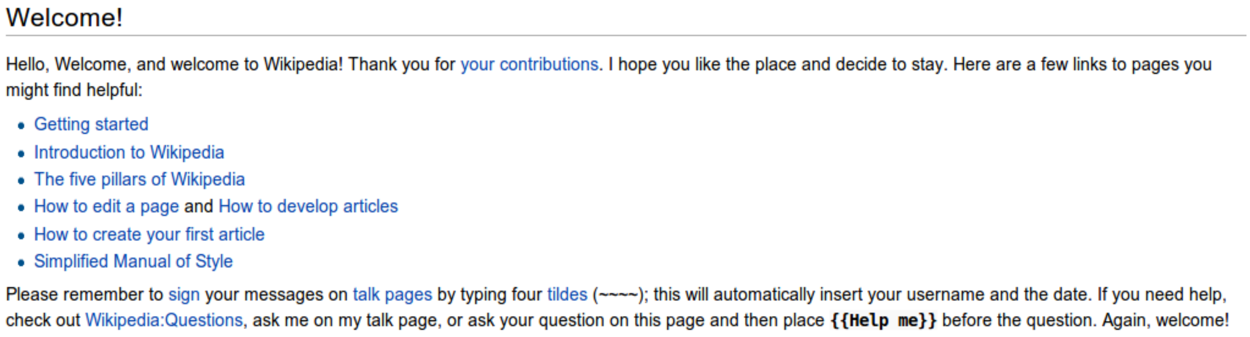
Independent variables

Productivity: As an independent variable, I consider the productive work output in the first month. This measure indicates each contributor's amount of productive activity over time (in this case, one month), or productivity (Kittur et al. 2009b; Kittur, Suh, et al. 2007). To be classified as productive work, a user's contribution must be aimed at producing content that is visible to readers of the site. This primarily occurs through edits to encyclopedia articles or creating news articles, but also includes preparation of multimedia content (such as images or audio files) and development and application of the categorization schema for a project's content. I transform the using the natural logarithm data because of their skewed distribution. As the preliminary analysis presented at the beginning of this chapter shows, the more productivity that a contributor engages in, the lower chance of discontinuation.

Peer-to-peer communication (user talk): From the analysis at the beginning of the chapter, it is incumbent upon us to consider how actively engaged a contributor is in the many different roles and types of social interaction that are part of peer production. This variable measures the amount of informal, peer-to-peer communication engaged in by the contributor. Because Wikipedia contributors engage in multiplex relationships with one another – that is to say, they communicate and interact in multiple ways depending on the roles and the types of exchanges in their social relationship – user talk is the least formalized type of social interactions on Wikipedia. However, even this interaction channel has the potential for becoming bureaucratized. For example, some new users will be greeted with a welcome message from another user on their user-talk page. These messages have been formalized in a number of pre-designed "templates" such as the one in Figure 12.

Nonetheless, most user talk is casual and informal. It is also the way in which contributors can reward one another (as with Barnstars, in Chapter 2), send words of appreciation when another is helpful or kind, and is generally the most "social" of all types of interactions on Wikipedia. Because these occur either on the focal user's talk page or on another user's talk, a conversation can often times become split between the two: for example, I may add a comment to another user's talk page, and that user in turn may reply on my talk page. Other times, the conversation may all take place on one user's talk page; and other users may chime in with their voices, too, since the conversation is public.

To construct this variable, I sum the monthly count of this type of activity for each user. I then add one to this count and transform it using the natural logarithm so that values of zero indicate no activity of this type.



Welcome!

Hello, Welcome, and welcome to Wikipedia! Thank you for your contributions. I hope you like the place and decide to stay. Here are a few links to pages you might find helpful:

- Getting started
- Introduction to Wikipedia
- The five pillars of Wikipedia
- How to edit a page and How to develop articles
- How to create your first article
- Simplified Manual of Style

Please remember to sign your messages on talk pages by typing four tildes (~~~~); this will automatically insert your username and the date. If you need help, check out Wikipedia:Questions, ask me on my talk page, or ask your question on this page and then place **{{Help me}}** before the question. Again, welcome!

Figure 12: Example of welcome message template

Coordination: I also consider how active a user is in coordinating his or her work efforts with other members of the organization. The primary channel for coordination is the article-talk pages, where users discuss improvements to an article. Each article has a corresponding talk page, which provides a forum where users can interact and collaborate with one another about their shared work. I also include efforts at group coordination such as WikiProjects, which are sub-communities within Wikipedia that aim to improve articles within a particular knowledge domain. Coordination is generally thought be enabling, since the purpose of coordination is to facilitate production. However, as I reviewed before, coordination work can also impede work and the work of others. I measure this variable as the amount of coordinating work performed in a month, and I transform by adding one and then taking the natural logarithm so that values of zero indicate no activity of this type.

Governance: I consider how active a user is in taking part in the basic self-governance activities of the organization. This includes discussion, development, application, and attempts at understanding and enforcement of community rules, procedures, norms, and standards, among other actions. Governance includes new users learning how the community works as well as discussing disagreements or disputes with other contributors, resolved through the process of consensus decision-making in the community. Governance work can be either coercive or enabling, depending on the particular type of interactions contributors have and whether they view these positively or negatively. Decline theories have focused on the negative consequences for coordination and governance work, but contributors who engage in these dimensions of activity also form ties within the contributor community and may come to better understand and fit in with its culture and practices. I measure this variable as the amount of governing work, and I transform it by adding one and then taking the natural logarithm, so that values of zero indicate no activity of this type.

Please see Figure 13 for a graphical representation of changes to distribution of work efforts on Wikipedia over time, which is consistent with the argument put forth by proponents of the bureaucratization theory of decline. Changes in the relative proportion of work going to different parts of the organization support the notion of a growing emphasis on formalized coordination and governance work as well as a large increase in peer to peer communication around 2005.

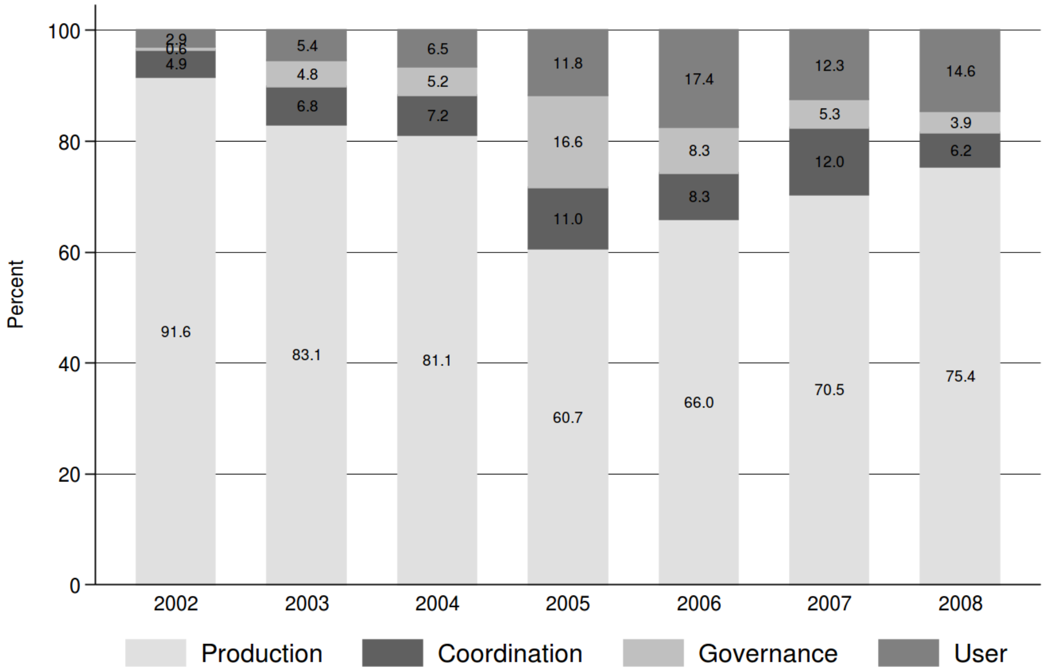


Figure 13: Distribution of work on Wikipedia for contributors in the sample.

Results for retention

The dependent variable is whether a user fails to continue to a second month (fail = 1). I use logistic regression to model this likelihood. I present the results in Table 3. Of the 4748 contributors in my sample, 3336 (70%) did not make it to their second month of contribution. Because we know that retention is lower, on average, over time, each regression equation includes time as a control variable. In Model 1, I include the covariate for the amount of productive work performed by a contributor in the first month. In Models 2-4, I include each of the types of activities (user talk, coordination, and governance) that are associated with bureaucratization in the overall work context of Wikipedia. In Models 5-7, I consider these covariates in pairwise combinations, and Model 8 contains all three predictors.

The results of the logistic regression yields odds ratios, which indicate the increase or decrease in odds of a contributor failing to continue into a second month. Intuitively, it is understood that an odds ratio greater than one increases the probability of failure, whereas an odds ratio less than one decreases the probability of failure. The general transformation from odds to percent change is (b - 1) * 100%. For example, in Model 1, the odds ratio of 1.047 for time indicates that, on average, for each calendar month that passes in Wikipedia's history before a person first starts contributing, there is a 4.7% increased chance of failure within the first month. The odds ratio of 0.581 for productivity can be interpreted as a percent change using the following transformation: (0.58 - 1) * 100, or in other words, a 41.9% decrease in chance failure for each doubling in productivity. The remaining coefficients can be interpreted in a similar manner.

For each additional variable I entered, I performed an incremental F test to determine whether the additional variance explained by the current model was a statistically significant improvement over the more simple model. I also used a likelihood ratio test at each stage. For the most part, these tests justify the modeling strategy I adopted, however they indicate that there is no likely improvement after Model 6.

The odds ratio for user talk in Model 2 is 0.511, suggesting that the more peer-to-peer communication that a contributor engages in during the first month, the lower the odds of failing to be retained until a second month. This is consistent with the notion that Wikipedia is not a social network, and contributors who spend more time talking to others are less likely to continue contributing after their first month.

In Model 3, we find the same direction of the relationship for coordinating activity, and Model 4 maintains this consistent pattern with governing activity. Models 5-7 consider the pairwise inclusion of these covariates, and in all cases, the results are consistent. Only in Model 8 is the coefficient for coordination non-significant, but this is likely due to high multicollinearity among all the predictors in the model. All of these factors are associated with an increase in contributor retention (i.e. lower odds of failure), contrary to expectations from theory. Simply put, the more that a contributor engages in the broad range of activities in peer production early in one's career, the more likely that person will continue contributing to the project past this initial phase.

Not shown in the table, I also considered a number of alternative model specifications to check the robustness of these findings. I tested a series of dummy variables for the year of a contributor's first edit, which yielded substantively the same results. I considered non-linearity for all the predictors, even though there is no good theoretical reason to think that these would exhibit a non-linear relationship with contributor failure in their first month. I checked whether there was a linear or non-linear interaction between time and the type of work performed during the first month. For example, if the organization became too bureaucratized during the period 2005-2006, as evinced in Figure 12, then bureaucratic activities such as coordination or governance may be correlated with higher failures during those years than either earlier or later in the project's history. None of these alternative specifications offered a significant improvement over the more parsimonious models presented here, suggesting that this is a fairly robust predictor of contributor retention.

The results of the analysis of contributor retention appears to be an example of a veridical paradox: contributors who are unable or unwilling to interact with the bureaucratic side of Wikipedia drop out, and their experiences likely form the basis for our belief that bureaucracy leads lower contributor retention. While this may be the case, it also suggests that contributors who do engage with the bureaucratic side of Wikipedia are themselves less likely to drop out. We rarely hear about the experiences of contributors who get over this initial hurdle.

	Model 1	Model 2	Model 3	Model 4
Calendar Time	1.047 *** (14.33)	1.046 *** (13.88)	1.047 *** (14.15)	1.047 *** (14.23)
Productivity (ln)	0.581 *** (-13.59)	0.589 *** (-13.12)	0.575 *** (-13.85)	0.574 *** (-14.01)
User talk (ln)		0.511 *** (-6.91)		
Coordination (ln)			0.700 *** (-3.57)	
Governance (ln)				0.670 *** (-6.27)
N obs.	4748	4748	4748	4748
loglikelihood	-2677.500	-2649.700	-2671.400	-2658.800

	Model 5	Model 6	Model 7	Model 8
Calendar Time	1.046 *** (13.75)	1.046 *** (13.85)	1.047 *** (14.13)	1.046 *** (13.77)
Productivity (ln)	0.584 *** (-13.31)	0.583 *** (-13.49)	0.571 *** (-14.15)	0.580 *** (-13.59)
User talk (ln)	0.522 *** (-6.54)	0.550 *** (-5.88)		0.555 *** (-5.71)
Coordination (ln)	0.756 ** (-2.66)		0.781 * (-2.32)	0.818 (-1.80)
Governance (ln)		0.723 *** (-4.85)	0.690 *** (-5.67)	0.738 *** (-4.44)
N obs.	4748	4748	4748	4748
loglikelihood	-2646.200	-2638.300	-2656.100	-2636.600

Notes: odds ratios with t statistics in parentheses; * p<0.05, ** p<0.01, *** p<0.001

These results are substantively meaningful because they add nuance to the dominant narrative that new users are being turned away from the project due to its increasingly abrasive bureaucracy. This may be true, for some users. In fact, given the fantastic growth in the overall number of people who give Wikipedia a try, this may be the experience of a large number of first-time users. But nonetheless, bureaucracy can also be seen as enabling contributor retention. How can we make sense out of this seeming contradiction? In the case of Wikipedia, there are two numbers to keep in mind. First is the absolute number of new users over time, which grew exponentially for years after the project began. The second number is the rate of contributor retention, which has decline since then as well. Decline theories suggest that a growing bureaucracy (necessary to deal with the first number) causes a decline in contributor retention (the second number). It is true that the absolute number of new users who begin Wikipedia and quit soon thereafter is much larger in later years of the project, when Wikipedia has a more bureaucratized structure in place. There are many more chances, then, for a contributor to encounter an over-zealous bureaucratic who makes the initial volunteering experience unpleasant. However, this evidence suggests that contributors who are more willing to engage in the different roles and social relationships that have became part of Wikipedia's more formalized structure are also the ones who are more likely to continue contributing. On the whole, when new users engage with the community, their retention is higher, even though the absolute number of hostile bureaucratic encounters will increase as well. These have garnered much of the spotlight and have been written about*,* ad nauseum, as the cause of Wikipedia's decline. But the same mechanism also seems to serve a stabilizing force in the community. I consider this possibility in the next section when I analyze bureaucracy's effect on contributor productivity over time.

Analysis of productivity

The results from the first month of the analysis, predicting contributor retention, are not wholly inconsistent with decline theories, because many contributors choose not to engage in the more bureaucratized aspects of the community. By some accounts (Butler et al. 2008) users may choose to avoid such situations because they do perceive them as hostile, coercive, or not the reason why they began volunteering to begin with. Nonetheless, having seen seen opposite results from what was expected – that in their first month of contribution, being active in bureaucratic work contexts increases contributor retention – I continue the analysis by considering how bureaucratic work contexts is related to changes in contributor productivity over time.

Using a longitudinal model, I estimate the productivity of contributor i at time t as a function of time, a series of time-varying covariates, and time-invariant characteristics. I transform both the dependent variable and covariates using the natural logarithm, which yields a 'double-log' model that is common in econometrics and many other social science fields. The coefficients are elasticity estimates that can be interpreted using a convenient rule-of-thumb: a one-percent change in x produces a b-percent change in the dependent variable, holding all other variables constant. Positive coefficients can be understood as driving factors yielding higher productivity. Coefficients larger than one suggest a greater percentage gain in productivity than percent expended on other types of work. Negative coefficients imply factors that hinder productive work.

In general, the log function has the desirable property that it tends to convert multiplicative relationships (such as growth trends) into additive relationships. By using the logarithm of variables that grow exponentially with respect to one another, this relationship appears linear on a log-log scale, which can be estimated using linear models. More conveniently, the difference in logged values between two time points represents the percent change over time.

The results presented here are random effects models. My rationale for using random effects is both methodological and substantive. Hausman tests do not support using fixed effects over random-effects estimation. In addition, since I control for some stable characteristics of individuals, these covariates would drop out of the equation in a classical fixed effects estimator since their variance is zero. The random effects models use a generalized least squares estimator with weighted average of results “between” and “within” effects. In other words, it includes both components of the variance in productivity for contributors over time as well as across different contributors.

Since the model estimates productivity in the current period, the covariates are measured at a one-period lag. There are several reasons for doing so (Cameron and Trivedi 2005; Finkel 1995). First, it increases the ability to make causal inferences since the model incorporates time-ordering of events, which helps rule out reciprocal effects. In a preliminary analysis of this problem (not presented), I tested whether the use of covariates contemporaneous to the dependent variable were different than the lagged terms. They were not. I also directly tested for reverse causality by substituting the dependent for the key independent variables and rerunning the analyses. These yielded non-significant results, suggesting that the models presented in this paper correctly capture the temporal-direction of the influence of work contexts on productivity and not vice versa. However, errors tend to be heteroskedastic, so I report robust standard errors to individual users.

I do not include a lagged term for the dependent variable since this would create a potential issue of endogeneity of the lagged error term. If I had simply including a one-period lagged measure of productivity into a standard model, this could lead to significantly biased estimates. Conceptually, we can see this limitation by realizing that the lagged dependent term in the current period would have been the dependent variable in the preceding period, which clearly makes it correlated with the error term from the prior period (Blundell, Bond, and Windmeijer 2000; Cameron and Trivedi 2005; Woolridge and Wooldridge 2002). Although some sophisticated solutions to this problem have been offered (Blundell and Bond 1998), no regression technique on non-experimental data can avoid all these potential pitfalls. I believe that a theoretical grounded model that maintains careful attention to not violating regression assumptions is the most prudent way approach this situation, rather than attempting to implement advanced techniques which may be difficult to assess how “well behaved” they are in this situation.

Independent variables

I retain the same independent variables used in the logistic regression predicting contributor retention. In this case, the variables are measured each month, over time, for each user (for the variables *user talk*, *coordination*, and *governance*). I use the lagged values from the period t-1, but in exploratory analyses (not presented) the contemporaneous predictors yield substantively the same results. I also control for the time trend using the calendar month, as above.

Dependent variable

Productivity: For the longitudinal analysis of productivity, I take a user’s monthly productive work output as the dependent variable. I add one, then take the natural logarithm of this count, so that values of zero indicate that a contributor did no productive work in a given month. (Please see Table 4 for descriptive statistics for productivity as well as the focal covariates.)

Additional control variables

Duration: As is standard in longitudinal regression models, I consider changes in the dependent variable over time. As a way of incorporating the effects of time in the analysis, I use the notion of a *career* which begin on the date of a user's first contribution to the wiki project and ends when they make their last edit, or the last day of observation, whichever comes first. I then standardize this measure for each user by their overall career length in order to measure how far users' have progressed through their careers in each period. Panciera et al. (2009) note that "nearly all editors begin with a burst of activity, then quickly tail off." For this reason, I also include a mean-centered quadratic term to account a possible non-linear relationship between individuals' productivity and duration into career.

Career date: In addition to contributors' duration into their careers, I also include a linear term for the number of months into a contributor's career tenure. Including this term allows me to control for the effect of careers of shorter or longer lengths. To understand this, consider a scenario with two contributors, one of whom contributes to Wikipedia for 24 months and the other who contributes for 6 months. If they are both half-way through their careers, both of their values on the duration variable will be .50 (i.e. half). On the other hand, their values on career date will be 12 and 3, respectively. By including these two variables together, the model controls for both *position* and *scale* of a contributor's career.

Administrator: Users may nominate themselves to gain administrator status, which is voted upon by other members of the community. Administrators have some technical capabilities, such as being able to temporarily prevent changes to articles to deter vandalism or temporarily block editors who violate the "three revision rule," which ordinary users cannot do. Despite the ostensible bureaucratic powers that administrators may yield, some research (Kittur, Chi, et al. 2007) has found that the significance of the administrator role in the community is modest and has decreased over time. In this view, administrators are described more as janitors, performing mundane clean-up tasks rather than exercising any authority over other contributors. When a user is promoted to administrator, I indicate this using dummy coding, with 1 indicating a user is an administrator and 0 as the reference category.

Core: Because this research is designed to assess if the effects of bureaucratic work contexts vary based on the extent of a users' contribution history, I control users who are core contributors. I use dummy-variables to identify users whose total work output puts them in the top 10% ("core") of Wikipedia contributors, with 1 indicating a core contributor and 0 as the reference category. This remains consistent with the approach in Chapter 2.

