Ecology of Online Communities

Nathan TeBlunthuis

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Benjamin Mako Hill

David McDonald

Kirsten Foot

Aaron Shaw

Program Authorized to Offer Degree:

Communication

University of Washington

Abstract

Ecology of Online Communities

Nathan TeBlunthuis

Chair of the Supervisory Committee:
Benjamin Mako Hill
Communication

How do competitive struggles for resources or symbiotic relationships that support a web of interdependent communities shape the evolution of online organizing? Most prior studies of online community success have focused almost exclusively on communities' internal features, but in biology and organization studies, ecological approaches have shown that success is largely---and sometimes overwhelmingly---a function of what others groups are doing. This dissertation contributes to the fields of Human Computer Interaction, Social Computing and Communication an ecological analysis that accounts for the complex dynamic interactions between communities and their environments and is important for understanding the successes and failures of online communities.

The theoretical foundations are in organizational ecology, a vast social scientific literature that applies ecology to human organizations. Online communities are very different from classical

organizations, so this investigation required empirically validating basic assumptions about when online communities will be competitive or mutualistic. It uses linear and nonlinear time series analysis of clusters of online communities to shows that mutualistic relationships are more common than competitive ones.

Interviews with members of overlapping online communities empirically support and explain this widespread mutualism in terms of tensions between the types of benefits of participating in online communities. Instead of resolving these tensions using a complex organization that provides the full range benefits, people build multiple relatively simple communities each specializing in a subset of benefits. Designers of platforms for online communities should cultivate ecosystems of overlapping and differentiated communities by supporting resource sharing and simultaneous participation in many communities.

Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to the many academic friends, colleagues, and mentors who have cultivated my intellectual development, helped me work on these ideas, and in every other way made possible my success. In particular I would like to thank members of the Community Data Science Collective and Aaron Shaw, Sohyeon Hwang, Jeremy Foote, Carl Colglazier, Floor Fiers, Sejal Khatri, Sefania Druga, Nicholas Vincent, and Kaylea Champion in particular for their helpful feedback on parts of this work. Also thanks to Mako and Aaron for their innovation, dedication and care in organizing this very special research group. I am also grateful to my collaborators I have not yet mentioned: Isabella Brown, Laura (Alia) Levi, Nicole McGinnis, Tilman Bayer, Olga Vasileva, and Aaron Halfaker. Special thanks to Daryn McElroy for her work to externally validate our clusters. Thanks to Mark Kott for his excellent course on mathematical ecology which inspired an important turning point in the direction of this work and to Carmen Gonzalez and Matthew Powers' for their fantastic course on fieldwork research methods. The importance of this education in qualitative research to this work suprised me, but I doubt it would suprise them. I am also grateful to the organizers and participants in the social computing reading group (SCRG) at the University of Washington. My participation in this reading group has been invaluable to any ability I have to make contributions to social computing or HCI. I owe special gratitude to my 20 interview participants for their time and knowledge. I am thankful to the organizers and members of UAW Local 4121 for their strength and solidarity. Thanks to Jason Baumgartner and pushshift.io for the Reddit data archive. This work was made possible by generous financial support from the National science foundation grants IIS-1908850 and IIS-1910202 and GRFP2016220885 and was facilitated through the use of the advanced computational infrastructure provided by the Hyak supercomputer system at the University of Washington.

Dedication

To Amanda, my dear full mutualist.

Preface to Chapter 1

Several paragraphs in beginning of the following chapter adapt from text I wrote for a grant proposal submitted to the National Science Foundation (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1910202, 1910202)

Chapter 1

Introduction: An Ecology of Digital Affiliation

Would Wikipedia be one of the most visited websites in the world if other online collaborative encyclopedia projects had been more established when it was founded? Or was Wikipedia helped by the fact that its predecessors had engaged and trained hundreds of its future contributors? Do new discussion communities on Reddit compete with existing communities for contributors? Is the evolving world of online communities better understood as a competitive struggle for resources or as symbiotic relationships that support a web of interdependent communities? How does the environment of existing online communities shape the growth, performance, and impact of new groups?

Answering these questions requires an *ecological understanding* of online communities that accounts for the complex dynamic interactions between communities and their environments. Prior studies of the growth, survival, and success of online communities have focused almost exclusively on communities' internal features (Kraut et al., 2012) and have largely neglected environmental factors (e.g., Halfaker et al., 2013; Kraut et al., 2012; Schweik & English, 2012; Shaw & Hill, 2014; TeBlunthuis et al., 2018). Analyses from this "focal organization perspective" (Hannan & Freeman, 1989) typically account for only a small amount of variation in communities' growth, longevity, and performance. Ecology provides a compelling alternative theoretical approach. In biology and organization studies, ecological approaches have shown that success is largely—and sometimes overwhelming—a function of what others groups are doing (Hannan & Freeman, 1989; Worster, 1994).

Ecology is a scientific approach to understanding how interdependence between individuals, collectives, and environments shapes the world (Worster, 1994). Although first developed to understand biological ecosystems, ecology's theories and methods influenced the development of human ecology, and later of organizational ecology (Hannan & Freeman, 1989; J. M. McPherson, 1983; Park, 1936). Organizational ecology is a vast field in social science that explains the success, failure, and evolution of newspapers, microbreweries, social movements, and voluntary organizations (Carroll, 1985; Carroll & Swaminathan, 2000; J. M. McPherson, 1983; Soule & King, 2008). Ecology can provide practical solutions to problems in complex systems like effective wildlife management, pest control, and sustainable utilization of renewable resources. In organization science, it provides compelling explanations for industrial life-cycles, organizational specialization, and patterns of collaborative partnerships.

Recent research in the social computing on interdependence between online communities suggests that ecological analyses can provide not only novel scientific understandings but also viable community management strategies (Chandrasekharan et al., 2017; Kiene et al., 2018; Tan, 2018; TeBlunthuis et al., 2017; Vincent et al., 2018; X. Wang et al., 2012; Zhu, Kraut, et al., 2014). For example Chandrasekharan et al. (2017) found evidence that banning hateful communities on Reddit decreased hate speech in related communities. Community outcomes such as growth and survival depend on membership overlaps between communities (X. Wang et al., 2012; Zhu, Kraut, et al., 2014), but the nature of the resulting relationships remains unclear. X. Wang et al. (2012) found that participant overlaps between Usenet groups were associated with *competition* and decreased participation in both communities. However, Zhu, Kraut, et al. (2014) found evidence that membership overlap between wikis is associated with *mutualism* and benefits for both communities. Such contradictory findings point to the need for deeper, more precise theories of how ecological dynamics play out in online communities.

Online communities are a dynamic, growing, and increasingly important form of organization that enable collaboration on public goods in contrast to the private goods production most studied in organizational ecology (Benkler et al., 2015). Through peer production, the Wikipedia community has produced the largest collaborative effort and most important reference work in human history. Free/libre open source software (FLOSS) communities have produced tens of billions of dollars worth of software made freely available online (Benkler et al., 2015). Other online communities like subreddits provide information, social support, and entertainment to millions of people. Ecological research into online communities may enable us to understand *why* and *how* of the millions of attempts to build communities, only a tiny percentage of manage to mobilize participants and to sustain collaboration (Hill & Shaw, 2019; Schweik & English, 2012; Shirky, 2008). However, online communities are vastly different from the organizations organizational ecology was developed to study. Classical hypotheses in organizational ecology are built on a system of interlinked assumptions that were informed by background knowledge of 20th century organizations. I argue that past applications of organizational ecology to online communities have not anticipated how this change in context could lead to changes in theoretical predictions.