Product terms (Core X activity): To complete the list of variables, I compute interaction terms that allow me to consider the different effect of different types of work between core and non-core contributors. This lets me assess a situation that seems particularly relevant for organizations like Wikipedia which combine heterogeneous groups of contributors. I construct the interaction term by multiplying the moderator variable (*core*) by the focal variables (*user talk, coordination,* or *governance*). The product term can be interpreted as the effect of these types of work on core contributors, relative to non-core contributors, whose effect is represented by the linear term.

Table 4: Descriptive statistics of variables in the model

Non-core (n=808)	Mean	St. Dev.	Min.	Max.
Productivity	2.56	1.28	0.00	114.00
User talk	0.25	0.76	0.00	92.00
Coordination	0.28	0.73	0.00	59.00
Governance	0.08	0.38	0.00	28.00
Core (n=144)	**Mean**	**St. Dev.**	**Min.**	**Max.**
Productivity	13.28	3.77	0.00	3119.01
User talk	1.55	3.01	0.00	1235.99
Coordination	1.64	2.56	0.00	905.00
Governance	0.71	2.11	0.00	623.00

Descriptive statistics of contributors activities

The number of contributors (N) in the sample is different than in the previous analysis of retention. Every one of the 4748 contributor in the sample has a “first” month of activity. For this analysis, I only retain contributors who did succeeded to reach a second month of activity (i.e. did not “fail” in their first month), which left me with 1412 contributors. However, a number of these were not suitable for this analysis since their careers started too close to the last date of observation in the dataset. For example, if a user began contributing after June 2008, even though that contributor may be active for two or more months, that user is still screened out of this analysis because I wanted to retain at least six months of data on their contribution histories. (The available data ends at the end of 2008.) For this reason, the number of contributors is N = 952. I follow these users contribution histories over time, yielding 3653 user-months of observation, with an average of 3.8 months of activity for each user. Please see Table 4 for descriptive statistics of the variables in the model. These represent the untransformed monthly edit counts (whereas in the regression models they are transformed using 1 + natural log). It is quite obvious, as would be expected, that core contributors have on average a higher monthly edit count across all domains of activity. I present the results of the longitudinal regression estimates of productivity in Table 5.

Modeling strategy

In exploratory analyses (not presented here), I developed a sufficiently robust baseline model that accounts for both temporal dimension of these data (calendar time and contributor’s career time).[19] These are retained as the control variables across all model specifications. In Model 1, I estimate the effect of different types of

[19] As an additional robustness check, I “de-trended” the temporal aspects of these data, then confirmed all the appropriate distributional assumptions and regression diagnostics were met. Using these de-trended data, I re-ran all the model specifications presented in Table 5, and the results were largely consistent. This suggests to me that the temporal aspects of the data have been dealt with in a sufficient manner.

contributor activities on contributor productivity. This includes the least bureaucratized activities, such as peer-to-peer communication (user talk), to the most bureaucratized activities, such as peer self-governance. In Model 2, I add the quadratic term for a contributor's "duration" into their career, which gives the estimator some flexibility in handling any initial "spike" in contributions that are thought to take place during the earliest stages of a contributor's duration (Panciera et al. 2009). In Models 3-5, by using a product term of type of activity and core contributors, I separately contrast the effects of user talk, coordination, and governance work for core and non-core contributors. Finally, I present a fully saturated equation in Model 6 that includes these three product terms in the same equation.

Models 3-6 attempt to model the differential effect of types of activity on users who are either core or non-core contributors – that is to say, the potential differences between the most active contributors compared to less active ones. Keep in mind that any contributor whose career length did not span at least two months was excluded from this analysis (i.e. the contributors who "failed" in their first month, from the analysis of retention presented earlier). This means that the non-core users included here are still sufficiently engaged and involved in Wikipedia over the course of many months. In other words, the non-core contributors here are not simply curious fellows who pop in for one day, never to be seen again; these are precisely the not-fully-engaged contributors who decline theories are implicitly referring to. By including these two groups in the analysis at the same time, Model 6 actually pushes the methodological appropriateness of this modeling strategy to the breaking point, since it contains multiple product terms which are collinear to one another in the regression equation at the same time.

To overcome this limitation, I performed a confirmatory regression analysis where I split the sample into core and non-core contributors and then re-ran all the models separately for these two groups. This obviously mitigates the issue of needing to include multiple product terms in the same model, since now the regression coefficient for the linear term for types of activity performed simply refer to the effect within each group. Although I did not report these results here due to space considerations, they are substantively identical to the results in Models 1-5, and only differed in Model 6, most likely due to the problem of multicollinearity as I have just described. Therefore, I am confident that the results of the regression analysis that I present here continue to hold true if we looked at these groups of contributors separately. I now turn to a presentation of the findings.

Results for productivity

Table 5 presents the results of my analysis of contributor productivity. I report unstandardized regression coefficients and t statistics, where the simple mathematical operation (b / t) can be used to produce the standard error of each coefficient. To interpret these coefficients, I remind the reader that such a "double log" model yields elasticities or percent change in the dependent variable with respect to percent change in the independent variables. For the non-logged covariates, such as time, the coefficients can be interpreted as a one-unit change yielding a b percent change in the dependent variable.

The very first result from my analysis remains consistent across all model specifications and may appear somewhat surprising: the more time that has elapsed since "calendar time" began in January 2003, the more productive contributors become on average. This is counter-intuitive until we remember that contributors who drop-out within one month are excluded from the analysis, and such drop-outs represent an increasingly large portion of new contributors. We see another consistent time pattern for the career date variable, where overall contributor productivity tends to decline as each contributors "career date" increases, where a value of one represents their month of first edit. Finally, the variable for duration (in Model 1) and the quadratic term for duration (Models 2-6) remain negative and statistically significant throughout all the equations. This suggests that after an initial burst of productivity at the start of one's contributor career, productivity tends to regress to the mean over a long decay.

Next, I turn to the substantively more interesting results from these regression models. Across all model specifications save for Model 3, the coefficient for Administrator is not statistically significant. This suggests that obtaining administrator status in the community and performing that role does not negatively affect contributors' productive work output. (I will return to the statistically significant coefficient in Model 3 in a moment.) We also see that across all models, core contributors have a positive and statistically significant coefficient, confirming what we tautologically know to be the case from the descriptive statistics in Table 4 regarding their higher average productivity.

Finally, let us consider the most theoretically relevant portion of these regression models: how performing work activities that are more or less bureaucratized affect productive output. These focal activities are all measured in the previous time period, which permits these coefficients to be intuitively interpreted using the following heuristic: *performing certain types of activity last month yields what effect on productivity this month.*

In Models 1-2, we see that the coefficients for user talk, coordination, and governance are all positive and statistically significant. This suggests that engaging in these types of activities in the prior month is correlated, on average, with contributors performing more productive work in the present month. On first blush, this would seem to suggest that one of the main thrusts of decline theories is incorrect: contributors continue to fulfill their roles as content producers even more so as they engage in other community roles, rather than the assertion that contributors would become small-minded bureaucrats focusing more on community roles at the expense of the productive activities.

Reading only through Model 2, however, this analysis remains incomplete. As I suggested in Chapter 1's discussion of specification error, we should be careful to look for a potential differential effect between groups that would otherwise remain hidden behind the overall trend when those groups are combined. In other words, it is incumbent upon us to see if Simpson's paradox is lurking in these findings.

I accomplish this in Models 3-5. Please refer to the portions of Models 3-5 in the table that highlight the contrasting effects between core and non-core contributors. I will walk through these results one at a time, since they can be somewhat tricky to interpret. Model 3 highlights the comparison of engaging in peer-to-peer user talk activity between core and non-core contributors. Since core contributors are coded as one, the product term is the effect of such activity for core contributors; consequently, the linear term represents the effect of such activity for non-core contributors. (The dummy variable for core, which is still included in the model, represents the mean difference between these groups net of the other variables in the model.)

Table 5: Longitudinal regression estimates of contributor productivity

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Calendar time	0.009 ***	0.008 **	0.007 *	0.007 *	0.008 **	0.007 *
	(3.29)	(2.76)	(2.44)	(2.48)	(2.79)	(2.31)
Career date	-0.011 ***	-0.014 ***	-0.013 ***	-0.013 ***	-0.014 ***	-0.013 ***
	(-3.30)	(-4.07)	(-3.94)	(-3.88)	(-4.09)	(-3.86)
Duration	-0.367 ***	-0.136	-0.121	-0.163	-0.144	-0.146
	(-4.15)	(-1.25)	(-1.12)	(-1.52)	(-1.32)	(-1.38)
Duration^2		-1.119 ***	-1.064 ***	-1.020 ***	-1.088 ***	-0.995 ***
		(-4.21)	(-4.09)	(-3.91)	(-4.09)	(-3.85)
Administrator (1=yes)	-0.262	-0.276	-0.427 *	-0.210	-0.345	-0.350
	(-1.26)	(-1.30)	(-2.05)	(-1.03)	(-1.61)	(-1.71)
Core (1=yes)	1.056 ***	1.047 ***	0.927 ***	0.897 ***	1.013 ***	0.831 ***
	(13.41)	(13.27)	(11.84)	(11.21)	(12.73)	(10.42)
User talk (t-1)	0.124 **	0.122 **	-0.159 ***	0.107 **	0.113 **	-0.094 *
	(3.13)	(3.11)	(-4.16)	(2.79)	(2.74)	(-2.28)
Coordination (t-1)	0.218 ***	0.213 ***	0.200 ***	-0.104 **	0.205 ***	-0.051
	(5.32)	(5.26)	(4.67)	(-2.64)	(4.84)	(-1.23)
Governance (t-1)	0.234 ***	0.235 ***	0.175 **	0.187 ***	-0.083	0.046
	(4.47)	(4.59)	(3.13)	(3.68)	(-1.20)	(0.65)
Core X User talk (t-1)			0.391 ***			0.279 ***
			(6.55)			(4.18)
Core X Coordination (t-1)				0.425 ***		0.335 ***
				(6.98)		(4.94)
Core X Governance (t-1)					0.367 ***	0.124
					(4.44)	(1.30)
Constant	-3.904 *	-2.991	-2.476	-2.471	-3.046	-2.190
	(-2.44)	(-1.88)	(-1.53)	(-1.55)	(-1.90)	(-1.36)
R^2	0.376	0.379	0.390	0.391	0.382	0.398
N (user-months)	3653	3653	3653	3653	3653	3653
N (users)	952	952	952	952	952	952

Notes: unstandardized coefficients with t statistics in parentheses; * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Model 3 shows the differential effect of user talk activity for core and non-core contributors. The linear term which represents the effect for non-core contributors is negative and statistically significant, while by contrast, the product term which represent the effect for core contributors is positive and statistically significant. Similarly, Model 4 shows the differential effect of coordinating activity between these groups of contributors, where again, we see a negative and statistically significant coefficient for non-core compared to a positive and statistically significant coefficient for core contributors. Intriguingly, in Model 5, governance activity for non-core contributors has no effect on contributor productivity, whereas it maintain a positive and statistically significant effect on productivity for core contributors.

Discussion of decline and/or stability

Stinchcombe (1965) reminds us that many new organizations tend to fail early on. One factor that Stinchcombe identified that predictably leads to failure was when key members of the organization have a difficult time performing unfamiliar roles and fulfilling different types of work relationships. As Wikipedia experienced its meteoric growth in the size of its contributor community over a very short amount of time, it created a problematic situation where the community was compelled to develop and formalize rules, policies, and types of working relationships to manage and coordinate their efforts. Ironically, this growing bureaucracy was thought to be the factor that contributed to a reversal in its growth trend, putting it onto a path of slow decline due to “calcification of rules against newcomers” (Halfaker et al. 2012). I believe that this existing narrative focuses too heavily on only half of the puzzle – that is, newcomers’ negative experiences with bureaucracy – and ignores the potentially enabling consequences of bureaucracy for other members of the community. I suggest that this is due to an overemphasis on the small decline (around 25%) since its peak size around 2007 in contrast to several prior consecutive years of exponential growth that led to an increase in the size of its community by several orders of magnitude.

Much of the focus in organizational studies has centered on the debate around organizations’ “life cycle stages” including periods of growth and organizational decline (Weitzel and Jonsson 1989; Whetten 1980, 1987). One of the most common themes in the literature that I encountered is the potentially negative consequence of size: large organizations can become too rigid, too formal, too inefficient, or too inaccessible. In this view, growth is seen to be beneficial up to a point (Perrow, Wilensky, and Reiss 1986), beyond which growth can be counterproductive. But we should recognize that, in general, bureaucracy can have both positive and negative effects on workers: when workers consider a rule to be good, they rarely notice it; on the other hand, rules that workers consider bad become the source of much dissatisfaction (Perrow, 1983). Thus effects of bureaucracy depend on the organizational contexts that workers find themselves in, as well as the social roles they enact and their structural positions within the organization.

The empirical studies in this chapter broadly support the contours of this argument while also providing a much needed refinement a refinement to this view. The development of an informal bureaucracy, and formalization in the bureaucratic side of Wikipedia’s community, can have both “enabling” and “coercive” effects (Adler and Borys 1996) – it can increase effectiveness, efficiency, productivity, and commitment of workers; or it can circumscribe their autonomy, potentially stifle creativity, innovation, and commitment. Individual-level commitments to volunteering and a contributor’s position within the structure of the organization play a central role in shaping patterns of participation.

Given the large number of research findings in this chapter, it is a good idea to review the implications of these results for the larger inequality story introduced in Chapter 1. Broadly speaking, we continue to see a huge gap between the stable, highly active core and the more peripheral and less integrated population of occasional contributors. This gap may endure and even widen as the core contributor community creates, utilizes, and reconstitutes through their actions the bureaucratic side of Wikipedia to further their work effort. This creates a higher "barrier of entry" to newcomers, requiring more commitment on their part in order to become integrated into the community. Failure to put for this effort can lead to failure to continue volunteering.

However, I find little support for the notion that there is a widespread pattern of small-minded bureaucrats who promulgate rules and policies on newcomers, and instead found that engagement across roles and types of work enhanced the overall rate of production for these core contributors. For example, Aaltonen and Lanzara (2010) suggest that individuals differentiate into specialized behavioral roles, which may be true for a small number of contributors. But if there was an emerging "managerial" class would govern the project and coordinate other users while contributing little productive work on their own. On average, this does not appear to be the case. These results imply that rather than a strict division of labor where users progress through discrete roles, growth in peer production entails users broadening the types of work they perform.

But this illustrates an underlying paradox of peer production; let us not lose track of the fact that all of this work, communication, and coordination is done on a voluntary basis, by contributors who may stop at any time. Without their continued support, the organization would essentially collapse under its own weight of work needing to be done and no one left to do it. The mechanisms that deepen commitment and engagement for some members of the community, hence providing some stability, also constrain the organization's growth. If we think about Wikipedia's growth (Figure 8) as an S-shaped curve, point at which its exponential growth halted and the size of the organization stabilized can be thought of as the "carrying capacity" of the organization with respect to the size of its bureaucracy.

Additionally, the analysis here suggests that the argument put forth by Panciera et al. (2009) may have overlooked some of the important contextual effects that can steer users into different "career paths." For example, it is undoubtedly true that the "free encyclopedia that anyone can edit" in 2001 is not the same as the one in 2007 or 2012, and new contributors' first encounters with the organization at these later dates requires different type and level of commitment. In particular, more engagement in the increasingly formalized work contexts, especially early in a volunteer's career (e.g. during the first month of volunteering) can powerfully shape their propensity to continue.

I believe that this chapter points to the type of future research that needs to be developed on this topic. On the one hand, having such detailed micro-level records whose patterns reveal the macro-level structure in Wikipedia permits the type of analysis here, that is representative of the population. On the other hand, this approach is limited by the inability to understand the meaning of bureaucracy to participants themselves. Perhaps new techniques in computational social science can assist us in being able to tease out the potentially different effect of unwelcome "encounters" with Wikipedia's bureaucracy versus voluntary "engagement" in these roles. While these subjective experiences are possible to discern on a small scale (Butler et al. 2008) using qualitative methods such as surveys, interviews, and even in-depth online ethnography, we have yet to develop a way to scale such findings to the population level in a reliable way.

Chapter 4. Gender, diversity, and quality

> Going on to Wikipedia and trying to edit stuff and getting into fights with dudes makes me too weary to even think about it. I spend enough of my life dealing with pompous men who didn't get the memo that their penises don't automatically make them smarter or more mature than any random woman.

> Even if I don't explicitly identify as female in my Wikipedia handle (and I don't), I still find myself facing attitudes of sexism and gender discrimination, attempts at silencing, 'tone' arguments, and an enforced, hegemonic viewpoint that attempts to erase my gender when editing.[20]

Introduction

Women's historical absences from scientific endeavors and other institutions that create knowledge is also seen in the contemporary practice of online peer production. Recent survey research of contributors to Wikipedia has found that less than 15 percent of its contributors are women. The causes for this high degree of gender inequality have been linked to both self-selection and barriers to participation: we continue to imagine expertise in creating knowledge as a masculine pursuit, and women may buy into this belief as well, undermining their confidence in participating in online knowledge production. However, as I mentioned, the gender gap has been shown to be more than simply self-selection: research suggests that some aspects of online participatory culture limits women's participation because of its excessive conflict or contentiousness, devaluation of certain topics or perspectives, and in some instances, overt hostility to women or other forms of misogyny.

This raises several interesting questions about whether online realms are open to a diverse range of participants and whether they can ever truly represent "disembodied spaces" if participants' socially-learned and embodied gender, and others' perceptions thereof, accompany them into virtual spaces. One way that gender may continue to be salient is through presentation – in particular, whether contributors choose to publicly disclose their gender to others in an online community, which can affect how group members perceive one another and alter group interactions. A second way that gender may be salient is through socially-learned gender performances in terms of the roles people adopt and the types of work they perform.

In this chapter, I use this insight about the many ways of doing gender to problematize the way the gender gap has been investigated by other Wikipedia scholars. I also aim to consider a practical consequence of the

[20] These quotations are from two separate women contributors to Wikipedia, from Gardner (2011).

gender gap that has previously gone unexplored: the relationship between gender diversity in participation and the quality of the work produced by peer production. Organizational research has shown that diversity, in respect to group composition, the types of tasks being performed, and other group dynamics, can affect group performance and shape its collective output. Recent research on Wikipedia (Anthony et al. 2009) found that the quality of work produced by Wikipedia contributors varies with functional diversity (differences in types of work roles performed) and cognitive diversity (differences in the knowledge bases) of group members. However, the potential interaction with gender not been explored in any detail, which is surprising since scholarship on gender differences in online communication and interaction points to the need to incorporate gender more thoroughly into our conceptualization of the problem.

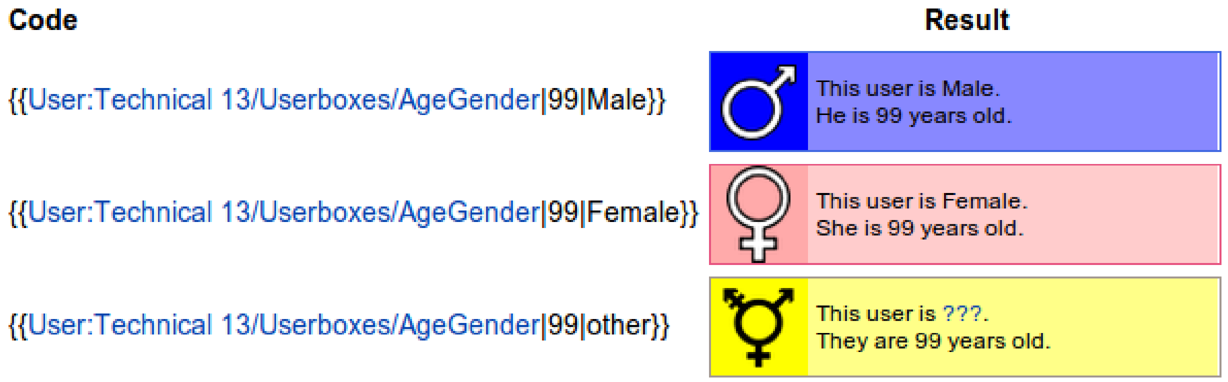
Studying the work of Wikipedia contributors provides an invaluable opportunity to assess how gender diversity, and the different ways of doing gender, affects group performance and shapes the quality of peer produced knowledge. To begin, I critically examine the ways in which Wikipedia's gender gap has been dealt with in the empirical literature.

Establishing the gender gap

It is imperative that we define what we mean by a "gender gap" in participation and establish the empirical foundations of this phenomenon. The two main sources of data for our understanding of the gender gap come from (1) surveys of Wikipedia contributors and (2) from analyses of the work actually being performed on Wikipedia. Both lines of scholarship on this topic establish that there is a gender gap in participation. My own research, that I present later in this chapter, is consistent with this as well. My goal is not to challenge the veracity of the broad claim, but to problematize the state of our knowledge of the gender gap by digging beneath the surface of how we conceptualize gender in an organization that exists online.

There have been two major surveys that establish the gender gap in participation – that is to say, involvement in the actual production of it's content, not just use of the service. The first survey took place in late 2008, which was a collaboration between the Wikimedia Foundation and the UN University at Maastrict (Glott et al. 2010). This survey of over 170,000 contributors was the first major study to report on the gender gap. Women were found to constitute only 13% of the contributor community. In 2011, the Wikimedia Foundation conducted another survey which estimated women's share of participation at 14% (Wikimedia Foundation (WMF), 2011).

Results from both surveys puts the percent of women contributors at under 15% (see Figure 15 for details). These surveys also asked contributors about "gendered" experiences on Wikipedia, which echo the experiences of women contributors featured in other qualitative research (Collier and Bear 2012b; Lam et al. 2011), as well as prominent newspaper articles on this issue (Cohen 2011), countless blogs, social media postings, and Wikipedia community discussions. (The two quotations that I used to start this chapter are taken from such sources.)



Code	Result
{{User:Technical 13/Userboxes/AgeGender|99|Male}}	This user is Male. He is 99 years old.
{{User:Technical 13/Userboxes/AgeGender|99|Female}}	This user is Female. She is 99 years old.
{{User:Technical 13/Userboxes/AgeGender|99|other}}	This user is ???. They are 99 years old.

Figure 14: Templates for Wikipedia contributors to disclose their gender on their user page. By using the code on the left, the associated image on the right will appear on their user page. The sex/gender distinction seems to be confused, with the code using the term “gender” whereas the image uses the male or female “sex” (as well as permitting users to classify their gender as “other.”

One major criticism of these surveys is that their results may be biased due to systematic survey non-response rates. Hill & Shaw (2013) attempt to correct for this bias by combining survey data of *contributors* with additional data from a nationally-representative phone sample of Internet users, including Wikipedia *readers*. Using a propensity score matching technique, they estimated that women’s share of participation (from survey research) may be too low by about 25%, which can be attributed to women respondents being underrepresented in the original survey (relative to their true participation rate in the Wikipedia population) or less likely to complete the survey questionnaire. Nonetheless, their corrected estimate of the true percentage of women contributors still constitutes “only” about 16% of all editors. Hill & Shaw’s work demonstrates the difficulty inherent in obtaining representativeness from a sample of contributors to online organizations.

In addition to survey research, a second source of data on the gender gap from the technical literature uses digital trace data, such as information taken from a user’s profile page or other records created by Wikipedians as they go about their work (Antin et al. 2011; Lim and Kwon 2010). Users can choose to disclose some information about themselves, such as what language they speak, what topics they work on, other details. Some choose to self-identify their gender, either by stating it outright or through the use of “userbox templates” (see Figure 14 for an example). In addition to contributors’ disclosure of gender on their user profile pages, contributors can also set their gender in their public profile in another way. The Wikipedia software (MediaWiki) allows users to set some personal preferences with the software – with the caveat that the choices were Male and Female. For scholars who recognize that gender is a social construct, these categories (limited by the software) are a revealing case in point of Lessig’s mantra that “code is law.” Nevertheless, the point of this line of scholarship is to understand similarities and differences in how men and women contribute to the project. For example, a major line of inquiry focuses on the causes of women’s relatively low participation and whether that leads to a gender bias in topics covered ((Reagle and Rhue 2011; Reagle 2012). Other topics included the use of gendered language (Thomson 2006), how the community handles conflict (Collier and Bear 2012a), and how to make Wikipedia a more welcome place for newcomers who may not be familiar with the traditionally “masculine” computer culture (Morgan et al. 2013).

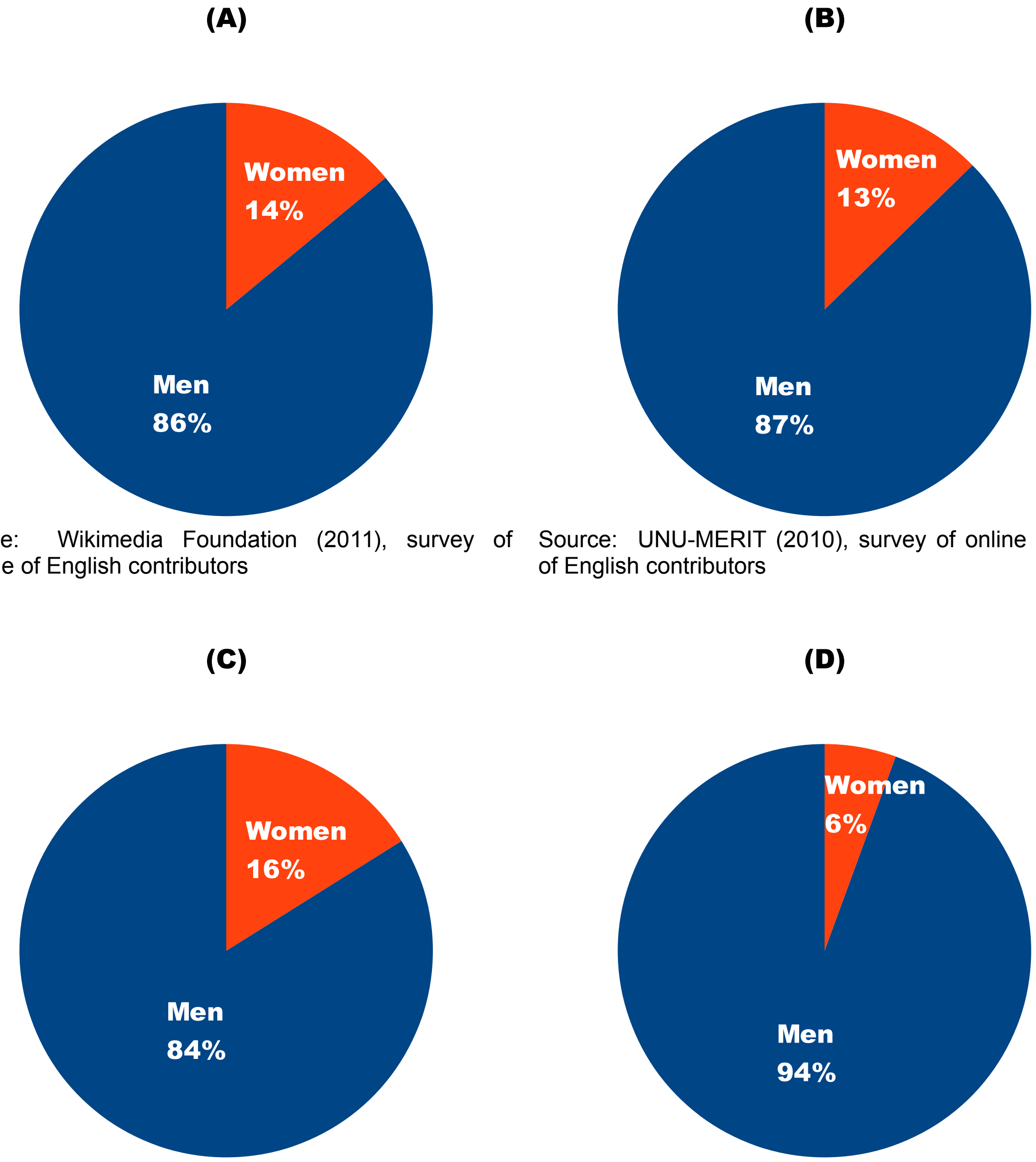
During the process of conducing a research project on Wikipedia contributors, researchers typically use the API (which can be used to retrieve data from the Wikipedia database) to obtain a random-sample of users. They then try to download data on the gender identity of users in this sample by reading the user profile setting or scraping the user's profile page for disclosure of gender. While it is fairly trivial to write the computer code to accomplish this task (I did it for this research), the treatment of missing data is usually brief. Here is fairly representative example from Antin et al. (2011) in their study of "gender differences" in editing:

> The base population from which we draw our sample consists of 256,190 users who created a valid new account on the English-language Wikipedia between September 9th, 2010 and February 14th, 2011. Our analysis requires that we can determine the gender of each user. As a result, we limited our study to the 13,598 users (18.8%) who optionally declared a gender in their Wikipedia profile. Of these, 11,194 (82%) were men, and 2,402 (18%) were women. We have no way to verify users' gender, and it is likely that some accounts were used by more than one person of multiple genders. However, we have no reason to expect that such reporting errors systematically differ by gender.

What I find interesting about this passage is that the authors demonstrate to the reader that they thought about the potential for there to be gender differences in *disclosing* one's gender. But at the same time, they failed to consider whether there might be gender differences in *editing* (the ostensible purpose of their paper) between the approximately 80% of contributors who choose to not disclose their gender and the 20% who do. In the technical research literature, the most common way of handling missing data on gender is to simply throw away data on this 80% of contributors.

I believe at this point it should be obvious the serious methodological limitations associated with both sources of gender data: surveys can be biased due to the difficulty of obtaining representative samples, whereas digital trace data can be biased due to a large percentage of missing data.

In my own research on the gender gap (Figure 15, Panel D; and Figure 16), I encountered both problems: first, using the digital trace data, I encountered a very high rate of gender non-disclosure – higher than most sources, in fact. I partly attribute this to my focus on Wikipedia's highest quality articles (e.g. "Featured Articles"), whose contributors may not perfectly represent the broader community of contributors. Second, I attempted to fill-in some of the missing data by contacting contributors to ask them to privately disclose their gender to me for my research. However, my preliminary efforts in this regard were largely unsuccessful, in that only a very small number of contributors I contacted replied to me with this additional information. My experience brought to light the difficulty of addressing this issue, and the research I present in this chapter offers only a partial and not altogether satisfying solution.



Source: Wikimedia Foundation (2011), survey of sample of English contributors

Source: UNU-MERIT (2010), survey of online sample of English contributors

Source: Hill & Shaw (2013), estimated correction for survey non-response bias to UNU-MERIT (2010)

Source: Restivo (2014), collected from self-disclosed gender of editors to Featured Article candidates in December 2010

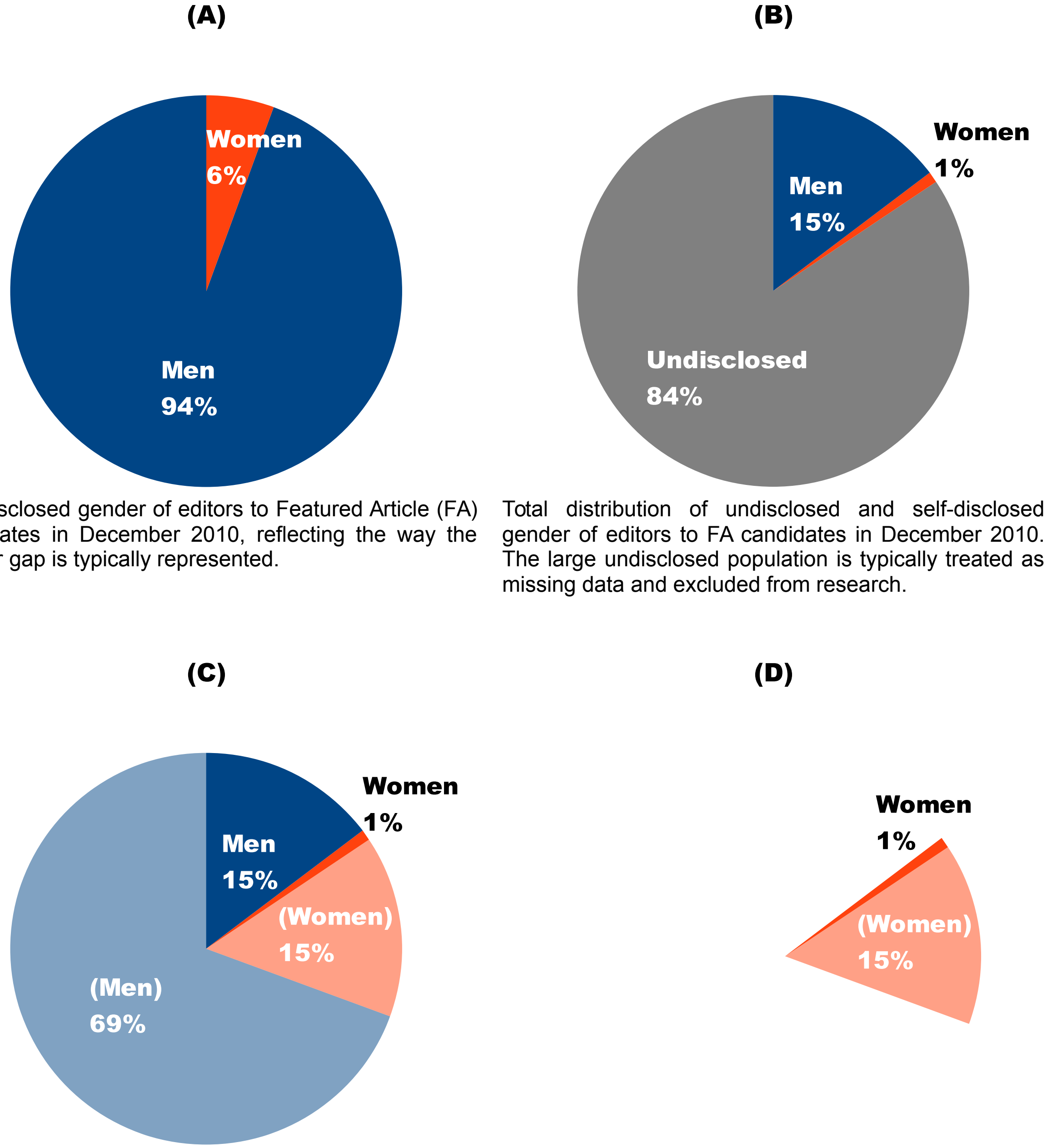
Figure 15: Gender gap in participation. Panel (A) and Panel (B) represent data drawn from two surveys of Wikipedia contributors. Panel (C) includes an adjustment for non-response bias. Panel (D) is self-disclosed gender among contributors to Featured Article candidates in December 2010 that I collected for this dissertation.

It is worth taking a closer look at Figure 16, in which I highlight the extent of the problem of understanding the true gender gap in Wikipedia. I analyze the gender distribution of contributors to articles nominated to become a Featured Article from December 2010. In Panel A, I recreate our existing "knowledge" of the gender gap by relying solely on publicly available data. In Panel B, I graph the true but hidden gender gap – the 16% of contributors who publicly disclose their gender compared to the remaining 86% of contributors who did not.

For Panel C, I provide a statistical imputation of the gender of these undisclosed contributors using mean imputation, which reveals the "hidden" gender gap in participation. Finally, in Panel D, I visually represent what I term the "invisibility" of women's participation in Wikipedia. Because only 6% of contributors whose gender is publicly disclosed are women, in actuality, the visibility of women on Wikipedia can be calculated as 6% x 16% = 1% of total contributors. That is to say, only 1% of contributors to these articles are "seen" to be women, even though their true numbers are far higher. While most attention has been focused on the overall share of women's participation, by taking into account these two different types of participation – gender publicly visible or hidden – I instead focus on an understudied dimension of the problem with sociological relevance. In this chapter, I develop an argument that this constitutes a second gender gap is related to the multiple ways of "doing gender" (West and Zimmerman 1987), which has not been theorized in a sufficient way in this online context (Herring 2003; Sussman and Tyson 2000).

So far I have presented evidence that women represent not just a small share of the total contributor community, but that they are also less likely to publicly disclose their gender. However, I found that this rate of gender disclosure is not the same across Wikipedians of different contribution levels. In Figure 17, I plot the proportion of contributors who disclose their gender, across different amounts of work performed. As the dashed line indicates, a majority of contributors – across all different levels – do not disclose their gender. However, we see there is a steady decline in this rate. (Please note that the x-axis is on a logarithmic scale.) When we compare the lines for 'male' and 'female' editors (recall that Wikipedia's API only gives users these two options), we see a stark contrast. Higher contribution levels invariably have a higher proportion of editors who disclose as men, whereas across all contribution levels, only around 1-2% contributors disclose as women.

One plausible interpretation is that men constitute a larger share of editors at the highest contribution levels, and since men are more likely to disclose, this accounts for the large gap that we see. However, this is only the case the highest contribution level (>100,000 edits), where men were found to be overrepresented. Interestingly, women were found to be overrepresented at the bottom category (<10 edits). But for contribution levels in between these two extremes, women's share is believed to be fairly constant at the proportion discussed above (around 15%). In other words, at nearly every level of contribution, women's true share of participation is around 15%, whereas their public disclosure is around 1-2%.



Self-disclosed gender of editors to Featured Article (FA) candidates in December 2010, reflecting the way the gender gap is typically represented.

Total distribution of undisclosed and self-disclosed gender of editors to FA candidates in December 2010. The large undisclosed population is typically treated as missing data and excluded from research.

Statistical imputation of gender of contributors who do not publicly disclose their gender, revealing the “hidden” gender gap on Wikipedia.

Highlighting invisibility of women’s participation. While women represent (at best estimates) around 16% of contributors, their visibility on Wikipedia is merely 1%.

Figure 16: Reconstructing a more accurate representation of the gender gap in participation. In this figure, data for all four panels come from Restivo (2014) information collected on editors to Featured Article candidates from December 2010. Panel (A) shows the distribution of gender among contributors who self-disclosed. Panel (B) represents gender data for all contributors, including the vast majority (84%) of contributors who do not disclose their gender. Panel (C) includes a statistical imputation of the gender composition of “undisclosed” contributors, represented by the lightly shaded regions. Panel (D) highlights the invisibility of women’s participation.

The two quotations I used to start this chapter highlight some of the reasons why women may choose to not disclose their gender publicly online. Let us take a step back and think about how women's choices to not let other contributors know that they are women can alter one's overall perception of the Wikipedia community. Women's participation – across different levels of contribution – is consistently lower than men's participation. But in addition, men are *seen* doing more work; and in particularly, the most prolific male editors are more likely to show to others in the community that they are men, multiplying their visibility. As a consequence, from the perspective of people who are involved with the Wikipedia community, women become virtually invisible. My research is the first to unveil this hidden form of gender inequality in Wikipedia, but it has been a concern of the community for some time. For example, the Wikipedia community has organized collective efforts to correct this invisibility through events such as the Women Edit-a-Thon (Hern 2014). In doing so, they are simultaneously addressing three connected issues: the first is women's underrepresentation in overall participation; the second is the lack of visibility of women contributors; and the third is the accompanying bias in topical coverage of Wikipedia's articles, where the articles on women scientists, athletes, politicians, and historical figures (among others topics) are less comprehensive, less complete, or missing altogether, relative to their male counterparts.

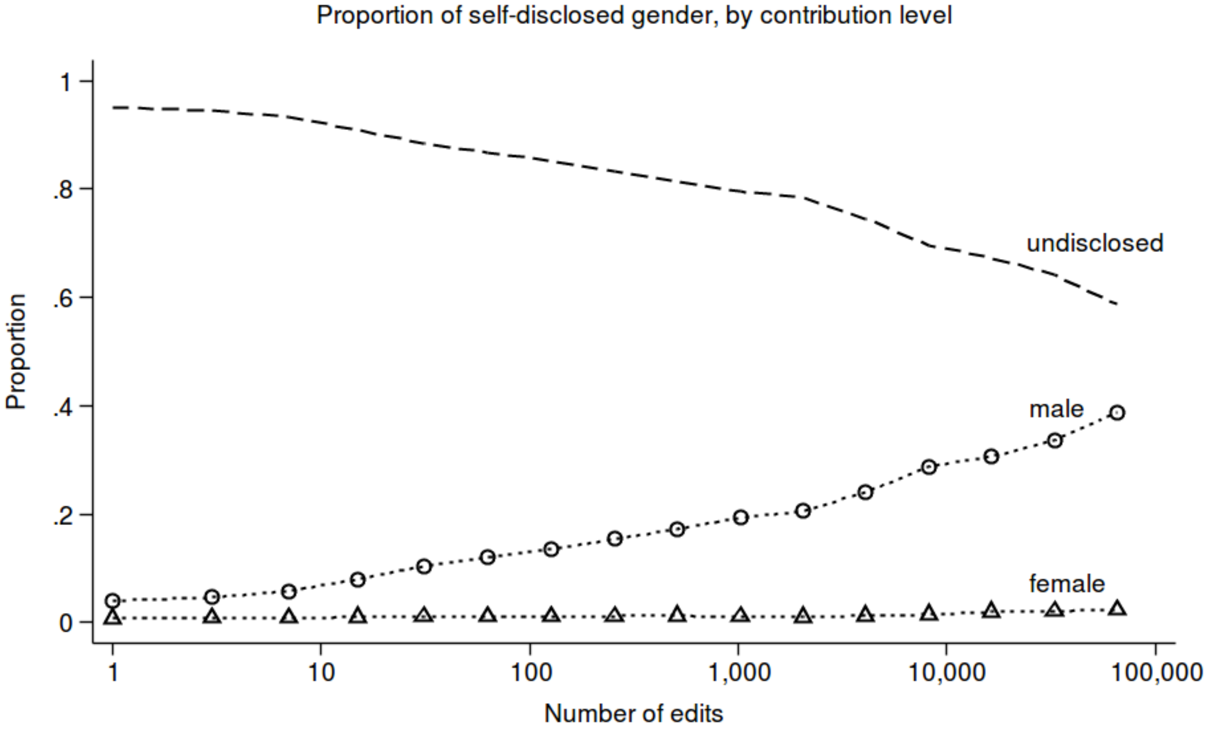


Figure 17: Self-disclosed gender by contribution levels. Data come from user:DaB and user:Dispenser from the Wikimedia Toolserver API.