Therefore, I do not pick up organizational ecology as an authoritative model or set of laws capable of explaining the growth and decline of online communities. Instead I drew some ideas directly from mathematical ecology, a subfield of applied mathematics, to better understand the foundational assumptions of an ecological perspective. On this foundation, I see this project as building an empirical basis for an ecological theory of online communities that starts by inferring competitive and mutualistic relationships between online communities. Once ecological dynamics between communities are demonstrated to have measurable relationships with the growth and performance of online communities, we can more fully explain their origins and consequences.

My empirical studies are framed in terms of how the ecological approach provides new insights and ways of studying interdependent online communities. However, these studies' methodological designs and empirical results also contribute to organizational ecology by expanding its application beyond the scope of its founding assumptions. The developers of organizational ecology developed strong intuitions about when organizations will complement or compete with one another based on claims from prior organization theory including that organizations compete over resources, are shaped into established forms by homogenizing pres-

sures, are defined by strong boundaries, and lack capacities for rational adaptation. Because they typically lacked sufficient longitudinal data to infer when organizations are competitors or mutualists, they have rarely tested these assumptions directly, but rather test theoretical predictions about outcomes like organizational formation, survival, and change (Baum & Shipilov, 2006; Hannan & Freeman, 1989). The literature on interdependence between online communities is relatively young and provides less background knowledge that can inform such assumptions, but data from online communities enables a stronger empirical basis for understanding relationships between groups.

Although the time series models in Chapters 2 and 4 depend on fewer assumptions about when competition or mutualism occur compared to the most influential frameworks of organizational ecology, Chapter 2's key finding, that mutualism is more common than competition among online groups with highly overlapping users, radically departs from organizational ecology which has found that both firms and voluntary organizations with highly overlapping resources typically compete (Hannan & Freeman, 1989; J. M. McPherson, 1983). Although I initially planned to continue developing model-based approaches to explaining the performance of online communities, to find widespread mutualism was surprising and demanded qualitative validation and explanation in terms of the experiences of online community members. Therefore, Chapter 3 reports on an interview-based study of members of highly overlapping online communities. It concludes that "no community can do everything" because groups of overlapping communities are characterized by high degrees of specialization. Each community seems to provide a different set of benefits. As Chapter 3 discusses, this is consistent with ecological theory which suggests that highly specialized groups with overlapping memberships are unlikely to compete and that groups provide complementary benefits that can "spill over" and drive mutualistic dynamics.

Knowledge from interviewees includes invaluable cases of mutualism, grounded descriptions of relationships between overlapping online communities, a strong sense that Chapter 2's models are right about the ubiquity of mutualism, and clues about the importance of specialization. However, the interviewees did not provide much to explain processes by which systems of specialized mutualistic overlapping communities develop. In Chapter 4, I draw from strands of organizational ecology that use evolutionary theory as a foundation for processes of change. In addition, the models from Chapter 2 are effectively the most simple time series models that might be used to infer ecological interactions. They depend on many assumptions that are probably unrealistic in the setting of online communities. Therefore, Chapter 4 adopts nonlinear time series models developed by mathematical ecologists to study nonlinear dynamics. These models are important to Chapter 4's study design for investigating change processes and also compel us to conceptualize competition and mutualism interactions that are not static and fixed, but that vary over time.

In sum, online communities are a kind of organization, at least in the sense that organizations are "constructed as tools for specific kinds of collective action" (Hannan & Freeman, 1989). Even when online communities are constructed to facilitate communication with strangers on the internet about a topic, this facilitation depends on the sustained contributions of members to keep the conversation going and structures for regulate behavior to maintain a suitable conversation space (Kraut et al., 2012). Online communities bear other similarities to organizations including their use of formalized roles, rules, and procedures and their use of boundaries defining the scope of activity (Foote, 2019).

That said, online communities are distinctive in that they are public-good producing voluntary groups constructed through computer-mediated communication. Features of online communities depart in important ways from the types of organizations that classical organizational ecology has studied the most. Online communities (1) are dependent on volunteer participation, (2) allow participation at very low levels of granularity (3) are weakly bounded and (4) face different potential sources of inertia. The remainder of this chapter discusses the methodological and theoretical implications of these interrelated features for ecological analysis.

1.1. Online Communities as Voluntary Organizations

Ecological theories conceive of dynamics among individuals that share resources needed for production and survival to explain change in the size and composition of groups over time. Organizational ecology explains macro-level social change in economies and industries through in an "evolutionary" style through mechanisms of the selection and adaptation of firms in a changing resource environment (Hannan & Freeman, 1989; Ven & Poole, 1995). Each organization's survival depends on its *niche* in the resource environment. The notion of a niche is central, if sometimes slippery, and aims to capture the position of an organization in an abstract, high-dimensional resource space (Hannan & Freeman, 1989). Organizational ecology was developed mainly to study commercial firms whose survival ultimately depends on their potential to offer returns on investment. Profitability of these firms typically hinged on expansion to control greater quantities of resources, provide economies of scale and create the potential for monopolistic rents (Hannan & Freeman, 1989). Niches for such organizations are often defined in terms of established categories of organizational forms (Carroll & Swaminathan, 2000), technological production factors (Dobrev et al., 2001), or economic outputs (Dobrev et al., 2003).

How should we define niches for online communities? They use the low-cost communication systems of the Internet to coordinate voluntary production of public information goods like encyclopedias, FLOSS programs, and cultural artifacts (Benkler, 2006). An online community might produce something damaging to the broader society, such as computer viruses or misinformation, but the types of online communities considered here produce public goods defined as *non-excludible* (in principle, an individual cannot be excluded from utilizing them) and *non-rival* (utilization does not diminish the good's value). Therefore, the survival of online communities depends not on capacities to generate revenues and capture profits, but on the consistent participation of volunteer members who have heterogeneous motivations for contributing to a public good (Lampe et al., 2010; Shah, 2006).

Dependence on volunteer members is something online communities have in common with voluntary organizations like social clubs, churches, or fraternal organizations (Bimber et al., 2012). Voluntary organizations have been studied in organizational ecology by J. Miller McPherson and collaborators who investigate overlapping niches defined by organizational members and associated demographic patterns (J. M. McPherson, 1983; J. M. McPherson & Ranger-Moore, 1991; J. M. McPherson & Rotolo, 1996; Popielarz & McPherson, 1995). For example, Popielarz and McPherson (1995) locate voluntary organizations' niches in "Blau Space" corresponding to the distribution of their members' demographic characteristics and explain how voluntary organizations tended to become racially or educationally homogeneous in terms of competitive dynamics over members' time and attention (Popielarz & McPherson, 1995). Similar to McPherson, ecological studies of online communities, including the present work, have

defined niches of online communities in terms of their participants (X. Wang et al., 2012; Zhu, Kraut, et al., 2014).