Much of the discussion of the gender gap in participation has focused on these types of consequences, and ways of rectifying them. The solutions are not entirely straightforward: for example, consider the article on "20th Century Artists." This article listed many prominent artists, but excluded many important artists who were women. Attempts to include the women's contribution in the artistic field to this article were met with some questioning of whether these women artists were truly notable enough to warrant inclusion in this article. Consequently, a second article was created, "20th Century Women Artists" where women's contributions could be celebrated and written about in depth, to correct the historical record where their accomplishments were undervalued in mainstream scholarship in this field. However, this effort came under some criticism as well, since by creating a separate article for women artists, this promoted the impression that women artists were not sufficiently notable to be included in the other, "real" article on 20th century artists. Quite a lot of words have been written about situations like this one and how they may arise due to the lack of diversity among Wikipedia's contributors.

Contributor diversity

Organizational research emphasizes that diversity in a group can affect the group's performance and shape its collective output (Hong and Page 2004; van Knippenberg and Schippers 2007; Luan, Katsikopoulos, and Reimer 2012; Williams and O'Reilly 1998). Overall the literature is broadly supportive of the notion that increased diversity can lead to better outcomes (van Knippenberg and Schippers 2007; Page 2007, 2008, 2010), but as with most social phenomena, there are enough counter-examples to realize that there is no general social law of diversity; rather, diversity's effects differ depending on how "diversity" is defined in respect to group composition, the types of tasks being performed by the group, and other group dynamics.

Given the open nature of collaborative work, such as Wikipedia, where anyone may participate either anonymously or through pseudonyms, reliable data on who contributors are on salient demographic characteristics and in terms of what credentials or expertise they possess is not readily available. As I reviewed above, survey research suggests that only about 15 percent of all editors are women. While there has been much focus on understanding the causes for this high degree of gender inequality – which has been linked to both self-selection and barriers to participation (Antin et al. 2011; Collier and Bear 2012a; Herring 2003; Lam et al. 2011) – considerably less attention has been given to the potential consequences stemming from the lack of gender diversity among its contributors. When the potentially negative consequences have been written about, they primarily focus on topic coverage, as my previous example showed. However, it is surprising to me that scholars have yet to take a broader perspective on how the lack of gender diversity may have negative consequences, since research on group processes suggests that the gender composition of a group (as a type of diversity) can influence its overall performance (van Knippenberg and Schippers 2007; Rogelberg and Rumery 1996). So I arrive at the central question of this chapter: does this gender gap in Wikipedia affect the quality of articles being created?

Wikipedia's articles are authored in small, self-organized groups of contributors who work together to write the encyclopedia's contents (Iba et al. 2010; Spek et al. 2006). Ideally, as they perform this work, useful material is retained, imperfections revised, and errata removed. However, the outcome of this process yields information that varies quite significantly in quality. Some articles suffer from being very poorly written, while in many instances, the accuracy and reliability of its articles matches or exceeds that found in traditional, expert-written sources like the Encyclopedia Britannica (Giles 2005). What accounts for these vast differences in collective output?

Given the variation in outcomes from the collaborative authoring process, recent research (Arazy and Kopak 2011; Arazy and Nov 2010) has begun to explore the connection between group composition and the quality of the information produced. I build upon this research and offer several improvements, both methodologically and theoretically. First, I focus on the consequences of gender diversity. The omission of gender from previous analysis of Wikipedia's group dynamics is a significant oversight. Two other forms of diversity have been found to lead to higher-quality Wikipedia articles: functional diversity (type of work roles performed by group members) and cognitive diversity (differences in the knowledge bases of group members). To my knowledge, no prior studies have systematically tied the gender composition of Wikipedia contributors to the quality of the work being produced, and few prior studies (with Arazy et al. being a notable exception) have directly tested the effects of diversity in general on task performance in peer production.

There is a small literature that deals with the question of what shapes the quality of Wikipedia's articles (Arazy and Kopak 2011; Blumenstock 2008; Giles 2005; Kubiszewski, Noordewier, and Costanza 2011; Twidale et al. 2008). A significant limitation common to much of this research is the way "information quality" is operationalized. One way of measuring quality is through Wikipedia's own article assessment project, where volunteers rate the accuracy and completeness of an article on a scale ranging from short, incomplete articles (called "Stubs") to those that are polished, complete and of the highest-quality (called "Featured Articles") as judged by the community of peer evaluators. A second way that quality has been measured is by independent assessments of Wikipedia's articles, typically conducted by a panel of expert judges. In both instances, it should not come as a surprise that some group-level characteristics are correlated with articles rated as lowest or highest quality, given the vast substantive differences between these articles.

Let me explain this point further. We should not be surprised if a *homogeneous group* of molecular biologists working at Cold Spring Harbor Laboratories produced "better" research than a *diverse group* third-grade student's science fair project. This is analogous to the qualitative differences between Wikipedia's best and worst content. Relying a such broad measures of information quality may obscure the more theoretically-meaningful consequences that arise from group differences, since what we should be focusing on is whether increased diversity within a group of molecular biologist (or within a group of students) yields better outcomes – and not so much across these two groups. The question for Wikipedia is not whether evaluators can discern good from bad articles, but more to the point, what group characteristics may make the difference between a good article and a great one.

I adopt a conservative approach to measuring "quality." Instead of beginning with a random sample of articles, and then measuring the quality of those articles, I focus instead only on Wikipedia's highest quality articles. My sample includes only those articles that have been nominated to become a "Featured Article." These represent articles that are at a mature stage of development and are considered the best content in the encyclopedia. Nonetheless, only a portion of the nominated articles, which have already surpassed a high threshold of information quality in order to become nominated, are successfully "promoted" to become a Featured Article. Since my sample of articles has less variance in quality between them, for this reason, my analysis may yield better insight into what group configurations perform best at achieving this milestone.

A note on the term quality: for this research, I am adopting the Wikipedia community's judgment of the quality of their best articles (Featured Article candidates). This relies on the somewhat problematic assumption that the measure of quality of an article, as determined by other Wikipedians, has any *face* validity. To most readers, it may not. But it should give us pause to reflect on why we believe that it does not. There is no denying that Wikipedia's Featured Articles have *construct* validity as a measure of quality – in this sense, whether an article is promoted to become a Featured Article is a valid measurement of the process by which the Wikipedia community determines what it considers to be its best work.[21] The question of what Wikipedia's own quality assessments are actually measuring, while explored elsewhere in some depth (Blumenstock 2008; Stvilia and Twidale 2008; Twidale et al. 2008), remains outside the scope of this project.

Using Wikipedia as a research site is advantageous for testing theories about diversity's effects. The software that runs the collaborative work platform records all actions performed by contributors, and this database is made publicly available for researchers to analyze. Methodologically speaking, having access to comprehensive micro-level data on patterns of interactions and communications among group members permits me to stay empirically grounded in the actual practices that contributors engage in while writing and constructing these articles, while at the same time permitting me to compare across multiple groups who are busy collaborating in a naturalistic setting.

The population of Featured Article candidates includes over 8000 articles, each representing a separate instance of group collaboration. Even a small sample of these articles represents an enormous research opportunity. A recent study of information quality in Wikipedia (Arazy and Kopak 2011) had a sample size of 96 articles. It is not unusual for research from the management and organizations literature to be based on the study 50-100 groups (Pelled, Eisenhardt, and Xin 1999; Rogelberg and Rumery 1996). However, such groups are often observed in laboratory settings, whereas on Wikipedia, we are observing group behavior where it is actually occurring.

Wikipedia is also advantageous to study in that participation is voluntary and groups of contributors are self-organized. However, as I discuss later in this chapter, this last point also poses a significant limitation to making generalizations from my findings, since it is difficult to rule out the possibility that self-selection accounts for a significant share of the differences between groups. Nonetheless, groups of Wikipedia collaborators do not suffer from many of the confounding factors that exist within corporate 'virtual' teams or traditional knowledge management systems that can affect group dynamics (Arazy et al. 2010). Similarly, survey research which derives from self-reports may be of questionable reliability, while laboratory studies of group collaboration can suffer from low external validity.

Theoretical frame

Group tasks

Analyzing group problem-solving entails four basic constructs: the type of task at hand, composition of the group, patterns of interactions among group members, and the collective output of groups. Steiner (1966) outlines several types of joint tasks that groups can perform. The first are additive tasks where group product is the sum of independent parts, requiring no social interaction among participants or collaboration on the part of a group; rather, separate contributions are combined to form a shared resource or output. Steiner also describes compensatory tasks, where the group product is the average of its member's guesses or estimations. This is the type of task envisioned by the "wisdom of the crowd" argument put forth by Surowiecki (2005) and popularized by Anderson (2006), Sunstein (2006) and others. Some aspects of peer

[21] Members of the American Sociological Association vote each year to award one student the honor of "Best PhD Dissertation." Similarly, members of the Academy of Motion Picture Arts and Sciences vote each year to award one movie the honor of "Best Picture of the Year." Nonetheless, we should not uncritically accept that these truly represent each year's best dissertation or movie. I contend that Wikipedia's "Featured Articles" represent an analogous situation.

production rely on compensatory tasks, where group members do not necessarily interact but are working together on a shared goal. For example, Wikipedia has been experimenting with a system whereby readers can rate the completeness or accuracy of an article. These multiple, independent assessments are combined to create an average rating or metric. No group deliberation or interaction is required for either additive or compensatory tasks.

Social interaction among group members does become central in three other types of group tasks. Conjunctive tasks require all members to succeed for the group to succeed; by contrast, if one group member succeeds at a disjunctive task, that member's success wins it for the entire group. Neither of these types of tasks are common in Wikipedia's peer production. The type of group work most salient for research on peer produced knowledge is a complementary task, which the literature on group processes has devoted most of its attention. Complementary tasks require group members to combine different skills, knowledge, abilities, and other resources to work together on a shared goal. Tasks of this sort require coordination among group members to produce an end product that is greater than what any member could produce alone. Given the collaborative nature of Wikipedia, it seems most appropriate to focus on the performance of groups engaged in solving complementary tasks.

Diversity within groups

Organizational theory stresses that the composition of groups engaged in collective problem solving is a factor that determines success in their collective endeavor. Group diversity refers to the composition of individuals in terms of characteristics that are meaningful to the relationships among group members (DiTomaso, Post, and Parks-Yancy 2007). I focus specifically on gender diversity and how it pertains to two other dimensions of diversity within a group: functional diversity and cognitive diversity. The overarching aim is to offer causal and meaningfully-adequate explanations for why particular configurations of groups produce the highest-quality encyclopedia articles.

The central analytic construct in my research is "diversity," and there is an extensive literature on group processes that explores the differing effects of diversity on collective outcomes. For example, group members who are in the minority on a salient demographic or functional characteristic can introduce new knowledge or approaches to solving a problem, but if group differences prevent their input from being heard, their contributions may not yield appreciable benefits for the group. Similarly, highly diverse groups tend to have greater intra-group conflict, less efficient communication, and lower trust among group members (DiTomaso et al. 2007) which can lead to lower overall group performance. Diversity, however, also been shown to benefit groups engaged in performing complex cognitive tasks, since these groups have access to a broader range of information sources, perspectives on the problem at hand, and access to potentially innovative solutions. While diversity can increase group conflict, this can also lead to higher quality outcomes if differing viewpoints can be reconciled (Pelled et al. 1999). This is because of the positive role of "creative friction" (Leonard-Barton and Leonard 1998) where differences within groups can bring into focus and help group members to clarify substantive disagreements, potentially yielding better solutions to their shared problems (Williams and Cothrel 2000). Because of these potentially contrasting effects stemming from group diversity, next I consider what prior research can inform us about diversity's effects in Wikipedia's editing community.

Gender diversity

Contributors to Wikipedia are not required to disclose their identities or other information about themselves, although some users choose to do so. My aim in this research was to combine publicly-available information about contributors with privately-obtained information that I solicited by email. In particular, I asked contributors to privately disclose their gender for my analysis. As such, my analysis benefits from valuable, hidden information about how gender can affect the collaborative editing process. One way gender may be salient is through presentation – whether contributors disclose their gender to others in the group, which can affect how members perceive one another and alter intra-group interactions. A second way that gender may be salient is through socially-learned gendered performance in terms of the role and style of contributions that individuals adopt for themselves. Given that I have access to information that is otherwise hidden from group members, my research provides an invaluable opportunity to study the effects of these contrasting ways of "doing gender" (West and Zimmerman 1987).

While gender diversity (as a demographic characteristic) is conceptually the most straightforward and ostensibly the easiest to measure, a serious challenge arises in the case of online collaboration. Reliable information about who contributors are in terms of their demographic characteristics is generally unavailable, since contributing can be done anonymously or, as is more frequently the case, through the use of a self-chosen pseudonym. As I discussed at the start of this chapter, editors do sometimes self-disclose their gender on their profiles pages or elsewhere, but for contributors who do not self-disclose their gender, I contacted them by email to ask them to fill out a confidential form asking about demographic information, including gender. By combining data about contributors' self-disclosed gender from their user page with my privately obtained data, I am able to construct two measures of diversity in the gender composition of a group.

Few studies explore the role of gender diversity in Wikipedia's collaborative process, and those that do rely solely on contributors' self-disclosed gender. While contributors' publicly disclosed gender is important because it can affect how others interact with them, these data are insufficient because as we have seen, only a small portion of editors choose to disclose their gender. Any analysis that relies solely on these data may be biased because missing data may be correlated with gender. My own research is no exception to this limitation. But by asking participants to privately disclose their gender for my analysis, my research aims to tease out the differences that may arise when group members can observe the gender of other actors in their group compared to when this remains unknown to others in the group. Although I am adopting Wikipedia's binary conception of gender (male/female) out of necessity, my conceptualization adds a second dimension to this categorization: whether a contributor publicly self-discloses or keeps one's gender private. (A possible fifth response, "publicly and privately undisclosed," I take to represent pure random variation in my study.)

The fact that women represent such a small share of the Wikipedia contributor community is surprising, since women regularly use Wikipedia in high proportion. Taniguchi (2006) suggests that in the broader society, women do more volunteering than men, and these differences may be explained by several underlying social factors: women, on average, exhibit higher degrees of altruism, place a higher value on helping others, and view volunteering as an important part of their social roles (Wilson 2000). However, women's historical absence from scientific endeavors (Schiebinger 1999) may be paralleled in peer produced knowledge online: we continue to imagine the "expert" as a male, and women may even buy into this belief, undermining their confidence in producing knowledge. The practice of knowledge creation has long been viewed as a masculine pursuit, and consequently women may be less likely to publicly reveal their gender to others online and instead choose to participate anonymously.

The gender gap in Wikipedia, however, has been shown to be more than simply self-selection. Notably, Lam et al. (2011) suggest that Wikipedia's culture may be resistant to female participation. The reasons they identify include frequent conflict or contentiousness between editors, choice of topics that are considered valuable, and in some instances, overt hostility. While *use* of Wikipedia is approximately equivalent among men and women Internet users (Lim and Kwon 2010; Zickuhr and Rainie 2010), contributions to the project remain highly unequal, consistent with Lam et al.'s propositions. This raises the interesting question of whether online realms truly represent “disembodied spaces” if contributors’ socially-learned and embodied gender, and other’s perceptions of it, accompanies them into virtual spaces (Nowak and Rauh 2005, 2008; Rosier and Pearce 2011). Reagle (2012) traces the origin of the gender gap in Wikipedia to several features of the culture of peer production, including the predominance of “geek” identities among early adopters, which is highly gendered; and the notion that because participation is nominally open to all, concerns about a gender imbalance can be attributed to matters of individual preference and choice rather to any broader cultural underpinnings of the problem. On this point, I urge the reader to go back and re-read the two quotations that I used to open this chapter. However important gender is as a dimension of diversity in peer production, there are other important types of diversity to consider as well.

Functional diversity

Online contributors often perform different roles in their collaborative work (Wexler 2011). Some users tend to favor community-oriented administrative and coordination work. Community-oriented group members play important roles to mediate conflict when it arises, facilitate the use of efficient procedures during the production process, act in leadership roles toward less experienced contributors, help focus and coordinate disparate individuals to work on collective tasks, and enforce norms related to quality control (Bryant et al. 2005; Kittur et al. 2009a; Reagle 2007; Stvilia and Twidale 2008). Thus, members of groups can vary on a functional dimension.

Successful mediation of conflict within groups that can arise out of group differences has been seen as an important factor that can boost the likelihood of group success (Hoffman and Maier 1961; van Knippenberg and Schippers 2007; Pelled et al. 1999). Conflict, when unresolved, can impede group collaboration (Saavedra, Earley, and Vandyne 1993), but lack of conflict within a group stemming from group homogeneity can also have negative consequences for group problem-solving (Jehn and Mannix 2001; Jehn 1995). Manageable intra-group conflict can expose members to a number of perspectives, foster a deeper understanding of the problem at hand, and lead to better solutions (Amason 1996; Pelled et al. 1999; Woodman, Sawyer, and Griffin 1993). This type of “creative friction” is a central aspect of the collaborative editing process (Kittur, Suh, et al. 2007).

Researchers looking for gender differences in Wikipedia’s community have focused on these differing roles. For example, editors may choose to explicitly coordinate with one another through the available communication channels, or they may favor implicit coordination (Collier and Bear 2012a) or leadership without overt discussion (Reagle 2007). These approaches have been shown to vary by gender (Antin et al. 2011); however, since we know that a majority of contributors do not disclose their gender, the true relationship between gender and functional role is not well understood. For example, the norm to “talk before you type” (Viégas et al. 2007) in order to coordinate work with others may be confounded by the observable gender composition of the group. Further study of this facet of collaboration is important to be able to discern how gender performance and hidden dimensions of gender diversity shape a group’s overall performance.

Cognitive diversity

Among Wikipedia contributors, some tend to engage in less coordinated activity and instead focuses on simply adding to an article’s content in their own ways. Contributors of this sort, who I label as “content-oriented,” tend to make highly reliable contributions to an article (Anthony et al. 2009), but alone may not possess sufficient knowledge to cover a topic in its entirety. Because contributors posses different sets of knowledge, experience, and mental models of the task at hand and strategies for successfully accomplishing their shared goals (Lee and Cole 2003), a certain degree of cognitive diversity may be necessary to successfully write a feature-rich and complete article. Measuring the cognitive diversity of a group, however, is not straightforward.

In the same way that we lack reliable information about demographic characteristics of contributors, we also do not have access to any objective measure of contributors' expertise, credentials, or sets of knowledge that they possess. Because we have no certainty regarding what contributors actually know, we can only consider what they *claim* to know by examining their patterns of contributions. For example, if a person adds a paragraph to an article on science, that person is making a claim that she knows something about science. By examining patterns of collaboration, we can find authors who claim knowledge about shared subject areas. Prior research has demonstrated the value of cognitive diversity within group problem-solving (Brown et al. 2005; West and Dellana 2009). However, in the case of peer production, cognitive characteristics of contributors, such as what they actually know, is an unmeasurable latent concept. Instead, I rely on behavioral indicators of their cognitive interests and expertise.

Contributors who work on the same articles are assumed to know similar things. I use network methods (Brandes et al. 2009; Halatchliyski et al. 2010) to measure the coincidence of contributors who claim to know similar subject matters by editing similar articles. For large networks, finding and describing the distribution of such features is an important way to understand patterns for the whole network and actors embedded in it (Hanneman and Riddle 2005). I use a broader measure of cognitive similarity, when contributors collaborate on *similar* articles (although not the exact same ones) by virtue of these articles belonging to the same topical category. Wikipedia has an extensive categorization schema that is used to organize its articles. Interlocking category memberships of articles indicate that their knowledge domains are categorically interrelated (Shiffrin and Börner 2004). Categorization divides the world into groups of concepts that are in some way similar to one another. It permits us to find order and meaning in novel experiences by imposing boundaries and extending meaning based on what is already known about the categories (Zerubavel 1993). Two contributors may be cognitively very similar, despite having never worked on precisely the same articles. However, they may have worked on separate articles that are categorically-related, which represents a 'second order' level of cognitive similarity between contributors because categorization represents constraint on what information people have access to (Converse 1964; Martin 2002) and structures who else they connect to (Bowker and Star 1999). Cognitive diversity, which can improve group performance by bringing differing perspectives and sets of knowledge to bear on the same problem, can be represented by the overlap of group members' claims to knowledge based on commonalities in their editing behaviors.