However, membership is not the only plausible way to define an online community's niche. As a consequence of their nature as public-good producing voluntary organizations, their survival does not depend on expansion. Although influential models of the growth of online communities have assumed that motivations to participate in online communities increase as communities grow (Butler, 2001; Kraut et al., 2012), recent surveys and interviews find that large and small communities provide different sorts of benefits (Foote et al., 2017; Hwang & Foote, 2021). As Chapter 3 finds, larger communities provide steady streams of content and larger potential audiences, but are less capable of providing tight-knit socialization or specialized information.

This kind of size-dependent specialization resembles "niche-width" arguments in organizational ecology. For example, Carroll (1985) seeks to explain the coexistence of large and small organizations within an industry by proposing that generalists, who have wide niches, underperform in certain areas of the resource space. Smaller organizations can exploit this under performance by specializing in these areas. However, as Dobrev et al. (2001) argue, specialist organizations can grow large in certain circumstances and then organizational size can be uncorrelated with niche width. This is the case with online communities. For example the subreddit `r/prequelmemes` is dedicated to making and sharing memes only about the Star Wars and is the largest Star Wars related community on Reddit. Therefore, it important to recognize that memberships may not capture all the relevant dimensions of an online community's niche. Indeed, Chapter 3 finds at least three dimensions of specialization in terms of the benefits that members obtain from online communities. These are (1) access to the largest possible audience, (2) socialization in a homophilous community and (3) ability to find specialized content or information.

Still, for the purposes of the studies in Chapters 2 and 4, membership overlaps provide a number of advantages. The benefits of participation may not be easily observed, so measuring online community niches in terms of participation, which is observable, is empirically tractable. Furthermore, findings in Chapter 3 suggest that community leaders do not normally seek to appropriate private value from their communities. If so, then it seems more likely that ecological dynamics that shape the growth and survival of online groups will have more to do with participation, the main rival resource on which online communities depend. Finally, studies in organizational ecology have set out to test models that depend on linear or curvilinear relationships between niche-overlap and competitive pressures and this required stronger assumptions around the measurement of niche width than those needed here (Carroll, 1985; Dobrev et al., 2003). Chapters 2 and 4 use membership overlaps to identify clusters of highly related communities while time-series models are used to infer competition and mutualism. These models bear their own assumptions, but the threat to scientific validity moves from the task of measurement to the task of statistical inference. Chapter 5 discusses how expanding definitions of an online community's niche to account for additional dimensions of specialization will be important for future work.

1.2. Openness Allows Dividing Time into Little Chunks

Although following McPherson's use of membership-based niches makes sense because online communities depend on voluntary contributions to produce public goods, a second key feature

of online communities departs from the voluntary organizations in McPherson's studies. This is that online communities provide opportunities for "tiny acts of participation" like signing a petition, fixing a typo on Wikipedia, or "liking" a post. When individuals can act in small granular ways they can easily participate in many online communities in rapid succession (Benkler, 2006; Margetts et al., 2015; Tan & Lee, 2015). By contrast, McPherson assumes that organizations conduct their activities in face-to-face in-person meetings and theorizes that constraints of time and space strongly limit the number of organizations to which an individual can belong (J. M. McPherson, 1983). After work and other obligations, it seems unlikely that many people would have time to belong to very many voluntary organizations at once, so participation in an organization is highly rival and overlaps in membership are tightly coupled with competition.

Chapter 2 and prior ecological studies of online community participation follow this intuition by considering membership to be a rival resource, and assuming that online communities with overlapping users are those likely to have significant ecological interactions (Butler, 2001; X. Wang et al., 2012). However Chapter 2 avoids assuming that these interactions will be competitive and instead finds that mutualism among highly related online communities is about 4 times as common as competition and in Chapter 3 interviewees described how these related communities have specialized roles. Together, these findings suggest that the growth and survival of a sufficiently established community is not often limited by competition over membership. Why is membership overlap so strongly associated with competition in the context of in-person voluntary organizations but highly overlapping online communities are often mutualists?

Online communities "transcend time and space" using asynchronous and low-cost telecommunications (Jarvenpaa & Leidner, 1998; Peters, 1999). Although individuals are fundamentally constrained in their available time and energy, they can finely divide their time over many communities. Margetts et al. (2015) suggest this less "lumpy" form of participation helps enable online collective action. Similarly, the fine-grained division of individuals' activities across communities is closely related to the success of online communities having "open" organizations with minimal barriers to participation. Benkler claims that the fact that information is nonrival is central to how online communities successfully peer-produce public information goods. This characteristic of information goods also enables open organizational structures so that peer-production projects can incorporate contributions from peripheral contributors (Benkler, 2006; Bryant et al., 2005). Together, these factors allow levels of participation that are even more unequal than those found in other voluntary organizations. For example, while "the top 20% of volunteering individuals contributed 50% of the time volunteered in the USA" in 2016, the top 1% of Wikipedia editors put in 77% of the effort into editing Wikipedia (Matei & Britt, 2017).

When people can spread their time across many open communities, this also shapes the nature of membership in a community and the boundaries between communities. Organizational ecology was developed with the relatively impermeable boundaries of commercial organizations in mind (Hannan & Freeman, 1989). This is a second reason why J. M. McPherson's studies of voluntary organizations provide a good model for studying ecology of online communities. While commercial firms have relatively strong boundaries around internal activities and control over much of their employees time, voluntary organizations open up more of their activities to outsiders in order to attract participants. As noted above, J. M. McPherson assumes that voluntary organizations with overlapping niches will compete. However, mathematical ecology shows that niche overlaps do not necessarily imply competition in complex systems involving multiple organizations or resource dimensions because factors other than the overlapping re-

sources can limit growth (Armstrong & McGehee, 1980).

Finally, by modeling community size as the "tiny act of participation" of commenting in a given week, the analysis of ecological dynamics in Chapters 2 and 4 might be predisposed to find mutualism. Although quantifying time spent on contributions might not be possible in the case of Reddit (how would we count the time someone spends creating art to share with an online community?), it is possible that a study of participation intensity might find weaker mutualism and stronger competition if small contributions from peripheral members are less rival than larger contributions from core members. On the other hand, if these contributions take the form of non-rival information goods, then communities will be unlikely to compete over them (an artist is likely to share their effortful creations with all communities from which they desire an audience). The findings of Chapters 2 and 3 both suggest that part of why subreddits with overlapping memberships can provide complementary benefits and form mutualistic ecological relationships is that membership in multiple online communities is relatively inexpensive. If subreddits became closed organizations, perhaps by introducing pricey membership fees, one would expect stronger competition over membership. In this way, openness appears to provide conditions less conducive to competition and more conducive to mutualism.

1.3. How Should Online Communities be Divided into Organizational Forms?

Related to openness and the predominance of mutualism is Chapter 3's finding of extensive specialization among online communities that have similar topics and similar members. One rarely observes more than one active subreddit with similar topics that is not differentiated in some significant way, often in size, rules or topic. Groups of related online communities thus depart from the organizational forms studied in organizational ecology in ways that trouble the specific strands of organization ecology used by prior research on online communities.

Chapter 2 defines its approach as community ecology because it focuses on relationships between different online communities. This may surprise readers of the organizational ecology literature in sociology which defines community ecology as the study interactions between populations of organizations, but I argue it is reasonable given the heterogeneity of overlapping online communities. I will also note that studies in Communication have applied the community ecology approach to study competition and mutualism between telecommunication companies (Barnett & Carroll, 1987; Dimmick & Rothenbuhler, 1984) or networks of organizational relationships (Dimmick & Rothenbuhler, 1984; Margolin et al., 2012). However, such studies are a small minority in the literature.