Information quality in Wikipedia

Finally, I briefly reflect on the primary outcome variable in this study: information quality in peer produced knowledge. Tollefsen (2009) suggests that because Wikipedia as a group is an intentional community whose members share rules, customs, traditions, and policies, we can assume that when the community produces a "mature" article, that reflects the consensus of its editors as a group and the norms of the community. In this case, we can view the quality of such articles as an emergent product that is greater than the sum of the contributions of its main authors. The limitation of such a view, however, is that it merely begs the question when we should consider an article to be "mature." Tollefsen argues that it happens through a lengthy discussion period where authors transform individual contributions into a more neutral, plural perspective. This corresponds closely to the criterion that I use to select my sample of articles: an indicator of an article's maturity is that it is formally nominated to become a "Featured Article."

The emergent properties of "mature" articles can be linked to the actual practices of its authors. In 1980s as scholars began to capture detailed descriptions of the practices scientists do while working (cf. Knorr Cetina, 1981), they also began to open up the black-box regarding the cognitive process of constructing knowledge. These investigations have had a wide-ranging influence how we think about cognition in general. Notably, Hutchins (1995) suggested that we deviate from the view of cognition as an individual, abstract, and solely mental process; individuals, he argued, utilize and manipulate sets of external representations of their mental processes as they perform many complex cognitive tasks. These external representations often are expressed as patterns of social relations among individuals who solve a complementary task. Viewed in this way, cognition does occur in the mind, but can also be understood as the product of a complex web of social relations and meaningful symbolic interactions. The unit of analysis shifts from the individual to the system in which individuals think (Giere and Moffatt 2003; Giere 2002; Hutchins 1995; Latour and Woolgar 1986). The structure of such a system, and how individuals work within it, shapes the cognitive outputs it produces.

Heylighen, Heath and Overwalle (2004) extend Hutchins's argument into an operational framework for analyzing systems of "distributed cognition," which seems highly applicable to peer produced knowledge. Individuals self-organize into meaningful communities intent on performing work that will "scratch their own itch." Collaborators interact by signaling their intentions through their direct work on the article or by communicating through discussion pages about the articles they are authoring. This creates a dynamic system of interactions that is altered under the impulse of individual actions. For example, one person may add a paragraph to the Wikipedia article on 19th century American history discussing demographic shifts in the

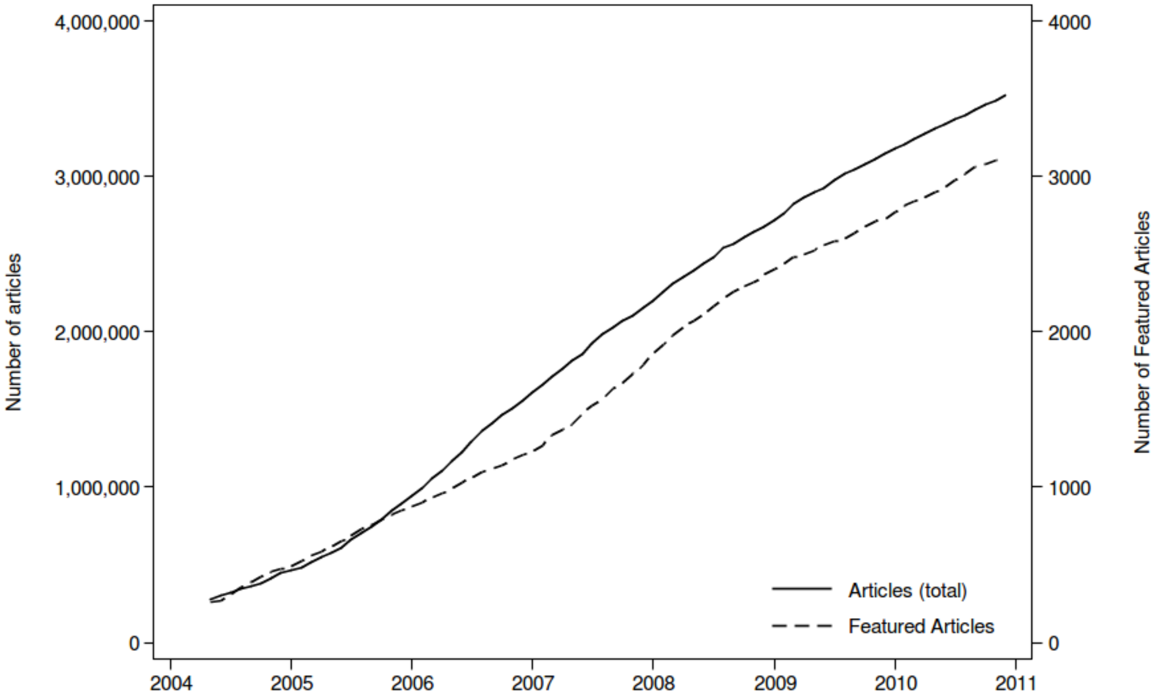


Figure 18: Growth in number of articles on Wikipedia and number of Featured Articles. By 2010, Wikipedia consisted of more than three million articles with more than 3000 Featured Articles. The ratio of total to Featured articles is roughly 1000:1. In other words, only one in every thousand articles on Wikipedia has been developed to the highest level of quality.

nation. In response, a second author may make note about the particular demographic categories used by the U.S. Census at the time and how they differ from today's. A third contributor suggests on the discussion page that the article needs to include more reliable sources, and the process continues. Not surprisingly, these interactions share similarities with simulations of evolution and cooperation (Axelrod 1984; Riolo, Cohen, and Axelrod 2001). Wikipedians created the neologism “darwikinism” to describe this process – a portmanteau of Darwinism and wiki that captures the evolutionary process to which articles are subjected (Boulos, Maramba, and Wheeler 2006; Crompton 2012). Through repeated interactions and selective propagation of information, groups become collectively capable of demonstrating expertise that as individuals they do not possess. Because of the large number of collaborators, the overall effect of their individual limitations are thought to be diminished; in this way, diversity in peer production enables the basic mechanisms for social construction of knowledge that is grounded in actual social practices and interactions among group members.

Diversity and group performance in Wikipedia

Having taken into account group composition, structure, interaction, and the task at hand, I can now formulate a theory of group performance in creating peer produced knowledge on Wikipedia. Groups and individuals within them exhibit a wide range of diversity in terms of the gender, functional orientation, and cognitive dimensions. Diversity can be thought of as a measure of variation on any salient dimension within a group, ranging from no diversity (perfect homogeneity) to complete diversity (perfect heterogeneity). The theory I propose follows the ‘Goldilocks principle’: too little or too much diversity is harmful to a group’s performance, while a medium amount is ‘just right’ for producing the highest quality outcomes. In other words, information quality produced will be lowest when groups are too homogeneous or heterogeneous and will be highest with intermediate levels of diversity on the three dimensions outlined above.

Since women are vastly underrepresented among Wikipedia contributors, this theory suggests that increasing women’s presence will enhance group performance. Information quality should increase as the ratio of women to men in a group approaches parity.[22] The performance of a group should also be higher when the true gender diversity is greater than the ostensible or apparent gender diversity, since this will minimize over conflict related to gender. As discussed above, my conceptualization of gender is more subtle than simply a variable for male/female. The purpose of my research is to contrast the effects of observable gender composition of groups (based on group members’ self-disclosure of their gender) with the true gender composition of groups (based additional information I have gathered on contributors’ private, undisclosed gender). For theoretically-sound reasons, I propose that group performance will be higher when its actual gender diversity is higher than its ostensible gender diversity from a group-member’s perspective. If I find this is the case, it would confirm the paradoxical and highly problematic situation that women’s participation is more valuable so long as no one knows they are women.

[22] Because there were no instances in the sample of articles analyzed for this research where women constituted the majority (let along the vast majority) of contributors, I can state this hypothesis in linear terms. However, there do exist on Wikipedia articles that have been edited predominantly by women; as mentioned, there have been several high-profile examples where women have organized a concerted effort to create and improve articles on specific topics such as women scientists and authors. If these articles were included in the sample, the hypothesis would suggest that the more unbalanced gender ratio in a group the worse performance.

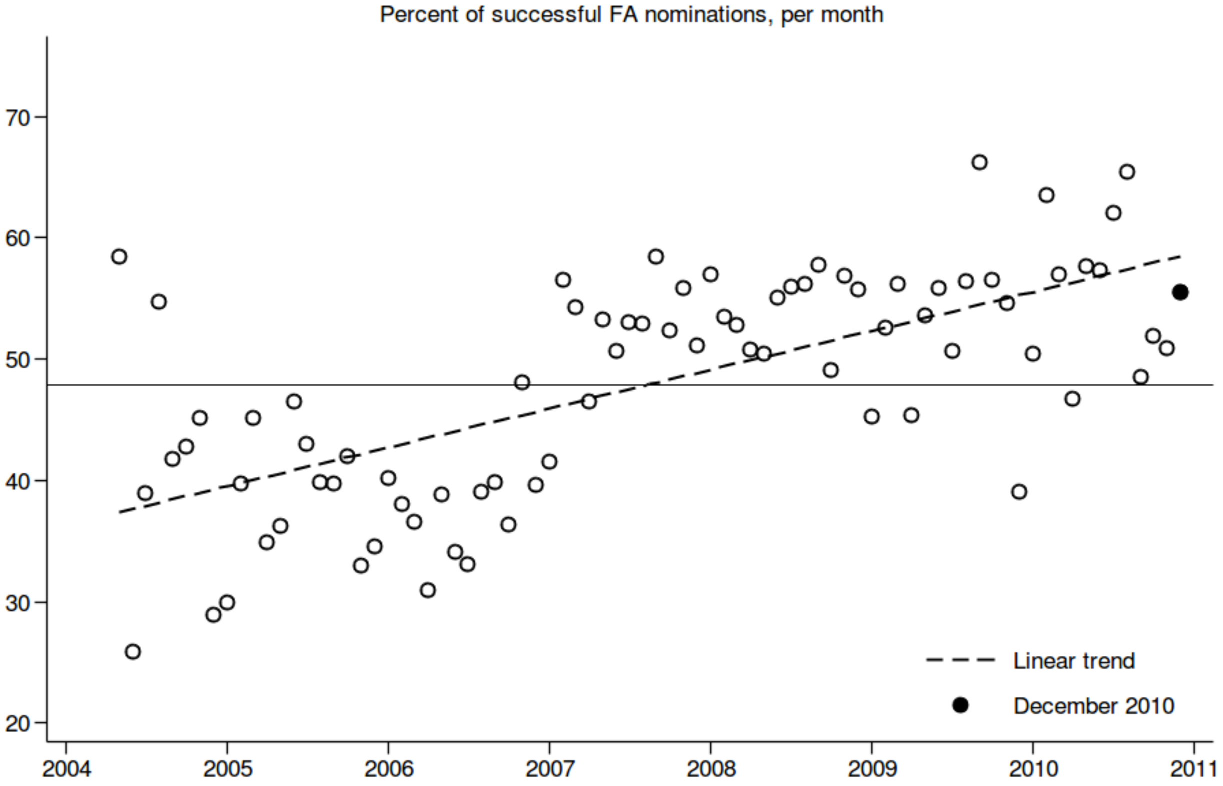


Figure 19: Percent of successful Featured Article nominations per month from 2004-2010. The mean line is 47.9%, and the monthly linear trend in success is positive. The highlighted month, December 2010, had 54 nominations and 30 successes (55%).

Research design

Data on articles

The primary unit of analysis is the nominated article, since groups are defined as contributors who work together on the same article. Since the Wikipedia community began their own internal article assessment project in 2004, through the end of 2010 (when the data were collected for this research), a total of 8009 articles have been nominated and 3804 have been successfully promoted to FA status (47.9% success). Articles are nominated on a monthly basis. As Figure 19 shows, the proportion of successful nominations has tended to increase over time.

My plan for the full analysis was to select a random sample of article candidates, stratified by year. As part of a preliminary analysis, I focused on only one month of nominated articles nominated from December 2010. While this one month's articles may not be representative of the historical population of nominated articles, by using this limited subset I can avoid the confounding effect of how the Featured Article process may have changed over time. After testing my theory about how diversity affects quality on this subset, I would then extend my analysis to a representative sample drawn from the population of all nominated articles.

A total of 54 candidate articles were nominated in December 2010. Of these, 30 were successfully promoted to become a Featured Article while 24 failed; for three of the articles, there was insufficient information available, yielding N=51 articles for my preliminary analysis. Using a custom written program, I downloaded data on the contribution histories of all the editors to these articles, as well as their associated discussion (talk) pages.

The articles included a diverse range of topics, including "The Walt Disney Company," "Missouri River," "Canadian heraldry," and "Star Trek IV: The Voyage Home." As is typical on Wikipedia, the distribution of overall work efforts within the group of contributors for each article was unequal. To measure this inequality, I calculate a Gini coefficient for each group, which ranges from 0 (perfect equality) to 1 (perfect inequality). Higher Gini coefficients indicate that a greater share of total work was performed by a smaller number of contributors. The Gini coefficient for articles ranged from .429 to .885, with a mean of .732. In exploratory analyses, I controlled for the level of inequality in share of work; however, I omitted this variable in the final analysis since I was able to substitute a more meaningful measure of share of work performed. For these articles, the mean number of edits made by contributors was 8.07, with the smallest number being only 1 edit and the largest 625 edits.

Data on contributors

In addition to data on the nominated articles themselves, I also collected the contribution histories of the editors to the articles under consideration. There were more than two thousand unique contributors to these articles (n=2394). For each contributor, I gathered the complete record of that contributor's prior work. I did this to discern the different functional roles the contributor performed, as well as the topics of other articles the contributor worked on. This permitted me to create variables for the functional and cognitive diversity of the group of editors for each article.

For the focal variable, gender, I combined data from three sources: (1) the public API from the English Wikipedia in which contributors can set their gender in the software preferences; (2) the user profile pages of contributors in the sample, where contributors can self-disclose their gender; (3) privately-obtained information from contributors solicited through email. As described before, the public data was more than 80% missing, with only 15% of contributors disclosing as male and 1% as female.

The key innovation that I have proposed in this chapter is a way to address the vast proportion of missing gender data. I introduced theoretically sound reasons why gender in online communities should be viewed as a multidimensional concept, that includes both an element of identity as well as performativity. To supplement the missing data in that is publicly available, I privately solicited information from contributors to these articles, asking them to disclose their gender (as well as optionally respond to an open-ended question that asked about their experiences editing Wikipedia). I created this using an anonymous survey so that no individually-identifiable information could be collected, and sent the survey to a targeted selection of contributors. The contributors who I targeted were the ones who performed large shares of the workload for these articles and yet whose gender was unknown. There were literally hundreds of other contributors whose gender was unknown as well, but these were people who only performed one or two total edits to these articles. My strategy was to start at the top of the list of people who did the most work and work my way downward, to collect as much data as possible.

However, the response rate to my survey was extremely low. Fewer than one percent of the contributors who I contacted completed the survey. This may be due to the relatively long time (more than one year) that had elapsed since these articles were nominated and when I conducted this research. It is possible that many contributors never received my solicitation for their input.

Despite my best efforts, this portion of my data collection proved to be a complete failure. I was only able to add information on twelve contributors' gender to my dataset. In other words, an overwhelming amount of missing data still remained. In light of this failure, I decided to proceed with the analysis anyway, albeit in a scaled back form. The original theoretical insight – that groups have both an "apparent" as well as a "true" gender composition, which may be influential in shaping the conduct of their work – cannot be tested with any sufficient statistical rigor given the limited amount of data available to me.[23]

This situation highlights the most significant limitation that I faced while conducting the research for this dissertation. While it is easy to obtain and analyze "big data," there is the possibility that there is theoretically important information, such as gender in this case, that cannot be reliably collected. In such instances, as I emphasized in Chapter 1, we should interpret with extreme caution any statistical findings that were derived in the absence of this information because of the possibility that the omitted variable would significantly change the results.

Nonetheless, I am still capable of continuing the analysis along the one dimension of gender which I was able to collect, which still may be salient to group outcomes. This is the ostensible or apparent gender composition of a group, from the point-of-view of a member of the group. The hypothesis stated above is that gender diversity leads to better quality outcomes. Unfortunately, without information about contributors' true (but unrevealed) gender, I am incapable of assessing whether there is a differential effect – as hypothesized – between the true but unknown gender diversity of a group and its ostensible, publicly visible gender diversity. This question, sadly, will remain unanswered until more information can be reliable acquired about unrevealed characteristics of contributors.

Dependent variables

Successful promotion: The dependent variable is whether an article is successful in being promoted to be a "Featured Article." This indicates the best performing groups in terms of producing an article judged to have the highest information quality. The variable is dichotomous, with 0 representing an article that fails to achieve this milestone and 1 an article that is successfully promoted. Figure 19 shows the rate of successful promotions over time. The average success rate for all nominated articles is just under half (47.9%), with a linear trend showing an increasing rate of successful nominations over time. In my analysis, I only include articles nominated in December 2010, so as to avoid the possible confounding effects of time.

Independent variables

Gender diversity: To test how the gender composition of a group affects its performance, I construct several measures of a group's gender diversity. I consider both women's membership within a group as well as

[23] As an aside, I should note that with the very limited data that I was able to obtain, it appears to be the case that women's participation, without revealing their gender, is additionally beneficial to group performance over and above the benefit of having a group that is publicly seen to have a diverse gender composition. However, "appears to be the case" is simply another way of saying, "not having enough evidence to say one way or the other with any statistical certainty."

share of work performed by women contributors. My measure of women's group membership is computed using the natural logarithm of the ratio of number of female (plus one) to male group members (plus one) [1]. This ensures that diversity is represented on the real number line with larger values indicating greater women's participation and zero indicating gender parity. (In no instances were there more women than men in the group.)

To measure the share of work performed by female group members, I take the logarithm of the proportion of women's edits (plus one) to men's edits (plus one) to an article [2]. Again, this measure ensures that gender diversity is represented as a real number with larger values indicating greater women's share of the overall work on an article, with zero indicating gender parity.

These alternative measures of gender diversity capture the potentially differing effects of *participation* in a group versus the actual share of *work performed* within a group. Since group-work is unequally distributed, merely having women in the group may not be as consequential as having women perform a large share of the group's work. I construct these two measures for each article as well as separately for each article's discussion (talk) page. This permits me to separately assess whether women's work on the article or women's participation in discussions about the article have the hypothesized consequences for the article's quality.

Functional diversity: Azary et al. (2011) measures functional diversity using information entropy from organizational theory (Cummings 2004; Murphy and Hasenjaeger 1973; Shannon 1963) to measure concentration of group members across different types of community-oriented roles of governing and coordinating work. To implement the measure here, I compute the entropy for each article group using [3], where N represents the total number of different functional roles and p_i represents the proportion of editors having acted in a particular role. I then standardize these scores with a standard deviation of one and a mean of zero. When entropy is above zero, contributors to an article perform many actions across other community roles, whereas when it is below zero, contributors tend to concentrate their behaviors into fewer roles.

Cognitive diversity: Contributors who edit the same or similar articles have a degree of cognitive similarity. To construct my measure of cognitive diversity, I begin with a matrix listing nominated article's contributors on one dimension and other articles edited by these contributors on the other dimension. Each cell in the matrix indicates whether a particular contributor edited the corresponding article. I then transform this matrix into a one-mode author affiliation matrix A_{ij} that counts when contributors i and j are both group members of an article. Finally, I use the network density of the author-affiliation matrix, where n is the number of contributors and A_{ij} is the number of ties between contributors for each article k [4]. This measure varies from zero (indicating no overlap in co-authorship network) to one (indicating complete overlap in co-authorship). I then standardize these scores with a standard deviation of one and a mean of zero.

Because network density will decrease as the number of articles increases, I decided that the initial author affiliation network proved to be to sparse since it yielded extremely large measure of cognitive diversity for each focal article with little variance. To illustrate why this was the case, consider a simplistic example where one person edits the article on "Emile Durkheim" and a second person edits the article on "Pierre Bourdieu." These two contributors would be considered cognitively dissimilar (diverse) because they did not edit the exact same article. Instead, I opted to construct a second-order measure of cognitive diversity based on co-authorship in categorically similar articles, and not just identical articles. In the formula, the number of author affiliations increases because $k_{articles} > k_{categories}$, since multiple articles in the same cognitive domain are contained within the same higher-level category. This reduces the measure of cognitive diversity within the group. I constructed this measure using the same procedure as above. In the example of "Emile Durkheim" and "Pierre Bourdieu," contributors to these articles would be considered cognitively similar since both articles are members of the category “French Sociologists.”