Aldrich and Ruef (2006), Hannan and Freeman (1989), and Astley (1985) all consider community ecology as having a distinct level of analysis from population ecology. They use levels of abstraction analogous to those used in biological ecology where a population is set of individual organisms of the same species and a community is a set of interacting populations. For these organizational ecologists, a population is a set of organizations having the same *organizational form* and a community corresponds to an *organizational field* of related organizational forms.

The identification of an organizational form is of central importance. Both organizational and mathematical ecologists are aware that population ecology models like density dependence depend on the assumption that the population under study is homogeneous in the sense all members of the population are equally subject to the same intra-population mutualistic and competitive forces. Organizational ecologists have justified these assumptions by carefully demarcating

different types of organizations into organizational forms theorizing that discrete boundaries around organizational forms are constructed by homogenizing features like efficient ways to bundle transactions (Williamson, 1981), external regulatory frameworks, or other mechanisms of institutional isomorphism (P. J. DiMaggio & Powell, 1983; Hannan & Freeman, 1989). Still, the definition of organizational forms in organizational ecology often amounts to accepting an established categorization. The fascinating question of how the processes by which such categorizations are socially constructed are related to the ecological dynamics within and between organizational forms has driven much work by Hannan and his collaborators in recent years (Hannan, 2019; Hannan et al., 2007; Pontikes & Hannan, 2014).

Although McPherson's series of papers on the ecology of voluntary organizations may best be described as a community ecology analysis of categories of voluntary organizations like "sports" or "youth serving" organizations, at times he resists analogizing organizations as biological populations: "A population of organizations, then, is not a set of discrete creatures who must mate with each other to reproduce, but a froth of bubbles, constantly sharing or exchanging members, growing and dying, and being absorbed and segmented in response to changing conditions" (J. M. McPherson, 1983). In this instance as well as others, McPherson's papers sometimes slip from discussing ecological dynamics among different organizational forms, which is measured in the data, and between different organizations, which is not. In the above quote, McPherson clearly has a dynamic ecosystem of differentiated organizations in mind. Perhaps the set of "sports" organizations contains too much heterogeneity to constitute an organizational form.

Later organizational ecologists studied diversity within an organizational population by appealing to a distinction between "core" features which define the organizational form and are mostly stable over time and "peripheral" features which are allowed to vary (Hannan & Freeman, 1989). Organizational ecologists have studied how variation and specialization of peripheral features shapes competition within an organizational form. For example, Dobrev et al. (2003) studies how degrees of overlap among automotive firms' technological niches, measured as engine horsepower, changed over time and affected organizational survival. Similarly, Chapter 2 and prior ecological studies of online communities measure user overlap density to quantify how much a community's members participate in other communities (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). Chapter 4 takes this a step further by studying how dynamically shifting niches are related to competitive and mutualistic interactions.

Organizational forms of online communities might be defined according to the platform hosting them. Indeed, prior ecological studies of online communities have done exactly this and treated sets of communities sharing a platform like Usenet or Wikia as a population. However, technological boundaries around platforms may not ensure sufficient homogeneity to justify treating these sets of communities as an organizational form. One finds enormous diversity in the topics and purposes of communities upon exploring a platform like Reddit, Facebook Groups, or Wikia. Chapter 3 finds that, even when topics and memberships are very similar, online communities are specialized in other dimensions. Although a platform clearly provides a set of common technological affordances, many platforms are flexible enough to allow a great deal of diversity in scopes, rules, and communities can greatly expand available affordances by using auxiliary technologies like bots (Kiene et al., 2019). It is thus questionable that overlapping features of online communities like memberships or topics are "peripheral" while the use of a platform is "core" and therefore it is difficult to identify populations of online communities *a*

priori.

When categorizations of organizations of interest are not well-understood, Hannan and Freeman (1989) recommend using numerical clustering to find divisions of organizational forms. The quantitative analyses in Chapters 2 and 4 are all based on a clustering algorithm that groups subreddits with similar kinds of users. I define these as "ecological communities" in a way that is consistent with the sense of Aldrich and Reuf, although they are interested in competition and mutualism between organizational forms. However, as Chapter 3 demonstrates, this results not in clusters of online communities having similar forms, but in groups of subreddits whose topics are related but whose forms vary along dimensions of scope, size, and internal structures like rules. Population ecology is designed to study the mutualistic and competitive processes among members of an organizational form. Community ecology is designed to study mutualism and competition between populations of organizations having different forms. Neither theory seems to fit exactly with subreddits, but Chapter 2 can be understood as advancing a community ecology analysis of organizational forms assumed to have a single member organization. If this seems overly nuanced, one can simply adopt the framing of Chapter 2 and ignore matters of organizational forms and fields and treat community ecology as a relational framework and population ecology as an environmental framework.

1.4. Inertia and Adaptation

Organizational ecologists have tended to emphasize selection processes because organizational cores appear to change relatively little. External homogenizing forces described and by internal factors like culture and routines that are difficult to change lead to "structural inertia." Structural inertia limits an organization's ability to rationally adapt to a changing environment. Organizations typically lack sufficient information about their environments and the ability to coordinate change with sufficient precision in order to rationally adapt, especially when it comes to change in the "core" aspects of an organization (Hannan & Freeman, 1984). However, they also experience exceptional transformational periods that accompany an increased risk of failure (Aldrich & Ruef, 2006). If organizations are adaptive, then a teleological or functionalist explanation of organizational change may be better than an ecological one (Ven & Poole, 1995) and theories of change in organizational fields should be based on Lamarkian adaptation-based evolution instead of Darwinian selection.

Whether online communities can adapt has important consequences for design interventions aimed at improving the quality or safety of online spaces. Adaptive online communities may adopt new tools for moderation or quality control or implement policy changes to address newly uncovered problems. But online communities having substantial structural inertia will struggle to adapt, problems that go unaddressed will contribute to communities' declines, and solutions will largely emerge through the construction of new communities. A selection-based change process may be slower than an adaptive one because it will be limited by rates of community formation and decline.

Prior research into online communities suggests a relatively high degree of structural inertia, at least when it comes to policy (Halfaker et al., 2013; TeBlunthuis et al., 2018), but the origins of this inertia are not obvious. One explanation looks to the composition of contributors to an online community and sees social barriers to diverse newcomers as limiting capacities for change (Lam et al., 2011; Menking et al., 2019; Tripodi, 2021). Another explanation is the

entrenchment of oligarchical leadership (Shaw & Hill, 2014), who may be conservative and resist change. Yet in classical organizations, leaders often seek purposeful adaptation, but are foiled by internal sources of inertia like organizational cultural, internal patronage networks, conflicts among stakeholders, and routines (Hannan & Freeman, 1984; Ven & Poole, 1995). Some of these inertial forces appear to have analogs in online communities such as the stability of emergent roles (Arazy et al., 2017; Arazy et al., 2015), routines (Keegan et al., 2016), and internal conflict that may stabilize policy (Shi et al., 2019).