[1] $$gender\ diversity_1 = \ln\left(\frac{women+1}{men+1}\right)$$

[2] $$gender\ diversity_2 = \ln\left(\frac{women's\ edits+1}{men's\ edits+1}\right)$$

[3] $$functional\ diversity = -\sum_{i=1}^{N} p_i \ln p_i$$

[4] $$cognitive\ diversity = \frac{\sum_k A_{ij}}{n(n-1)}$$

Figure 20: Summary of diversity measures

Control variables

Prior research on information quality in Wikipedia suggest additional factors that can affect the quality of an article or its success at the promotion process: the *number of contributors* has been shown to decrease article quality by raising coordination costs (Kittur and Kraut 2008) and the overall *amount of work on an article* has been shown to increase quality (Blumenstock 2008; Poderi 2009; Wilkinson and Huberman 2007). When a group has too many contributors, coordination and dispute resolution may overwhelm the group’s ability to perform their task at hand. While this relationship is thought to be curvilinear – too few contributors can be detrimental to an article’s quality – in this case since I am only focusing on Featured Article candidates, there are no articles with only a very few number of contributors. I nevertheless tested for such a curvilinear relationship; there was none.

Method

Because the dependent variable (whether an article is promoted) is dichotomous, I use binary logistic regression in my analysis. Tabachnick & Fidell (2012) suggest a minimum of ten cases to every one predictor with a minimum sample size of fifty when using logistic regression. Therefore, with the sample size I am working with, I am limited to about five predictors per model.

In Table 6, I present the binary logistic regression estimates of gender diversity’s effects on article promotion. Before conducting my analysis, I checked several diagnostics to guard against violations of the logistic regression assumptions. First, I checked for multicollinearity by referring to the variance inflation factor (VIF) scores. According to conservative guidelines, VIF scores should not exceed a value of 2.5 for any independent variable in the model (Allison 1999). Because of the high correlation between article-group size and discussion-group size, my modeling strategy is to consider the effect of these predictors in separate models. I also could not include both article-group and discussion-group measures in the same model due to the limited number of predictors. Second, I checked that there were no influential cases by examining standardized residuals and confirming that none had a value greater than two.

Models 1-4 include predictors for the article group, and Models 5-8 include predictors for the discussion group. In Model 1, I estimates a baseline model that includes a measure of a group’s functional diversity, its cognitive diversity, and the total amount of work performed on the article. In Model 2, I include the total number of contributors. I retain all these predictors as control variables for Models 3 and 4. In Model 3, I include the measure of women’s share among contributors, and in Model 4, I include the measure for women’s share of work performed. Models 5-8 follow the same modeling strategy but instead focus on the discussion group.

I tested Model 3, which included the independent variable for gender diversity, against the control-variable model (Model 2). The result of the Wald test was statistically significant ($X^2 = 7.88$, $p < .01$), indicating that gender diversity increased the ability to reliably distinguish between successful and unsuccessful article promotions. I performed the same for Model 5 versus Model 2, which again indicated that the inclusion of gender diversity resulted in a statistically-significant improvement over the control variables-only model.

Not reported in the table, the Cox-Snell R^2 for Model 3 indicates that the model accounted for around 31% of the variance in nomination success, whereas the Cox-Snell R^2 in Model 4 indicated that the model accounted for around 28%. This suggests a small preference for measuring gender diversity as women's participation in a group rather than women's share of work performed. Overall, these models are able to discriminate between success and failure in an article's promotion decision, but a significant portion of unexplained variance remains. This points to the need to develop a reliable way of measuring information on contributors' gender that was otherwise missing in this analysis. Next, I move on to specific discussion of the results of my analysis.

Results

Table 6 presents the results of the logistic regression predicting whether a candidate article is promoted to "Featured" status. To facilitate interpretation, I report exponentiated coefficients (odds ratios) where values greater than one indicate increased likelihood that an article will be promoted. Model 1 is a baseline model that controls for the functional and cognitive diversity of the group of contributors and the total amount of work performed on the article at the time of nomination. None of these predictors are statistically significant.

In Model 2, consistent with theory, groups with relatively more contributors have lower performance. The mean number of contributors in this sample was 44 per group. In other words, each additional contributor was associated with a 6.4% lower chance of success (e.g. a group with 54 contributors had a 30% lower chance of success). The coefficient for functional diversity, but not cognitive diversity, becomes statistically significant in this model. Not merely having more contributors, but having more that have experience performing various roles in Wikipedia's community, becomes a significant predictor of article success.

In Model 3, which incorporates the test of gender diversity, we see the coefficient for women's share among contributors in the group is positive (above one) and statistically significant. This indicates that when more women are represented in the group, the odds of a successful outcome are significantly higher. In this model, the total work variable is also statistically significant. This suggests that not merely more work being performed, but more of women's work, may be the mechanism that is driving the increased chance of success.

Model 4 explores this possibility by testing an alternate form of gender diversity – women's share of the overall work effort – which yields a consistent result as the previous model. The coefficient for total work is no longer statistically significant, while the coefficient for women's share of work is positive and statistically significant. This provides support for the notion that women's contributions increase the article's chance of success. In Models 3 and 4, we also notice that the measure of functional diversity remains positive and statistically significant.

Models 5-8 focus on the article's discussion group. In Model 5, none of the control variables are statistically significant, which echo the results from Model 1. In Model 6, we see that the coefficient for discussion size is above one and statistically significant, indicating that increased discussion about an article's contents yields a greater chance of successful promotion. This is to be expected, since discussion pages are one of the primary vehicles though which editors can plan and organize their collaborative work. However, we also see in Model 6 that the coefficient for number of discussants is negative (less than one) and statistically significant. This remains consistent with the results we saw in Model 2, which suggests that larger discussion groups necessitate extra coordination work which can be detrimental to the group's success. In Models 7 and 8, neither measure of gender diversity – women's share in the discussion group, and the actual amount of the discussion engaged in by women – are good predictors of success.

Given the limited sample size and hence the restrictions on the acceptable number of parameters in the model, I conducted a few exploratory analyses (not presented here) which are suggestive of directions for future research. I considered the possibility that these findings may be being driven by the gender of the contributor who nominated the article or by the gender of the evaluator of the article. Controlling for the nominator's gender (which was only available for 28 of the 51 articles), the results remained consistent; I still find that gender diversity in the group was a positive and statistically significant predictor of successful article promotion. As for the gender of the evaluator, the evaluator's role is not to judge the article, but rather to summarize the review of the article by other members of the community and to synthesize their arguments for or against promotion. As such, the evaluator reports the judgment of the community about whether an article warrants promotion. Given that some evaluators may render judgment on several articles, I estimate a multi-level logistic with random-effects, which treats within-evaluator association as a random variable. I also simply controlled for the gender of the evaluator. Neither yielded significant results.

Discussion

Although prior research (Arazy and Kopak 2011) shown that functional and cognitive diversity of groups are associated with higher-quality outcomes, I do not initially find support to confirm this notion in my research. One reason that I suspect this is so is because of the more conservative way that I measure quality: functional and cognitive diversity of groups may be useful for discriminating between the lowest and the highest quality articles, but alone, these factors alone do not help us to discriminate the Featured Article candidates from the Featured Article winners. However, the mechanism may become activated as the group size increases, bringing a more diverse set of contributors with experience in Wikipedia's myriad community roles. On the other hand, more contributors also tend to increase coordination costs and task conflict, which can be detrimental to a group's success. There appears to be a trade-off between these facilitating and hindering factors.

We also see that women's participation in a group, as well as their share of the group's work, significantly increases the chance of an article being promoted. This finding confirms the intuition that when women participate in the editing process, their inclusion leads to improved overall performance of the group. The gender diversity may mediate the the effect of functional diversity, since prior research suggests that women who participate in Wikipedia, at least when visible to others, tend to engage in a more diverse range of functional roles (Antin et al. 2011; Arazy et al. 2011). Disentangling the ways in which gender and functional diversity are interrelated remains an open question for future research.

Interestingly, we find that gender plays less of a direct role in the discussion groups associated with each article, which are essential sites where work can be coordinated and contributors can identify and plan how to improve the article's content. However, gender may play an indirect role in this regard, since research on computer-mediated interactions suggests that online discussions still cleave along gender lines where power displays are common (Sussman and Tyson 2000). If women bring valuable perspectives and approaches to online discussions, possibly due to their aforementioned experience in many community roles, then this still may not benefit the group if their voices are undervalued in group discussions.

These findings are extremely suggestive of the deleterious effects of the gender gap in participation. However, they are still based on analysis that includes quite a large amount of missing data, which precludes me from considering the possible difference between women's actual participation and their visible presence in the group. Such a nuanced analysis is truly warranted by these findings, but given the limitations posed by the available data and the difficulty of acquiring this information speaks to the tentative state of knowledge regarding this important question.

Much of what has been written about the gender gap in participation has focused on one problematic consequence: underrepresentation of articles about women (and other "gendered" topics). However, as I demonstrate in this chapter, there are less obvious but equally important consequences of the gender gap regardless of the topic. Stated quite plainly, women's inclusion in Wikipedia seems to be beneficial, despite the persistence of practices of the majority of contributors who may not be open to women's voices, experiences, and ways of participating. This suggests the importance of fostering a culture of inclusive participation if the organization is serious about achieving its stated goals of creating a high-quality, free encyclopedia available to all.

Table 6: Logistic regression estimates of successful promotion to Featured Article for 51 articles nominated from December 2010

	Model 1	Model 2	Model 3	Model 4
Functional diversity	1.378	2.826 *	3.895 *	3.166 *
	(0.93)	(2.27)	(2.49)	(2.16)
Cognitive diversity	1.089	1.378	1.472	1.456
	(0.28)	(0.80)	(0.85)	(0.86)
Total work	0.999	1.002	1.004 *	1.003
	(-0.71)	(1.83)	(2.12)	(1.53)
Number of contributors		0.936 **	0.902 **	0.922 *
		(-2.61)	(-2.94)	(-2.07)
Women’s share of contributors			4.45 **	
			(2.81)	
Women’s share of work				1.013 *
				(2.02)
N obs.	51	51	51	51
loglikelihood	-33.73	-27.01	-24.16	-25.08

	Model 5	Model 6	Model 7	Model 8
Functional diversity	1.632	1.464	1.477	1.522
	(0.70)	(0.46)	(0.48)	(0.49)
Cognitive diversity	1.021	1.699	1.877	1.538
	(0.03)	(0.65)	(0.73)	(0.48)
Discussion size	0.996	1.035 *	1.036 *	1.04 *
	(-0.64)	(2.07)	(1.96)	(2.21)
Number of discussants		0.845 **	0.842 *	0.833 **
		(-2.72)	(-2.56)	(-2.79)
Women’s share of discussants			1.763	
			(0.45)	
Women’s share of discussion				0.935
				(-0.97)
N obs.	51	51	51	51
loglikelihood	-34.12	-29.35	-29.21	-28.94

Notes: odds ratios with t statistics in parentheses; * p<0.05, ** p<0.01, *** p<0.001

Chapter 5. Conclusion, limitations, and future directions

In this chapter, I revisit the main contributions of my empirical studies, review the limitations that I encountered, and suggest some possible avenues for future. Across all the questions addressed in this dissertation, the underlying theme was that of participation inequality. My interest is not in determining whether inequality is "good" or "bad," nor whether it is "functional" or "dysfunctional" for the organization. Rather, I sought to better understand the the origins and consequences of different forms of inequality – and in particular, how these relate to questions of who volunteers and why, as well as how this shapes the product of their efforts.

My investigation began with the question of why people volunteer to work for free to produce a public good. The solution that has been suggested – that social considerations, such as status and recognition, substitute for absent material incentives – is largely supported by my Barnstar experiments in Chapter 2. Among top contributors, status-based social rewards foster a virtuous cycle that further binds these contributors to the community. But this also points to a process whereby cumulative advantage can exacerbate status inequalities, since when all other things are equal, status grantors tend to disproportionately recognize the efforts of higher status individuals in the community.

My evidence suggests that this gap is not closed when rewards flow to contributors whose efforts are lower relative to their more prolific peers. This suggests that rewards are not constitutive of the overall degree of participation inequality that we see, and thus the incentive structure in Wikipedia appears to be broadly meritocratic. The vast differences we see in productive effort can be largely attributed to differential intrinsic motivation, and the key to retaining these enthusiastic volunteers appears to be sufficient social recognition of their generosity.

In a similar study of Barnstars, Shaw (2012) found that the effectiveness of social rewards to reinforce participation may be moderated by interactional aspects of participation. For example, contributors who prominently displayed their awards to others in the community also tended to exhibit more sustained contributions than those who did not. In my study, I found that an individual's prior level of contribution moderated the effect of receiving a Barnstar. In both cases, the solution to the underlying puzzle – why people volunteer and what sustains their efforts against the tendency to free ride – is that social incentives can be a sufficient counter-weight against this tendency, so long as as the virtuous cycle that Willer (2009) posits tends to strengthen contributors' social identification with the community and the organization's reason for being.

In Chapter 3, I take aim at another possible explanation for high participation inequality. The emergent organizational level structures of Wikipedia's informal bureaucracy, which developed to cope with the remarkable growth in size of the contributor community, may drive away new volunteers while propping up existing contributors. My analysis suggests that as time goes on, new contributors indeed experience greater difficulty in entering and integrating with the community. Yet this does not seem to be caused by a commensurate rise of a managerial class. The development and evolution of Wikipedia's bureaucracy instead broadly supports the existing efforts of volunteers, even as it makes it more difficult for others to begin this endeavor.

Let us take a moment to consider how the Wikipedia community has tried to address concerns about its growing bureaucracy. For example, there have been various efforts to institute "welcoming committees" and other ways of reaching out to new contributors to make initial encounters with Wikipedia's increasingly formalized procedures and work-flows a little less scary to new recruits. Such efforts can go a long way to repairing Wikipedia's reputation as being hostile to newcomers. But on the other hand, I do not believe that such efforts can do much to close the wide gap in participation that we see. The reason is that only a small number of new contributors have sufficient enthusiasm and high intrinsic motivation to become lasting members of the community. As Panciera et al. (2009) suggest, the initial user experience can be improved to channel these small number of contributors' intense energies, but it cannot instill this in others. The consequence, ironically, is that participation inequality can be exacerbated as the harmful effects of bureaucracy are attenuated.

I explore a second consequence of inequality – in this case, the large and persistent gender gap on Wikipedia – in Chapter 5. Peer production would seem to be an ideal location where the beneficial aspects of diversity have an opportunity to to flourish, since participation is open to all in a non-compulsory and self-directed manner. However, women are vastly underrepresented in the community. The best estimates of women's share of participation put that figure at under 20%, but this obscures the existence of an even larger gender gap. Contributors may optionally disclose their gender identity to others, and since women do so in lower proportions than men, women's visibility in the community represents a mere one percent of all contributors. I term this divide the invisible gender gap.

Consistent with theories that suggest diversity can be beneficial to a group's performance, I find support for this notion among the groups that author Wikipedia's articles. Women's participation in a group yields more positive outcomes for the group's ability to develop the highest quality articles. This should not be surprising, since other research (Antin et al. 2011) suggests that women perform a more diverse range of community roles in the community; yet the finding that women's inclusion in a group is beneficial for their shared work continues to hold true even when controlling for the functional diversity that they may bring to the group.

Much of the existing concern regarding the gender gap in Wikipedia focuses on the consequences it has for topical coverage, such as whether articles will reflect women's historical contributions to science or literature. My findings speak to a second negative consequence which is much broader: women's participation appears to be beneficial regardless of the topic being covered. If the Wikipedia community is intent on producing a truly comprehensive and high-quality encyclopedia, increasing women's participation is vital. Not only will this ensure equitable coverage of a vast array of topics, but it may also more generally improve the quality of any articles. Confronting the sexist culture (Reagle 2012) of online collectives is one of the main challenges toward inclusive participation.

There are a number of important limitations to this dissertation that prevent me from providing a more comprehensive understanding of inequality in peer production. To begin, the most obvious one is that this dissertation focuses on merely one organization and community. While I argued that Wikipedia is a strategic research site, it still limits me from being able to make inferences across different organizations that employ peer production without a comparative case or an even more comprehensive statistical analysis across multiple organizations. I can say, however, that at least part of this dissertation's findings hold for another similar organization, WikiNews (a project of collaborative journalism), where I replicated the analysis in Chapter 3 and found very similar results, although I have not presented them here. Comparison across one or more other organizations would yield further insights into whether the same social dynamics in other contexts create similar patterns of inequality.

To this end, with Arnout van de Rijt and in collaboration with Soong Moon Kang and Akshay Patil, we replicated the Barnstar experiment and extended it to several other online contexts. The purpose was to investigate the "success breeds success" dynamic in reward systems, which we saw in Chapter 2 as contributors began to accumulate additional social recognition disproportionate to merit. These follow-up experiments, across the domains of social status, crowdsourced funding, endorsement, and reputation, found that an initial success bestowed upon otherwise undifferentiated individuals produced a comparable increase in rates of subsequent successes. The findings from these series of experiments are scheduled to be published in the *Proceedings of the National Academy of Sciences* shortly after the completion of this dissertation. This demonstrates one of the ways that the work I performed here points to directions for future research.

Future work should further refine the analysis to isolate and test causal mechanisms that explain the origin and maintenance of the high degree of participation inequality. In particular, we need a more systematic way of measuring the valence of social interactions; in principle this can be done using small datasets, but as we scale up our data collection, it becomes more difficult to measure. For example, it undoubtedly matters to a new users whether bureaucracy is experienced in a positive or negative light, that is to say, as enabling or constraining. However, to systematically measure the 'valence' of bureaucratic encounters and work contexts on a large scale remains a challenge, since such qualities inhere in the subject experiences of contributors and cannot easily be discerned in digital trace data. On this point, I refer back to Weber's two components of an explanation – adequacy on the level of causality and adequacy on the level of meaning. In this case, it may only be possible to obtain data on meaningful adequacy from thick, rich descriptions, which do not lend themselves to "big data" approaches. Bridging this divide remains a significant, and perhaps intractable, obstacle. Such distinctions could further refine the findings in Chapter 3, or could potentially yield the opposite results altogether. Therefore, the empirical results from my studies remain tentative but highly suggestive.

Finally, my research on the gender gap in Wikipedia remains equally inconclusive due to the large amount of missing data. Although studies of gender differences in online communication and interaction would benefit from a more thorough conceptualization of gender rather than simply including it as a dichotomous variable, even this latter approach hinges on having access to such information. As I have argued, when we are presented with enormous datasets, as can be the case with online organizations, it is imperative that we do not lose sight of what information we are missing, nor of the enormous consequences that can stem from such missing data.

Even with full information, however, we face a second challenge: it may be the case that women select work for themselves that also happens to be associated with articles of higher quality. In other words, we need a way to disentangle the effect of women's participation on a group, from women self-selecting into groups doing higher quality work. The research in my study here only points to their correlation. To address this limitation, I am fortunate to have had the opportunity to work with Arnout van de Rijt and Hyanggi Song on an experiment where we randomly assigned students into either gender mixed or gender segregated online groups, who were then given tasks to complete. Because we could manipulate the gender composition of groups, we could rule out the possibility that the resulting quality of their work was due to self-selection. Our results, which are in preparation, suggest that women's participation does indeed yield higher quality outcomes. My study on Wikipedia demonstrates how correlational research can inform the design of experimental studies which can tease out how these mechanisms operate.

In sum, I hope you now know more, but with less certainty, about how and why Wikipedia works.

References

Aaltonon, Aleksi, and Giovan Francesco Lanzara. 2010. "Unpacking Wikipedia Governance: The Emergence of a Bureaucracy of Peers?" in *Latin American and European Meeting on Organizational Studies*. Buenos Aires, Argentina: European Group for Organizational Studies.

Adler, Paul, and Bryan Borys. 1996. "Two Types of Bureaucracy: Enabling and Coercive." *Administrative Science Quarterly* 41(1):61 – 89.

Allison, Paul. 1999. *Multiple Regression: A Primer*. Thousand Oaks, CA: Sage.

Allison, Paul D. 1984. *Event History Analysis: Regression for Longitudinal Event Data*. Newbury Park, CA: Sage.

Allison, Paul D. 2001. *Missing Data*. Newbury Park, CA: Sage.

Amason, A. C. 1996. "Distinguishing the Effects of Functional and Dysfunctional Conflict on Strategic Decision Making: Resolving a Paradox for Top Management Teams." *Academy of Management Journal* 39(1):123–48.