Chapter 4 explore the relationship between ecological dynamics and adaptive processes in online communities by relaxing assumptions of the model in Chapter 2 to allow ecological interactions between online communities vary over time. This allows us to explain that mutualism is more common than competition in Chapter 2 because periods of mutualistic interaction last longer than periods of competitive interaction. Finding that competitive and mutualistic dynamics in online communities are not static, but dynamic and vary over time sets up hypotheses tests about how online communities might adapt to avoid competition or increase mutualism. While I find evidence that communities increase their specialization by decreasing their user and topic overlaps in competitive conditions, I do not find that this decreases competition and increases mutualism. This suggests that variations in competitive and mutualistic dynamics are driven by exogenous events and that at least when it comes to positioning themselves with respect to one another, that successful online communities have "selected an effective niche" (Zhu, Chen, et al., 2014). As discussed further in Chapter 5, the evidence from Chapters 4 does not support strong claims about whether mutualism is common because of adaptation or selection. Future work should seek to demonstrate the selection process in action.

1.5. Conclusion: Contributions to Organizational Ecology

Organizational ecology began by asking "Why are there so many kinds of organizations?" (Hannan & Freeman, 1977, 1989). It provides a conceptual model of how people build systems of interdependent social structures within organizational fields, and a vast and rich literature that was initially developed to study firms in long-running commercial industries. Although Hannan and Freeman (1989) account for the demography of industrial unions in their theory, these unions had key characteristics in common with the firms including strong boundaries, pursuit of monopoly, and dependence on institutional legitimacy. In general, they had their ideological and historical origins in the age of bureaucratic rationalism (Hannan & Freeman, 1989). Theories of organizational ecology have been widely applied to organizations in other contexts, most importantly voluntary organizations and social movements (J. M. McPherson, 1983; Minkoff, 1995; Olzak & Uhrig, 2001; Soule & King, 2008).

The best work of this kind meaningfully adapts organizational ecology to the new context. For example, Soule and King (2008) link organizational ecology to the resource mobilization theory of social movement organizations. Such works use organizational ecology as a "theory of the middle range" that is empirically grounded but has sufficient generality to bridge across multiple domains. However, organizational ecology is not mature paradigm like thermodynamics where models can be treated as "scientific laws" and expected to make accurate predictions about new contexts without any conceptual modification (Kuhn, 1970). As discussed above, some basic concepts of theory, like that of the organizational form, are difficult to apply to online communities. When virtually all organizations in an organizational field are highly distinc-

tive and no established system for categorization can be found, the concept of "organizational form" breaks down and so may the usefulness of distinguishing between the "population" and "community" levels of analysis.

Despite these ontological concerns, as I argue in Chapter 2, density dependence theory's environmental perspective is still useful because the relationship between user overlap density and growth or survival seems to reflect the hospitality of an environment. However, one must keep in mind that tests of density dependence theory in online communities have provided evidence in the form of weak correlations derived from observational data. I suggest that a project to synthesizes foundational concepts from organizational ecology with new empirically supported ideas about the interdependence between online communities will be a more effective strategy.

The most important empirical finding, that mutualism is widespread, is empirically supported by quantitative-qualitative triangulation. Using statistical methods, I have found that mutualism is much more common than competition among subreddits with highly overlapping users. Based on interviews with members of these subreddits, I have found that this widespread mutualism is consistent with their intuitions and I have surfaced a plausible explanation for it in how individuals seek multiple benefits from online communities and that communities with similar topics and overlapping users specialize in providing different types of benefits.

Online communities provide granular longitudinal data of individual behaviors in overlapping groups that make it possible to effectively model and test such propositions. Studies in organizational ecology have generally been limited to one organizational form or organizational field at a time. This has made it difficult to test hypotheses about the scope conditions for ecological dynamics or their consequences. The time series analysis strategies advanced in chapters 2 and 4 make it possible to study ecological interactions on much larger scale, and to justify statements about what kinds of relationships are typical and to model antecedents and consequences of these relationships. It is important to recognize the limits of prior theories and quantitative tools. When results are puzzling or dead-ends are reached, talking to community members is likely to yield insights that open the way toward a solution. The project of this dissertation is to begin reconstructing organizational ecology in the relatively theory-poor but data-rich context of online communities.

Preface to Chapter 2

The following chapter is a collaborative work with Benjamin Mako Hill.
It was honored with a Top Paper award from the Computational Methods Division of the International Communication Association's 2021 annual meeting. An early version of this chapter was presented at the 2020 International Conference for Computational Social Science (IC2S2 2020).

Chapter 2

Identifying Competition and Mutualism Between Online Groups

We introduce a method for inferring competitive and mutualistic interactions between online groups from time series participation data based on the theoretical framework of community ecology. Platforms often host multiple online groups with highly overlapping topics and members. How can researchers and designers understand how interactions between related groups affect measures of group health? Inspired by population ecology, prior social computing research has studied competition and mutualism among related groups by correlating group size with degrees of overlap in content and membership. The resulting body of evidence is puzzling as overlaps seem sometimes to help and other times to hurt. We suggest that this confusion results from aggregating intergroup relationships into an overall environmental effect instead of focusing on networks of competition and mutualism among groups as our approach does. We compare population and community ecology analyses of online community growth by analyzing clusters of subreddits with high user overlap but varying degrees of competition and mutualism.

2.1. Introduction

Although the fact is frequently ignored in social computing scholarship, online groups do not exist in isolation.[1] Indeed, although studying interdependence between online groups is different and complex (Hill & Shaw, 2019), research in social computing has sought to quantify how online groups share users or topics (Datta et al., 2017; Del Tredici & Fernández, 2018; Hessel et al., 2016; Tan & Lee, 2015), and how such interactions relate to outcomes like the emergence of new groups (Tan, 2018), contributions to peer-produced knowledge (Vincent et al., 2018), and the spread of hate speech (Chandrasekharan et al., 2017). Although this work has demonstrated that intergroup interactions matter very little intergroup research has tackled questions of group success—i.e., why some online groups succeed in maintaining active and long-lived participation while most do not. Can intergroup relationships explain whether online groups will grow or decline?

Studies in social computing have drawn from organizational ecology to answer this question (Resnick et al., 2012; X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). Inspired by the ecological study of biological systems, organizational ecology is an in-

[1]We use the term "online group" instead of "online community" to help avoid confusion with our term "community ecology" which plays an important conceptual and analytic role in our paper.

fluential body of theory in sociology that studies competition and mutualism among human organizations Although ecological studies of firms and social movements have developed a clear and established body of theory with strong empirical support (Baum & Shipilov, 2006), similar studies of online groups have yielded inconsistent results that differ both from one context to another and from theoretical predictions. For example, wikis whose memberships overlap with other wikis survived longer (Zhu, Chen, et al., 2014), but Usenet groups with overlapping memberships failed more quickly (X. Wang et al., 2012).

We argue that these confusing results are the result of a conflation of concepts and measures from two distinct strands of theory in organizational ecology: *population ecology* and *community ecology*. Both define competition as a form of interdependence that *decreases* growth and mutualism as one that *increases* growth. However, population ecology focuses on modeling the how overlapping resources among groups affect their subsequent growth, decline, or survival (Astley, 1985; Baum & Shipilov, 2006; Dobrev et al., 2001). It does not attempt to directly study competitive and mutualistic interactions. On the other hand, community ecology recognizes that groups often exist within "ecological communities," or clusters of highly related entities, and provides an approach for inferring competitive and mutualistic interactions among these. Although the stated goal of ecological research in social computing has been to understand how groups influence each others' ability to sustain participation, ecological research in social computing has relied exclusively on concepts and measures from population ecology. This paper seeks to explain the puzzling set of findings in ecological social computing research by introducing community ecology.

We do so in a three-part empirical study using a dataset drawn from the 10,000 communities on Reddit with the most contributors to analyze 641 clusters of online groups with overlapping participants. In Study A, we conduct the most important type of population ecology analysis, a test of what is called density dependence theory, and find support for the theory. This analysis suggests that high degrees of user overlap are associated with competition. In Study B, we introduce our method for community ecology analysis that infers networks of competitive and mutualistic interactions by using clustering analysis and vector autoregression (VAR) models of group size over time (Canova, 2007; Ives et al., 2003; Sims, 1980). We illustrate the method in four case studies and present a large-scale computational analysis showing that mutualistic interactions are far more common than competitive ones. Finally, in Study C, we bring Study A and Study B together to compare population ecology and community ecology by extending the density dependence model from Study A with a variable accounting for competition and mutualism. While we find that adding this variable does not help predict growth, including ecological interactions in our VAR models improves time series forecasting.

We discuss how these findings illuminate the differences between population ecology and community ecology and show how the two perspectives are complementary. While Study A suggests that competition is strongest when user overlap is high, Study B finds widespread mutualism among groups with overlapping membership. Although these findings might seem contradictory, they reflect how population ecology studies overlapping resources related to favorable or unfavorable environmental conditions, while community ecology studies competitive and mutualistic interactions playing out in local networks of specific groups. By demonstrating that mutualistic and competitive interactions within clusters of highly related groups are important—and by describing how to measure them—this paper lays the groundwork for future research to investigate and design for interdependence between online groups that supports their growth

and success.

2.2. Related Work

Online groups are important sites for social support (De Choudhury & De, 2014), entertainment (Ducheneaut et al., 2006), information sharing (Benkler, 2006), and political mobilization of disinformation campaigns and protest movements (Benkler et al., 2013; Choudhury et al., 2016; Krafft & Donovan, 2020). Although an online group's ability to achieve its goals depends on attracting and retaining contributors, few develop a sizable group of participants (Benkler, 2006; P. DiMaggio et al., 2001; S. L. Johnson et al., 2014; Koh et al., 2007; Kraut & Fiore, 2014). Many attempts to explain the success and growth of online groups look to properties of individual groups like characteristics of founders (Kraut & Fiore, 2014), language use (Danescu-Niculescu-Mizil et al., 2013), turnover (Dabbish et al., 2012), and designs for regulating behavior (Halfaker et al., 2013; TeBlunthuis et al., 2018).

Recent research suggests that interdependence among online groups is also important to explain success and failure (Cunha et al., 2019; Kairam et al., 2012; Tan, 2018; Tan & Lee, 2015). For example, banning hate subreddits reduced hate speech in related subreddits (Chandrasekharan et al., 2017). In a very different context, there is evidence that Reddit and Stack Overflow receive substantial benefits from activity on Wikipedia (Vincent et al., 2018). Our work contributes to this literature by providing a new conceptual lens and statistical method for studying competition and mutualism between online groups.

Online Groups Depend on Resources

Like prior ecological research in social computing and information systems, we build on resource dependence theory (RDT) (Butler, 2001; X. Wang et al., 2012). Butler (2001) introduces RDT to argue that growth in online groups is driven by positive feedback as participants contribute resources such as content, information, attention, or social interactions, which motivate further contributions by subsequent participants. That said, online groups do not grow forever and RDT explains that growth is self-limiting because costs of participation increase in larger groups (Butler, 2001; Butler et al., 2014).

Ecological approaches recognize that interrelated online groups may share resources with one another in ways that constrain their growth and survival. *Rival* resources like participants' time, attention, and efforts raise the possibility of competition because they become unavailable to others when used by one group (Benkler, 2006; Kubiszewski et al., 2010; Ostrom & Ostrom, 1977; Romer, 1990). RDT suggests that declines in online participation can be explained in terms of competition over important rival resources (X. Wang et al., 2012).

On the other hand, online groups also rely on *nonrival* resources. They can even produce connective and communal public goods like opportunities to communicate or collections of information (Fulk et al., 1996) which can be "antirival" when their usefulness increases as a result of others using them (Kubiszewski et al., 2010; Weber, 2000). For example, the usefulness of a communication network increases as more people join it (Fulk et al., 1996; Katz & Shapiro, 1985). Similarly, the usefulness of an information good can increase as more people come to know, refer to, and depend upon it (Kubiszewski et al., 2010; Weber, 2000). If multiple online groups help build the same connective or communal public goods, they may form mutualistic

interactions where contributions to one group may "spill over" and motivate participation in mutualist groups (Zhu, Kraut, et al., 2014). Ecological approaches seek to understand how different types of resources will limit or promote growth.

Population Ecology, Density Dependence and Overlapping Resources

While this paper focuses on the ecological study of online groups, other social computing and HCI scholars have used the term "ecology" (and related concepts like "ecoystem" and "environment") to denote an assemblage of sites, devices, or platforms (Nardi & O'Day, 1999; Y. Wang et al., 2015). We use the term more narrowly to refer to conceptual and mathematical models of ecological dynamics. In particular, our work builds on a tradition rooted in *organizational ecology*. First developed in the late 1970s by sociologists studying interactions between firms, organizational ecology was inspired by, and has drawn closely from, ecological studies in biology (Hannan & Freeman, 1977).

Because online groups bear similarities to traditional organizations, organizational ecology provides a compelling theoretical framework for understanding interdependence among online groups. It has inspired at least three high-quality empirical studies of how resources shared by online groups shared shape their growth, decline, or survival (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). These studies draw from the *population ecology* strand of organizational ecology that studies ecological dynamics within a population of groups. In organizational ecology, populations have been defined as sets of organizations sharing an organizational industry or business model (Hannan & Freeman, 1989). In social computing, populations have been defined as online groups sharing a given social media platform (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014).

While population ecology involves several distinct theoretical propositions, *density dependence theory* (DDT) is perhaps the most prominent and is the subject of all three prior ecological studies of online groups (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). DDT models competitive or mutualistic forces in a population of groups as a function of *density* which, in the earliest and most influential studies of DDT, is simply the size of the population. In this way, DDT assumes that every group in the population is facing the same competitive and mutualistic pressures (Aldrich & Ruef, 2006). However, online groups sharing a platform have diverse topics (Kairam et al., 2012), norms (Chandrasekharan et al., 2018; Fiesler et al., 2018), and user bases (Tan & Lee, 2015). Because groups sharing few resources are unlikely to be strongly interdependent, ecological studies of online groups have modeled density dependence based on the concept of *overlap density* (Baum & Shipilov, 2006; Dobrev et al., 2001; X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). Rather than the number of groups that exist in a population, overlap density measures the extent to which an one group's members or topics overlap with all other groups'. Overlap density thus characterizes a group's *niche* or local *resource environment* defined by its distinctive topic and membership.