Andersen, Per Kragh, and Richard D. Gill. 1982. "Cox's Regression Model for Counting Processes: A Large Sample Study." *The annals of statistics* 1100–1120.

Anderson, Chris. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More*. New York: Hyperion.

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired magazine* 16.

Andreoni, James. 1988. "Why Free Ride? Strategies and Learning in Public Goods Experiments." *Journal of Public Economics* 37(3):291–304.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *The economic journal* 100(401):464–77.

Andreoni, James, and John Karl Scholz. 1998. "An Econometric Analysis of Charitable Giving with Interdependent Preferences." *Economic Inquiry* 36(3):410–28.

Angwin, Julia, and Geoffrey A. Fowler. 2009. "Volunteers Log Off as Wikipedia Ages." *Wall Street Journal*, A3.

Anthony, D., S. W. Smith, and T. Williamson. 2009. "Reputation and Reliability in Collective Goods: The Case of the Online Encyclopedia Wikipedia." *Rationality and Society* 21(3):283–306.

Antin, Judd, Raymond Yee, Coye Cheshire, and Oded Nov. 2011. “Gender Differences in Wikipedia Editing.” P. 11 in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym ’11*. New York, New York, USA: ACM Press.

Arazy, Ofer et al. 2010. “Recognizing Contributions in Wikis: Authorship Categories, Algorithms, and Visualizations.” *Journal of the American Society for Information Science and Technology* 61(6):1166–79.

Arazy, Ofer, and R. Kopak. 2011. “On the Measurability of Information Quality.” *Journal of the American Society for Information Science and Technology* 62(1):89–99.

Arazy, Ofer, and Oded Nov. 2010. “Determinants of Wikipedia Quality.” Pp. 233–66 in *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW ’10*. New York, New York, USA: ACM Press.

Arazy, Ofer, Oded Nov, and Felipe Ortega. 2014. “The [Wikipedia] World Is Not Flat: On the Organizational Structure of Online Production Communities.” in *Twenty Second European Conference on Information Systems*. Tel Aviv.

Arazy, Ofer, Oded Nov, Raymond Patterson, and Lisa Yeo. 2011. “Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict.” *Journal of Management Information Systems* 71–98.

Arquilla, John, and David Ronfeldt. 1996. *The Advent Of Netwar*. RAND Corporation.

Arquilla, John, and David Ronfeldt. 2001. *Networks and Netwars*. RAND Corporation.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Barabasi, A. L., and R. Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286(5439):509–12.

Bauwens, Michel. 2005. “The Political Economy of Peer Production.” *CTheory* 1.

Bauwens, Michel. 2009. “Class and Capital in Peer Production.” *Capital & Class* 33(1):121–41.

Benkler, Yochai. 2002. “Coase’s Penguin, Or, Linux and ‘The Nature of the Firm.’” *Yale Law Journal* 112(3):369–446.

Benkler, Yochai. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press.

Bitzer, Jürgen, Wolfram Schrettl, and Philipp J. H. Schröder. 2007. “Intrinsic Motivation in Open Source Software Development.” *Journal of Comparative Economics* 35(1):160–69.

Blumenstock, Joshua E. 2008. “Size Matters: Word Count as a Measure of Quality on Wikipedia.” *Proceedings of the 17th international conference on World Wide Web* 1095–96.

Blundell, Richard, and Stephen Bond. 1998. “Initial Conditions and Moment Restrictions in Dynamic Panel Data Models.” *Journal of Econometrics* 87(1):115–43.

Blundell, Richard, Stephen Bond, and Frank Windmeijer. 2000. “Estimation in Dynamic Panel Data Models: Improving on the Performance of the Standard GMM Estimator.” *Advances in Econometrics* 15:53–91.

Bonaccorsi, Andrea, and Cristina Rossi. 2006. “Comparing Motivations of Individual Programmers and Firms to Take Part in the Open Source Movement: From Community to Business.” *Knowledge, Technology & Policy* 18(4):40–64.

Boulding, Kenneth. 1973. “Energy Reorganization Act of 1973: Hearings of the 93rd Congress on H.R. 11510.” 248.

Boulos, Maged N. K., Inocencio Maramba, and Steve Wheeler. 2006. “Wikis, Blogs and Podcasts: A New Generation of Web-Based Tools for Virtual Collaborative Clinical Practice and Education.” *BMC medical education* 6(1):41.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Brandes, Ulrik, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. 2009. “Network Analysis of Collaboration Structure in Wikipedia.” P. 731 in *Proceedings of the 18th international conference on World wide web - WWW ’09*. New York, New York, USA: ACM Press.

Brandes, Ulrik, and Jürgen Lerner. 2008. “Visual Analysis of Controversy in User-Generated Encyclopedias.” *Information Visualization* 7(1):34–48.

Breusch, T. S., and A. R. Pagan. 1979. “A Simple Test for Heteroscedasticity and Random Coefficient Variation.” *Econometrica* 47(5):1287–94.

Brothers, Laurence et al. 1992. “Supporting Informal Communication via Ephemeral Interest Groups.” Pp. 84–90 in *Proceedings of the 1992 ACM conference on Computer-supported cooperative work - CSCW ’92*, *CSCW ’92*. New York, New York, USA, NY, USA: ACM Press.

Brown, Gavin, Jeremy Wyatt, Rachel Harris, and Xin Yao. 2005. “Diversity Creation Methods: A Survey and Categorisation.” *Journal of Information Fusion* 6:5–20.

Bryant, Susan L., Andrea Forte, and Amy Bruckman. 2005. “Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia.” *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* 1–10.

Butler, Brian, Elisabeth Joyce, and Jacqueline Pike. 2008. “Don’t Look Now, but We've Created a Bureaucracy.” in *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI ’08*. New York, New York, USA: ACM Press.

Cameron, Adrian Colin, and P. K. Trivedi. 2005. *Microeconometrics : Methods and Applications: A. Colin Cameron, Pravin K ...* Cambridge University Press.

Campbell, Donald Thomas, and M. Jean Russo. 1999. *Social Experimentation*. Sage Publications Thousand Oaks CA.

Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. "Power-Law Distributions in Empirical Data." *Siam Review* 51(4):661–703.

Coase, Ronald H. 1937. "The Nature of the Firm." *Economica* 4(16):386–405.

Cohen, Noam. 2007. "After False Claim, Wikipedia to Check Degrees." *The New York Times* 12.

Cohen, Noam. 2011. "Define Gender Gap? Look Up Wikipedia's Contributor List." *New York Times, January* 30(362):1050–56.

Collier, Benjamin, and Julia Bear. 2012a. "Conflict, Criticism, or Confidence." P. 383 in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. New York, New York, USA: ACM.

Collier, Benjamin, and Julia Bear. 2012b. "Conflict, Criticism, or Confidence: An Empirical Examination of the Gender Gap in Wikipedia Contributions." Pp. 383–92 in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM.

Collins, Randall. 1981. "On the Microfoundations of Macrosociology." *American Journal of Sociology* 86(5):984–1014.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." Pp. 206–61 in *Ideology and Discontent*, edited by David E. Apter. New York: Free Press.

Cooper, J. 2006. "The Digital Divide: The Special Case of Gender." *Journal of Computer Assisted Learning* 22(5):320–34.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 187–220.

Crompton, Helen. 2012. "How Web 2.0 Is Changing the Way Students Learn: The Darwikinism and Folksonomy Revolution." *E Learning* 2013:2–7.

Cummings, J. N. 2004. "Work Groups, Structural Diversity, and Knowledge Sharing in a Global Organization." *Management Science* 50(3):352–64.

Davis, Murray S. 1971. "That's Interesting." *Philosophy of the Social Sciences* 1(2):309–44.

Demil, Benoit, and Xavier Lecocq. 2006. "Neither Market nor Hierarchy nor Network: The Emergence of Bazaar Governance." *Organization Studies* 27(10):1447–66.

DiTomaso, Nancy, Corinne Post, and Rochelle Parks-Yancy. 2007. "Workforce Diversity and Inequality: Power, Status, and Numbers." *Annual Review of Sociology* 33(1):473–501.

Dolnicar, Sara, and Melanie Randle. 2007. "What Moves Which Volunteers to Donate Their Time? An Investigation of Psychographic Heterogeneity Among Volunteers in Australia." *Faculty of Commerce - Papers*.

Enders, Craig K. 2010. *Applied Missing Data Analysis*. Guilford Press.

Etzioni, Amitai. 1988. *The Moral Dimension: Toward a New Economics*. New York, NY: Free Press.

Finkel, Steven E. 1995. *Causal Analysis with Panel Data*. Thousand Oaks, CA: Sage.

Forte, Andrea, and Amy Bruckman. 2008. "Scaling Consensus: Increasing Decentralization in Wikipedia Governance." in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE Computer Society.

Forte, Andrea, Vanesa Larco, and Amy Bruckman. 2009. "Decentralization in Wikipedia Governance." *Journal of Management Information Systems* 26(1):49–72.

Frey, Bruno S., and Stephan Meier. 2004. "Social Comparisons and pro-Social Behavior: Testing' Conditional Cooperation' in a Field Experiment." *The American Economic Review* 94(5):1717–22.

Gardner, Sue. 2011. "Nine Reasons Women Don't Edit Wikipedia (in Their Own Words)." Retrieved March 28, 2014 (http://suegardner.org/2011/02/19/nine-reasons-why-women-dont-edit-wikipedia-in-their-own-words/).

Geiger, R. Stuart, and Aaron Halfaker. 2013. "Using Edit Sessions to Measure Participation in Wikipedia." Pp. 861–70 in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM.

Ghosh, Rishab Aiyer, and Vipul Ved Prakash. 2000. "The Orbiten Free Software Survey." *First Monday* 5(7).

Giere, Ronald. 2002. "Distributed Cognition in Epistemic Culture." *Philosophy of Science* 69:637–44.

Giere, Ronald N., and Barton Moffatt. 2003. "Distributed Cognition: Where the Cognitive and the Social Merge." *Social Studies of Science* 33(2):1–10.

Giles, J. 2005. "Internet Encyclopaedias Go Head to Head." *Nature* 438(7070):900–901.

Giles, Jim. 2012. "Making the Links." *Nature* 488(7412):448–50.

Gloor, Peter, and Scott Cooper. 2011. "The New Principles of a Swarm Business." *MIT Sloan Management Review* 48(3).

Glott, Ruediger, Rishab Aiyer Ghosh, and Philipp Schmidt. 2010. *Wikipedia Survey, Technical Report.* Maastricht, Netherlands.

Goldspink, Christopher. 2010. "Normative Behaviour in Wikipedia." *Information, Communication & Society* 13(5):652–73.

Goldstein, Michel L., Steven A. Morris, and Gary G. Yen. 2004. "Problems with Fitting to the Power-Law Distribution." *The European Physical Journal B-Condensed Matter and Complex Systems* 41(2):255–58.

Gomm, Roger. 2004. *Social Research Methodology*. Palgrave Macmillan New York.

Goodin, Robert E. 1980. "Making Moral Incentives Pay." *Policy Sciences* 12(2):131–45.

Graham, John W. 2009. “Missing Data Analysis: Making It Work in the Real World.” *Annual review of psychology* 60:549–76.

Granovetter, Mark. 1973. “The Strength of Weak Ties.” *American Journal of Sociology* 78(6):1360 – 1380.

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge University Press.

Halatchliyski, Iassen, Johannes Moskaliuk, Joachim Kimmerle, and Ulrike Cress. 2010. “Who Integrates the Networks of Knowledge in Wikipedia?” P. 1 in *Proceedings of the 6th International Symposium on Wikis and Open Collaboration - WikiSym ’10*. New York, New York, USA: ACM Press.

Halavais, Alexander, and Derek Lackaff. 2008. “An Analysis of Topical Coverage of Wikipedia.” *Journal of Computer-Mediated Communication* 13(2):429–40.

Halevi, G., and H. Moed. 2012. “The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature.” *Research Trends, Special Issue on Big Data* 30:3–6.

Halfaker, A., R. S. Geiger, J. T. Morgan, and J. Riedl. 2012. “The Rise and Decline of an Open Collaboration System: How Wikipedia’s Reaction to Popularity Is Causing Its Decline.” *American Behavioral Scientist* 57(5):664–88.

Hanneman, Robert A., and Mark Riddle. 2005. *Introduction to Social Network Methods*. Riverside, CA: University of California Press.

Hansen, Sean, Nicholas Berente, and Kalle Lyytinen. 2009. “Wikipedia, Critical Social Theory, and the Possibility of Rational Discourse.” *The Information Society* 25(1):38–59.

Hardin, Garrett. 1968. “The Tragedy of the Commons.” *Science (New York, N.Y.)* 162(3859):1243–48.

Hars, Alexander, and Shaosong Ou. 2002. “Working for Free? Motivations for Participating in Open-Source Projects.” *International Journal of Electronic Commerce* 6(3):25–39.

Haythornthwaite, C., and L. Kendall. 2010. “Internet and Community.” *American Behavioral Scientist* 53(8):1083–94.

Healy, Kieran, and Alan Schussman. 2003. “The Ecology of Open-Source Software Development.”

Heckathorn, Douglas D. 1993. “Collective Action and Group Heterogeneity: Voluntary Provision versus Selective Incentives.” *American Sociological Review* 329–50.

Heisenberg, Werner Autor. 1958. *Physics and Philosophy: The Revolution in Modern Science*. Prometheus Books, Publishers.

Hemetsberger, Andrea, and Rik Pieters. 2001. “When Consumers Produce on the Internet: An Inquiry into Motivational Sources of Contribution to Joint-Innovation.” Pp. 274–91 in *Proceedings of the Fourth International Research Seminar on Marketing Communications and Consumer Behavior*. La Londe.

Hemetsberger, Andrea, and Christian Reinhardt. 2006. “Learning and Knowledge-Building in Open-Source Communities a Social-Experiential Approach.” *Management learning* 37(2):187–214.

Hern, Alex. 2014. “Wikipedia ‘Edit-a-Thon’ Seeks to Boost Number of Women Editors.” *The Guardian*, March 4.

Herring, Susan C. 2003. “Gender and Power in Online Communication.” Pp. 202–28 in *The Handbook of Language and Gender*, edited by Janet Holmes and Miriam Meyerhoff. Oxford, UK: Blackwell Publishing Ltd.

Heylighen, Francis, Margaret Heath, and Frank van Overwalle. 2004. “The Emergence of Distributed Cognition: A Conceptual Framework.” in *Conference on Collective Intentionality IV*. Sienna, Italy.

Hill, Benjamin Mako, and Aaron Shaw. 2013. “The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation.” *PloS one* 8(6):e65782.

Hill, Benjamin Mako, Aaron Shaw, and Yochai Benkler. 2012a. *Status, Social Signalling and Collective Action in a Peer Production Community*. American Sociological Association Annual Meeting (Denver, CO).

Hill, Benjamin Mako, Aaron Shaw, and Yochai Benkler. 2012b. “Status, Social Signalling and Collective Action in a Peer Production Community.” Denver, CO: American Sociological Association Annual Meeting.

Hoffman, L. R., and N. R. F. Maier. 1961. “Quality and Acceptance of Problem Solutions by Members of Homogeneous and Heterogeneous Groups.” *Journal of Abnormal and Social Psychology* 62(2):401–&.

Hong, Lu, and Scott E. Page. 2004. “Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers.” *Proceedings of the National Academy of Sciences of the United States of America* 101(46):16385–89.

Huberman, Bernardo A. 2012. “Sociology of Science: Big Data Deserve a Bigger Audience.” *Nature* 482(7385):308.

Humphreys, Ashlee, and Kent Grayson. 2008. “The Intersecting Roles of Consumer and Producer: A Critical Perspective on Co-Production, Co-Creation and Prosumption.” *Sociology Compass* 2(3):963–80.

Hutchins, Edw. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.

Iba, Takashi, Keiichi Nemoto, Bernd Peters, and Peter A. Gloor. 2010. “Analyzing the Creative Editing Behavior of Wikipedia Editors.” *Procedia - Social and Behavioral Sciences* 2(4):6441–56.

Jehn, K. A., and E. A. Mannix. 2001. “The Dynamic Nature of Conflict: A Longitudinal Study of Intragroup Conflict and Group Performance.” *Academy of Management Journal* 44(2):238–51.

Jehn, Karen A. 1995. “A Multimethod Examination of the Benefits and Detriments of Intragroup Conflict.” *Administrative Science Quarterly* 40(2):256–82.

Kirtley, Jane. 2006. “Web of Lies: A Vicious Wikipedia Entry Underscores the Difficulty of Holding Anyone Responsible for Misinformation on the Internet.” *American Journalism Review* 28(1):66.

Kittur, Aniket, Ed Chi, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. 2007. “Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie.” in *Alt.CHI*. San Jose, CA.

Kittur, Aniket, and Robert E. Kraut. 2008. “Harnessing the Wisdom of Crowds in Wikipedia.” P. 37 in *Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW ’08*. New York, New York, USA: ACM Press.

Kittur, Aniket, Bryan Pendleton, and Robert E. Kraut. 2009a. “Herding the Cats.” P. 1 in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration - WikiSym ’09*. New York, New York, USA: ACM Press.

Kittur, Aniket, Bryan Pendleton, and Robert E. Kraut. 2009b. “Herding the Cats: The Influence of Groups in Coordinating Peer Production.” Pp. 7:1–7:9 in.

Kittur, Aniket, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. “He Says, She Says: Conflict and Coordination in Wikipedia.” *Proceedings of the SIGCHI conference on Human factors in computing systems* 453–62.

Van Knippenberg, Daan, and Michaéla C. Schippers. 2007. “Work Group Diversity.” *Annual review of psychology* 58:515–41.

Knorr Cetina, Karin. 1981. *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Oxford, UK: Pergamon Press.

Kollock, Peter. 1998. “Social Dilemmas: The Anatomy of Cooperation.” *Annual review of sociology* 183–214.

Konieczny, Piotr. 2009a. “Governance, Organization, and Democracy on the Internet: The Iron Law and the Evolution of Wikipedia.” *Sociological Forum* 24(1):162–92.

Konieczny, Piotr. 2009b. “Wikipedia: Community or Social Movement?”.” *Interface: A Journal for and about Social Movements* 1(2):212–32.

Konieczny, Piotr. 2010. “Adhocratic Governance in the Internet Age: A Case of Wikipedia.” *Journal of Information Technology & Politics* 7(4):263–83.

Kossinets, Gueorgi. 2006. “Effects of Missing Data in Social Networks.” *Social networks* 28(3):247–68.

Kostakis, Vasilis. 2010. “Identifying and Understanding Problems of Wikipedia’s Peer Governance: The Case of Inclusionists Versus Deletionists.” *First Monday* 15(3).

Kriplean, Travis, Ivan Beschastnikh, and David W. McDonald. 2008. “Articulations of WikiWork: Uncovering Valued Work in Wikipedia through Barnstars.” P. 47 in *Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW ’08*. New York, New York, USA: ACM Press.

Kriplean, Travis, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. “Community, Consensus, Coercion, Control: Cs* W or How Policy Mediates Mass Participation.” Pp. 167–76 in *Proceedings of the 2007 international ACM conference on Supporting group work*. ACM.

Krishnamurthy, Sandeep. 2002. “Cave or Community?: An Empirical Examination of 100 Mature Open Source Projects.” *First Monday*.

Kubiszewski, Ida, Thomas Noordewier, and Robert Costanza. 2011. “Perceived Credibility of Internet Encyclopedias.” *Computers & Education* 56(3):659–67.

Kuznetsov, Stacey. 2006. “Motivations of Contributors to Wikipedia.” *ACM SIGCAS Computers and Society* 36(2):1–7.

Lakhani, Karim R., and Eric von Hippel. 2003. “How Open Source Software Works: ‘free’ User-to-User Assistance.” *Research Policy* 32(6):923–43.

Lam, Shyong (Tony) K. et al. 2011. “WP:clubhouse?” P. 1 in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym ’11*. New York, New York, USA: ACM Press.

Lanzara, Giovan Francesco, and Michele Morner. 2003. “The Knowledge Ecology of Open-Source Software Projects.” in *19th EGOS Colloquium, Copenhagen*.

Latour, Bruno, and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Lazer, David et al. 2009. “Computational Social Science.” *Science (New York, N.Y.)* 323(5915):721–23.

Lee, G. K., and R. E. Cole. 2003. “From a Firm-Based to a Community-Based Model of Knowledge Creation: The Case of the Linux Kernel Development.” *Organization Science* 14(6):633–49.

Leonard-Barton, D., and D. Leonard. 1998. *Wellsprings of Knowledge: Building and Sustaining the Sources of Innovation*. Harvard Business Press.

Lerner, Josh, and Jean Tirole. 2002. “Some Simple Economics of Open Source.” *The Journal of Industrial Economics* 50(2):197–234.

Lim, Sook, and Nahyun Kwon. 2010. “Gender Differences in Information Behavior Concerning Wikipedia, an Unorthodox Information Source?” *Library & Information Science Research* 32(3):212–20.

Luan, Shenghua, Konstantinos V Katsikopoulos, and Torsten Reimer. 2012. “When Does Diversity Trump Ability (and Vice Versa) in Group Decision Making? A Simulation Study.” edited by Alex Mesoudi. *PloS one* 7(2):e31043.

Lucas, Jeffrey W. 2003. “Theory-testing, Generalization, and the Problem of External Validity.” *Sociological Theory* 21(3):236–53.

Mann, H. B., and D. R. Whitney. 1947. “On a Test Whether One of Two Random Variables Is Stochastically Larger than the Other.” *Annals of Mathematical Statistics* 18:50–60.

Markus, M. Lynne. 2007. “The Governance of Free/open Source Software Projects: Monolithic, Multidimensional, or Configurational?” *Journal of Management & Governance* 11(2):151–63.

Martin, John Levi. 2002. “Power, Authority, and the Constraint of Belief Systems.” *American Journal of Sociology2* 107(4):861–904.

Martin, Owen S. 2011. “A Wikipedia Literature Review.”

Mason, Winter, and Duncan J. Watts. 2010. "Financial Incentives and the Performance of Crowds." *ACM SigKDD Explorations Newsletter* 11(2):100–108.

McCrae, Robert R., and Oliver P. John. 1992. "An Introduction to the Five-factor Model and Its Applications." *Journal of Personality* 60(2):175–215.

Medelyan, Olena, David Milne, Catherine Legg, and Ian H. Witten. 2009. "Mining Meaning from Wikipedia." *International Journal of Human-Computer Studies* 67(9):716–54.

Mehra, Amit, and Vijay Mookerjee. 2012. "Human Capital Development for Programmers Using Open Source Software." *MIS QUARTERLY* 36(1):107–22.

Merton, Robert K. 1987. "Three Fragments from a Sociologist's Notebooks: Establishing the Phenomenon, Specified Ignorance, and Strategic Research Materials." *Annual Review of Sociology* 13:1–28.

Michels, Robert. 1915. *Political Parties: A Sociological Study of the Olicarchical Tendencies of Modern Democracy*. Translated. New York: Fre Press.

Mills, C. Wright. 1959. *The Sociological Imagination*. New York: Oxford University Press.

Mockus, Audris, Roy T. Fielding, and James Herbsleb. 2000. "A Case Study of Open Source Software Development: The Apache Server." Pp. 263–72 in *Software Engineering, 2000. Proceedings of the 2000 International Conference on*. IEEE.

Mockus, Audris, Roy T. Fielding, and James D. Herbsleb. 2002. "Two Case Studies of Open Source Software Development: Apache and Mozilla." *ACM Transactions on Software Engineering and Methodology (TOSEM)* 11(3):309–46.

Morell, Mayo Fuster. 2012. "The Free Culture and 15M Movements in Spain: Composition, Social Networks and Synergies." *Social Movement Studies* 11(3-4):386–92.

Morgan, Jonathan T., Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. "Tea and Sympathy: Crafting Positive New User Experiences on Wikipedia." Pp. 839–48 in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM.

Muchnik, Lev, Sinan Aral, and Sean J. Taylor. 2013. "Social Influence Bias: A Randomized Experiment." *Science* 341 (6146):647–51.

Murphy, D. C., and J. T. Hasenjaeger. 1973. *Entropy as a Measure of Decentralization*. Boston: American Institute for Decision Science.

Newman, M. E. J. 2005. "Power Laws, Pareto Distributions and Zipf's Law." *Contemporary Physics* 46(5):323–51.

Nov, Oded. 2007. "What Motivates Wikipedians?" *Communications of the ACM* 50(11):60–64.

Nowak, Kristine L., and Christian Rauh. 2005. "The Influence of the Avatar on Online Perceptions of Anthropomorphism, Androgyny, Credibility, Homophily, and Attraction." *Journal of Computer-Mediated Communication* 11(1):153–78.

Nowak, Kristine L., and Christian Rauh. 2008. “Choose Your ‘Buddy Icon’ Carefully: The Influence of Avatar Androgyny, Anthropomorphism and Credibility in Online Interactions.” *Computers in Human Behavior* 24(4):1473–93.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Cambridge, MA: Harvard University Press.

Oreg, S., and O. Nov. 2008. “Exploring Motivations for Contributing to Open Source Initiatives: The Roles of Contribution Context and Personal Values.” *Computers in Human Behavior* 24(5):2055–73.

Ortega, Felipe, Jesus M. Gonzalez-Barahona, and Gregorio Robles. 2008. “On the Inequality of Contributions to Wikipedia.” Pp. 304–304 in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE Computer Society.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge university press.

Page, Scott E. 2007. “Making the Difference: Applying a Logic of Diversity.” *The Academy of Management Perspectives* 21(4):6–20.

Page, Scott E. 2008. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition)*. Princeton University Press.

Page, Scott E. 2010. *Diversity and Complexity*. Princeton University Press.

Panciera, Katherine, Aaron Halfaker, and Loren Terveen. 2009. “Wikipedians Are Born, Not Made.” in *Proceedings of the ACM 2009 international conference on Supporting group work - GROUP ’09*. New York, New York, USA: ACM Press.

Pelled, Lisa Hope, Kathleen M. Eisenhardt, and Katherine R. Xin. 1999. “Exploring the Black Box: An Analysis of Work Group Diversity, Conflict and Performance.” *Administrative Science Quarterly* 44(1):1–28.

Pentzold, Christian. 2011. “Imagining the Wikipedia Community: What Do Wikipedia Authors Mean When They Write about Their ‘community’?” *New Media & Society* 13(5):704–21.

Perrow, Charles. 1983. “The Organizational Context of Human Factors Engineering.” *Administrative Science Quarterly* 28(4):521.

Perrow, Charles, Harold L. Wilensky, and Albert J. Reiss. 1986. *Complex Organizations: A Critical Essay*. McGraw-Hill New York.

Pigliucci, Massimo. 2009. “The End of Theory in Science?” *EMBO reports* 10(6):534.

Poderi, Giacomo. 2009. “Comparing Featured Article Groups and Revision Patterns Correlations in Wikipedia.” *First Monday* 14(5).

Podolny, Joel M., and Karen L. Page. 1998. “Network Forms of Organization.” *Organization* (Kanter 1991):57–76.

Polletta, Francesca. 2012. *Freedom Is an Endless Meeting: Democracy in American Social Movements*. University of Chicago Press.

Potthast, Martin, Benno Stein, and Robert Gerling. 2008. "Automatic Vandalism Detection in Wikipedia." Pp. 663–68 in *Advances in Information Retrieval*. Springer.

Powell, W. 2003. "Neither Market nor Hierarchy." *The sociology of organizations: classic, contemporary, and critical readings* 315:104–17.

Powell, W. W., D. R. White, K. W. Koput, and J. Owen-Smith. 2005. "Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences." *American Journal of Sociology* 110(4):1132–1205.

Pregibon, Daryl. 1980. "Goodness of Link Tests for Generalized Linear Models." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29(1):14–15.

Priedhorsky, Reid et al. 2007. "Creating, Destroying, and Restoring Value in Wikipedia." Pp. 259–68 in *Proceedings of the 2007 international ACM conference on Supporting group work*. ACM.

Quine, Willard van Orman. 1966. *Ways of Paradox*. New York: Random House.

Ramsey, J. B. 1969. "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statistical Society. Series B (Methodological)* 31(2):350–71.

Raudenbush, Stephen W., Andres Martinez, and Jessaca Spybrook. 2007. "Strategies for Improving Precision in Group-Randomized Experiments." *Educational Evaluation and Policy Analysis* 29(1):5–29.

Raymond, Eric. 1999. "The Cathedral and the Bazaar." *Knowledge, Technology & Policy* 12(3):23–49.

Reagle, Joseph. 2012. "'Free as in Sexist?' Free Culture and the Gender Gap." *First Monday* 18(1).

Reagle, Joseph M. 2007. "Do as I Do: Authorial Leadership in Wikipedia." Pp. 143–56 in *Proceedings of the 2007 international symposium on Wikis - WikiSym '07*. New York, New York, USA: ACM Press.

Reagle, Joseph, and Lauren Rhue. 2011. "Gender Bias in Wikipedia and Britannica." *International Journal of Communication* 5:1138–58.

Restivo, Michael, and Arnout van de Rijt. 2012. "Experimental Study of Informal Rewards in Peer Production." *PLoS ONE*.

Riolo, Rick, Michael D. Cohen, and Robert Axelrod. 2001. "Evolution of Cooperation without Reciprocity." *Nature* 414:441–43.

Ritzer, George, and Nathan Jurgenson. 2010. "Production, Consumption, Prosumption: The Nature of Capitalism in the Age of the Digital 'Prosumer.'" *Journal of Consumer Culture* 10(1):13–36.

Rogelberg, S. G., and S. M. Rumery. 1996. "Gender Diversity, Team Decision Quality, Time on Task, and Interpersonal Cohesion." *Small Group Research* 27(1):79–90.

Rosier, Kate, and Celia Pearce. 2011. “Doing Gender Versus Playing Gender in Online Worlds: Masculinity and Femininity in Second Life and Guild Wars.” *Journal of Gaming &amp; Virtual Worlds* 3(2):20.

Saavedra, R., P. C. Earley, and L. Vandyne. 1993. “Complex Interdependence in Task-Performing Groups.” *Journal of Applied Psychology* 78(1):61–72.

Salganik, M. J., P. S. Dodds, and D. J. Watts. 2006. “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market.” *Science* 311(5762):854–56.

Salganik, M. J., and D. J. Watts. 2008. “Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market.” *Social Psychology Quarterly* 71(4):338–55.

Schiebinger, Londa. 1999. *Has Feminism Changed Science?* Cambridge, MA: Harvard University Press.

Schroer, Joachim, and Guido Hertel. 2009. “Voluntary Engagement in an Open Web-Based Encyclopedia: Wikipedians and Why They Do It.” *Media Psychology* 12(1):96–120.

Shang, Jen, and Rachel Croson. 2009. “A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods.” *The Economic Journal* 119(540):1422–39.

Shannon, Claude E. 1963. “The Mathematical Theory of Communication.” in *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Shaw, Aaron. 2012. “An Interactional Account of Online Collective Action.” *University of California, Berkeley*.

Shiffrin, Richard M., and Katy Börner. 2004. “Mapping Knowledge Domains.” *Proceedings of the National Academy of Sciences of the United States of America* 101 Suppl(suppl_1):5183–85.

Simonite, Tom. 2013. “The Decline of Wikipedia: Even As More People Than Ever Rely on It, Fewer People Create It.” *Technology Review*. Retrieved March 9, 2014 (http://www.technologyreview.com/featuredstory/520446/the-decline-of-wikipedia/).

Simpson, E. H. 1951. “The Interpretation of Interaction in Contingency Tables.” *Journal of the Royal Statistical Society. Series B (Methodological)* 13(2):238–41.

Smith, D. H. 1994. “Determinants of Voluntary Association Participation and Volunteering: A Literature Review.” *Nonprofit and Voluntary Sector Quarterly* 23(3):243–63.

Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. 2012. “‘ Big Data’: Big Gaps of Knowledge in the Field of Internet Science.” *International Journal of Internet Science* 7(1).

Spek, Sander, Eric Postma, and H. Jaap van den Herik. 2006. “Wikipedia: Organisation from a Bottom-Up Approach.” in *ACM SIGWEB*. New York, NY: ACM Press.

Stallman, Richard. 1999. “The GNU Operating System and the Free Software Movement.” *Open sources: Voices from the open source revolution* 1:280.

Steiner, Ivan D. 1966. “Models for Inferring Relationships between Group Size and Potential Group Productivity.” *Behavioral Science* 11(4):273–83.

Stewart, Daniel. 2005. “Social Status in an Open-Source Community.” *American Sociological Review* 70(5):823–42.

Stinchcombe, Arthur L. 1965. “Social Structure and Organizations.” *Handbook of organizations* 142:193.

Stvilia, Besiki, and Michael B. Twidale. 2008. “Information Quality Work Organization in Wikipedia.” *Journal of the American Society for Information Science and Technology* 59(6):983–1001.

Suh, Bongwon, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. 2009. “The Singularity Is Not near.” P. 1 in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration - WikiSym ’09*. New York, New York, USA: ACM Press.

Sumi, Robert, Taha Yasseri, Andr&#x0B4;s Rung, Andr&#x0B4;s Kornai, and J&#x0B4;nos Kertesz. 2011. “Edit Wars in Wikipedia.” *2011 IEEE Third Int’l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing* 724–27.

Sunstein, Cass R. 2006. *Infotopia: How Many Minds Produce Knowledge*. New York: Oxford University Press.

Surowiecki, J. 2005. *The Wisdom of Crowds*. New York: Random House.

Sussman, N. M., and D. H. Tyson. 2000. “Sex and Power: Gender Differences in Computer-Mediated Interactions.” *Computers in Human Behavior* 16(4):381–94.

Tabachnick, Barbara G., and Linda S. Fidell. 2012. *Using Multivariate Statistics*. 6th ed. New York: Pearson.

Taniguchi, H. 2006. “Men’s and Women's Volunteering: Gender Differences in the Effects of Employment and Family Characteristics.” *Nonprofit and Voluntary Sector Quarterly* 35(1):83–101.

Tapscott, Don, and Anthony D. Williams. 2008. *Wikinomics: How Mass Collaboration Changes Everything*. New York, NY: Portfolio.

Thomson, Rob. 2006. “The Effect of Topic of Discussion on Gendered Language in Computer-Mediated Communication Discussion.” *Journal of Language and Social Psychology* 25(2):167–78.

Thursby, Jerry G., and Peter Schmidt. 1977. “Some Properties of Tests for Specification Error in a Linear Regression Model.” *Journal of the American Statistical Association* 72(359):635–41.

Tollefsen, Deborah Perron. 2009. “Wikipedia and the Epistemology of Testimony.” *Episteme* 6(1):8–24.

Travica, Bob. 1999. *New Organizational Designs: Information Aspects*. Stamford, CT: Greenwood Publishing Group.

Tukey, John W. 1949. “One Degree of Freedom for Non-Additivity.” *Biometrics* 5(3):232–242 CR – Copyright &#169; 1949 International .

Twidale, Michael B., Linda C. Smith, Les Gasser, and Besiki Stvilia. 2008. “Information Quality Work Organization in Wikipedia.” *Journal of the American Society for Information Science* 59(6):983–1001.

Viégas, A. B., Wattenberg Martin, Kriss Jesse, and Ham Frank Van. 2007. “Talk Before You Type: Coordination in Wikipedia.”

Viégas, Fernanda B., Martin Wattenberg, and Kushal Dave. 2004. “Studying Cooperation and Conflict between Authors with History Flow Visualizations.” Pp. 575–82 in *Proceedings of the 2004 conference on Human factors in computing systems - CHI ’04*. New York, New York, USA: ACM Press.

Viégas, Fernanda B., Martin Wattenberg, and Matthew M. McKeon. 2007. “The Hidden Order of Wikipedia.” 445–54.

Voss, Jakob. 2005. “Measuring Wikipedia.” in *Proceedings of the 10th international conference of the International Society for Scientometrics and Informetrics*. Stockholm, Sweeden.

Wagner, Clifford H. 1982. “Simpson’s Paradox in Real Life.” *The American Statistician* 36(1):46–48.

Weber, Max. 1922. *Economy and Society: An Outline of Interpretive Sociology (2 Volume Set)*. University of California Press.

Weber, Steven. 2000. “The Political Economy of Open Source Software.”

Weber, Steven, R. Latham, and S. Sassen. 2005. “The Political Economy of Open Source Software and Why It Matters.” *Digital Formations: IT and New Architectures in the Global Realm (Princeton University Press, USA, 2005)* 178–211.

Weitzel, William, and Ellen Jonsson. 1989. “Decline in Organizations: A Literature Integration and Extension.” *Administrative Science Quarterly* 91–109.

West, Candice, and Don H. Zimmerman. 1987. “Doing Gender.” *Gender & Society* 1(2):125–51.

West, David, and Scott Dellana. 2009. “Diversity of Ability and Cognitive Style for Group Decision Processes.” *Information Sciences* 179(5):542–58.

Wexler, Mark N. 2011. “Reconfiguring the Sociology of the Crowd: Exploring Crowdsourcing.” *International Journal of Sociology and Social Policy* 31(1/2):6–20.

Whetten, David A. 1980. “Organizational Decline: A Neglected Topic in Organizational Science1.” *Academy of Management review* 5(4):577–88.

Whetten, David A. 1987. “Organizational Growth and Decline Processes.” *Annual review of sociology* 335–58.

Whittaker, Steve, Loren Terveen, Will Hill, and Lynn Cherny. 1998. “The Dynamics of Mass Interaction.” Pp. 257–64 in *Proceedings of the 1998 ACM conference on Computer supported cooperative work - CSCW ’98*. New York, New York, USA: ACM Press.

Wikimedia Foundation. 2010. *Former Contributors Survey Results*.

Wikimedia Foundation (WMF). 2011. “Wikipedia Editors Study: Results from the Editor Survey, April 2011.” Retrieved August 1, 2011 (http://upload.wikimedia.org/wikipedia/commons/5/51/Editor_Survey_Report_April_2011.pdf).

Wilkinson, Dennis M., and Bernardo A. Huberman. 2007. “Assessing the Value of Cooperation in Wikipedia.” Pp. 157–64 in *Proceedings of the 2007 international symposium on Wikis*, vol. 12. ACM.

Willer, Robb. 2009. “Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem.” *American Sociological Review* 74(1):23–43.

Williams, K., and C. O’Reilly. 1998. “The Complexity of Diversity: A Review of Forty Years of Research.” Pp. 77–140 in *Research in Organizational Behavior.*, vol. 20, edited by B and Sutton Staw R. Greenwich, CT.: JAI Press.

Williams, R. L., and J. Cothrel. 2000. “Four Smart Ways to Run Online Communities.” *Sloan Management Review* 41(4):81–92.

Williamson, Oliver E. 1979. “Transaction-Cost Economics: The Governance of Contractual Relations.” *JL & Econ.* 22:233.

Williamson, Oliver E. 1981. “The Economics of Organization: The Transaction Cost Approach.” *American journal of sociology* 548–77.

Williamson, Oliver E. 1998. “Transaction Cost Economics: How It Works; Where It Is Headed.” *De economist* 146(1):23–58.

Wilson, John. 2000. “Volunteering.” *Annual Review of Sociology* 26(1):215–40.

Woodman, R. W., J. E. Sawyer, and R. W. Griffin. 1993. “Toward a Theory of Organizational Creativity.” *Academy of Management Review* 18(2):293–321.

Woolridge, Jeffrey, and Jeffrey M. Wooldridge. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wuthnow, Robert. 1998. *Loose Connections: Joining Together in America’s Fragmented Communities*. Cambridge, MA: Harvard University Press.

Yasseri, Taha, Robert Sumi, and János Kertész. 2012. “Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis” edited by Attila Szolnoki. *PloS one* 7(1):e30091.

Ye, Yunwen, and Kouichi Kishida. 2003. “Toward an Understanding of the Motivation of Open Source Software Developers.” Pp. 419–29 in *Software Engineering, 2003. Proceedings. 25th International Conference on*. IEEE.

Ye, Yunwen, Yasuhiro Yamamoto, and Kouichi Kishida. 2004. “Dynamic Community: A New Conceptual Framework for Supporting Knowledge Collaboration in Software Development.” Pp. 472–81 in *Software Engineering Conference, 2004. 11th Asia-Pacific*. IEEE.

Zeitlyn, David. 2003. “Gift Economies in the Development of Open Source Software: Anthropological Reflections.” *Research policy* 32(7):1287–91.

Zerubavel, Eviatar. 1993. *The Fine Line: Making Distinctions in Everyday Life*. Chicago: University of Chicago Press.

Zhao, Wenjia. 2012. “The Wikipedia Bureaucracy - Forbes.” *Forbes*. Retrieved March 9, 2014 (http://www.forbes.com/sites/wenjiazhao/2012/07/23/the-wikipedia-bureaucracy/).

Zhu, Haiyi, Robert E. Kraut, and Aniket Kittur. 2013. “Effects of Peer Feedback on Contribution: A Field Experiment in Wikipedia.” in *CHI'2013: Proceedings of the 2013 annual conference on Human factors in computing systems*. New York: ACM Press.

Zickuhr, Kathryn, and Lee Rainie. 2010. “Wikipedia, Past and Present.” *Pew Internet & American Life Project*