DDT proposes a model for the growth of organizational populations that has a similar structure to Butler (2001) RDT model for the growth of online groups. In DDT, mutualism is the engine of positive feedback driving population growth. Organizational ecologists show how successful organizations in an emerging industry develop nonrival resources like the legitimacy of a business model or industrial know-how that attract new organizations to enter the market (Carroll & Hannan, 1989; Hannan & Freeman, 1989). Similarly, a population of online groups,

such as those sharing a platform, may grow in size as their platform gains in popularity, as established groups spin off new ones, and as useful knowledge develops that can be shared between groups (Tan, 2018; Zhu, Kraut, et al., 2014).

In RDT, growth of online groups is self-limiting because of the challenges in managing large groups (Butler, 2001). In DDT, competition among population members over rival resources limits growth (Hannan & Freeman, 1989). DDT thus proposes a trade-off in which low density reflects limited opportunities for mutualistic contributions of nonrival resources like legitimacy, connectivity, and knowledge, but high density reflects competition over rival resources. Therefore, DDT predicts that the relationship between density and positive outcomes like growth or survival is ∩-shaped (inverse-U-shaped) (Baum & Shipilov, 2006; Carroll & Hannan, 1989).

Tests of DDT in populations of online groups yield inconsistent results. In X. Wang et al. (2012), user overlap in Usenet newsgroups is associated with decreasing numbers of participants. Similarly, TeBlunthuis et al. (2020) find that topical overlaps between online petitions are negatively associated with participation. By contrast, Zhu, Kraut, et al. (2014) find that membership overlap is positively associated with increasing survival of new Wikia wikis. Only Zhu, Chen, et al. (2014) find support for the ∩-shaped relationship predicted by DDT in an enterprise social media platform.

In Study A, we provide a test of DDT using data from Reddit. The classical logic of DDT appears reasonable in the context of Reddit because low overlap density is likely to reflect an impoverished environment lacking in non-rival resources like skills and knowledge of experienced users, while a group with high overlap is likely to face competition over its members (Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014): ***(H1)*** *The relationship between overlap density and the growth of online groups is ∩-shaped (inverse-U-shaped).*

DDT proposes that very high levels of density will decrease growth because of increasing forces of competition within a niche. However, to conclude that groups with the greatest membership overlap are likely competitors would be to commit a well-known statistical fallacy (Piantadosi et al., 1988; Robinson, 1950). The density of a group's environment suggests that it faces competition or mutualism, but it does not tell us which overlapping communities are competitors and which are mutualists. Community ecology overcomes this limitation of DDT.

Introducing Community Ecology

Perhaps the most natural way to understand the distinction between population ecology and community ecology is in where they believe ecological dynamics like competition and mutualism play out (Astley, 1985). While population ecology locates competition and mutualism within an environmental niche, community ecology locates competition and mutualism in networks of interdependent groups called *ecological communities* (Aldrich & Ruef, 2006). In organizational ecology, this can mean studying interactions between different organizational populations (e.g. J. M. McPherson, 1983; Sørensen, 2004), or networks of interactions between organizations (e.g Margolin et al., 2012; Powell et al., 2005). While varying conceptions of community ecology are found in the organizational ecology literature (J. H. Freeman & Audia, 2006), the approach we describe is identical in structure to that taken by Aldrich and Ruef (2006) and Hawley (1986).

Community ecology focuses on *ecological interactions* (Aldrich & Ruef, 2006). Ecological interactions can be mutualistic when one group has a positive influence on the second such that

growth in the first group leads to growth in the second. They can also be competitive if one group has a negative effect on the second such that growth in the first group leads to decline in the second. Ecological interactions can be reciprocated if mutualism (or competition) from one group to another group is returned in kind. An ecological interaction can also be mutualistic in one direction and competitive in the other. The competitive or mutualistic interactions in an ecological community are quantified by the *community matrix*, a central analytical object in community ecology in both biology and organization science (Aldrich & Ruef, 2006; Novak et al., 2016; Verhoef & Morin, 2010).

In Study B, we demonstrate community ecology by inferring networks of ecological interactions in ecological communities on Reddit. Because our understanding of community ecology theory does not suggest hypotheses about what we will find, we conduct an exploratory data analysis to determine whether mutualism or competition among subreddits is more common on Reddit and present case studies illustrating the types of ecological communities we identify.

Predicting Growth

In Study C we build upon our analyses from Study A and Study B by testing whether community ecology can explain the growth and decline of online groups in ways that population ecology can not. We do this by analyzing in two different ways whether accounting for ecological interactions helps predict future group sizes. In general, competition for overlapping resources will have no effect on group growth if something besides the overlapping resource limits growth (Verhoef & Morin, 2010). For example, two wikis might share a large number of contributors (they have high user overlap), but their growth might be limited by a lack of core contributors who perform important administrative tasks like policy making and software administration (Zhu, Kraut, et al., 2014). Community ecology relaxes the assumption that competition and mutualism are caused by user overlap density and instead seeks to infer these relationships from data. We test the importance of this conceptual shift for predicting growth by testing two hypotheses. The first uses a model comparison approach to test if adding a measure of ecological interactions to the density dependence model in Study A improves prediction of growth: (***H2)*** *A model with ecological interactions and density dependence predicts growth in online groups better than density dependence alone.*

Support for H2 may be a relatively low bar for assessing whether ecological interactions are important factors shaping the growth of online groups because of confounding moderator or mediator variables related to the occurrence of ecological interactions. Therefore, we also use a time series forecasting approach to test whether modeling ecological interactions is useful for making time series forecasts of participation in online groups: (***H3)*** *The addition of ecological interactions to a baseline time series model improves the forecasting performance.* While this does not directly compare population ecology and community ecology, it validates that ecological interactions are important.

2.3. Materials & Methods

Data

Our data are drawn from the publicly available Pushshift archive of Reddit submissions and comments which we obtained from December 5th 2005 to April 13th 2020 Baumgartner et al. (2020). Within this dataset, we limit our analysis to submissions and comments from the 10,000 subreddits with the highest number of comments. There are 702 subreddits larger than the smallest subreddit included in our dataset having a majority of submissions marked "NSFW," which typically indicates pornographic material. As others have done in large-scale studies of Reddit (e.g., Datta et al., 2017), we exclude these subreddits to avoid asking members of our research team to inspect clusters including pornography. The top 10,000 subreddits provide a sufficiently large number of ecological communities for our statistical analysis.

Study A: Density Dependence Theory

User overlap $o_{i,j}$ quantifies the degree to which two subreddits (i and j) share users. Zhu, Kraut, et al. (2014) and X. Wang et al. (2012) both measure user overlap between two groups by counting the number of users contributing to both groups at least once and exclude users who appear in more than 10 groups. In our preliminary analysis, we found that this measure led to similarity measures and clusters with poor face validity. These issues may have stemmed from how Reddit users often peripherally participate in many groups while participating heavily in few (Hamilton et al., 2017; Tan & Lee, 2015; J. Zhang et al., 2017). Therefore, our measure of user overlap follows Datta et al. (2017) by using the number of comments each user makes in each pair of groups.

To measure user overlap between subreddits, we first build user frequency vectors by counting the number of times each user comments in each subreddit. We prevent giving undue weight to subreddits with higher overall activity levels by normalizing the comment counts for each subreddit by the maximum number of comments by a single author in the subreddit:

$$f_{u,j} = \frac{n_{\mathrm{u,j}}}{max_{v \in \mathrm{J}} n_{v,j}} \tag{2.1}$$

where $n_{u,j}$, the user frequency, is the number of times that user u authors a comment in subreddit j.

This results in a user frequency vector F_j for each subreddit that is sparse and high-dimensional, having one element for each user account that comments in any subreddit in our dataset. Next, we use LSA to reduce the dimensionality of the user frequency vectors. LSA is based on the singular value decomposition and is common in natural language processing and information retrieval. LSA preserves subreddit similarities while removing noise and dealing with sparsity (Dumais, 2004):

$$\begin{aligned} \mathbf{F} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ \widetilde{F}_j &= \mathbf{U_k}^T F_j \end{aligned} \tag{2.2}$$

$\mathbf{F}$ is the matrix where columns are author frequency vectors F_j and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is its singular value decomposition. Truncating the singular value decomposition to use only the first k left-singular

vectors gives $\mathbf{U_k}$. Left-multiplying a subreddit's author frequency vector by $\mathbf{U_k}$ transforms the high-dimensional author frequencies into $\widetilde{F}_j$, their approximation in the k-dimensional space.

We then obtain our measure of *user overlap* by taking the cosine similarities between the resulting vectors for a pair of subreddits:

$$o_{i,j} = \frac{\widetilde{F}_j \cdot \widetilde{F}_i}{\|\widetilde{F}_i\|\|\widetilde{F}_j\|} \tag{2.3}$$

where $\|\widetilde{F}_i\| = \sqrt{\sum_{x=1}^{k} \widetilde{f_{x,i}}^2}$ is the euclidean norm of the transformed user frequencies for subreddit i.

Growth is the dependent variable in our density dependence model testing H1 and is also used in our test of H2 as part of Study B. Growth is measured as the change in the (log-transformed) size of a subreddit over the final 24 weeks of our data, from to November 4th 2019 to April 13th 2020.

Overlap density d_i is the normalized average user overlap for a given subreddit. It is the independent variable in our density dependence model testing H1:

$$\begin{aligned} d_i^* &= \frac{1}{|S| - 1} \sum_{j \in R; j \neq i} \mathrm{o}_{i,j} \\ d_i &= \frac{d_i^*}{\max_j d_j^*} \end{aligned} \tag{2.4}$$

where S is the set of groups in our dataset.

Regression model for H1 To test H1, we fit Model 1 which has first and second-order terms for overlap density to allow for a curvilinear relationship between *overlap density* and *growth*.

$$\text{Model 1} \qquad\qquad Y_i = B_0 + B_1 d_i + B_2 d_i^2 \tag{2.5}$$

where Y_i is the growth of subreddit i and d_i is its overlap density.

Study B: Introducing Community Ecology

Clustering to identify ecological communities Analyzing networks of ecological interactions is the key difference between community ecology and population ecology. To identify ecological communities of related subreddits, we use a clustering procedure based on the user overlap measure described above in §4.3. We selected a clustering model using grid search to obtain a high silhouette coefficient (Rousseeuw, 1987). The silhouette coefficient captures the degree to which a clustering creates groups of subreddits with high within-cluster similarity.

Our description of our measure for user overlap in §4.3 does not explain how we choose the number of LSA dimensions k. To do so, we ran the affinity propagation (B. J. Frey & Dueck, 2007), HDBSCAN (McInnes et al., 2017) and *k*-means clustering algorithms and selected the

algorithm, hyperparameters, and LSA dimensions k that resulted in the clustering with a high silhouette coefficient having less than 5,000 isolated subreddits, and at least 50 clusters. We limit the number of isolated subreddits because some choices of hyperparamters for the HDBSCAN algorithm could improve the silhouette coefficient, but at the cost of greatly increasing numbers of isolated subreddits. Choosing a relatively high limit to the number of isolates helps ensure that our clusters contain highly related communities. We chose an HDBSCAN clustering with 731 clusters, 4964 isolated subreddits, $k = 600$ LSI dimensions, and a silhouette score of 0.48. We exclude the isolated subreddits from our analysis. More details about our clustering selection process are found in the online supplement.

We evaluate the external validity of the chosen clustering using the purity evaluation criterion (Manning et al., 2018) To do so, an undergraduate research assistant examined a random sample of 100 clusters including 744 subreddits. By visiting the subreddits and using her own judgment, the assistant flagged subreddits that did not seem like a good fit for their assigned cluster. Using these labels and excluding 25 subreddits that have been deleted, made private, or banned, we calculated the purity of our clustering as 0.92. This means that we believe that 92% of subreddits belong to their assigned cluster.

Group size is the dependent variable of the models we use to infer ecological interactions. Measured as the number of distinct commenting users in a subreddit each week, group size quantifies the number of people who participate in a subreddit over time. Typical of social media participation data, group size is highly skewed. Therefore, we transform it by adding 1 and taking the natural logarithm.

Inferring ecological interactions using Vector Auto Regression The community matrix $\mathbf{\Phi}$ of ecological interactions can be inferred from time series data using vector autoregression models (VAR models). VAR models are a workhorse in biological ecology because VAR(1) models (i.e., VAR models with a single autoregressive term) have a close relationship with the Gompertz of population growth which is widely used in ecology (Ives et al., 2003). Even in the presence of unmodeled nonlinearities, VAR(1) models can reliably identify competition or mutualism in empirically realistic scenarios (Certain et al., 2018). VAR models also been widely adopted in the social sciences, particularly in political science and in macroeconomics (Box-Steffensmeier, 2014).

VAR(1) models can be intuitively understood as a generalization of auto-regressive AR(1) models in time series analysis. But while AR(1) models predict the state of a single time series as a function of its previous value, VAR(1) models simultaneously predict multiple time series as a function of the values of every other variable in the system (Canova, 2007; Ives et al., 2003):

$$Y_t = B_0 + B_1 t + \sum_{k \in K} A_k x_{k,t} + \sum_{j \in M} \Phi_j y_{j,t-1} + \epsilon_t \tag{2.6}$$

where Y_t is a vector containing the sizes of a set of online groups (M) at time t. B_0 is the vector of intercept terms and B_1 is the vector of linear time trends ($b_{1,j}$) for each community (j). Φ_j represents the influence of $y_{j,t-1}$, the size of the j^{th} online group at time $t-1$ on Y_t. Φ_j is a column of $\mathbf{\Phi}$, a matrix of coefficients in which the diagonal elements correspond to intrinsic

growth rates (marginal to the trend) for each online group and the off-diagonal elements are intergroup influences, and ϵ_t is the vector of error terms

Additional time-dependent predictors ($x_{k,t}$) can be included in the vectors X_k with coefficients a_k. Because subreddits are created at different times, growth trends must begin only after the subreddit is created. We use X_k to introduce a counter-trend during the period prior to the creation of subreddits so that each group's growth trend begins in the period the group is created. For each group j created at time t_j^0 we fill X_j with the sequence $[1, 2, 3, \dots, t_j^0 - 1, 0, 0, 0, \dots]$. In other words, X_j adds a counter-trend only during the period prior to the first comment in subreddit j. We fix the elements $a_{j,i}$ of A_j equal to `0` unless $i = j$, so the counter trend only influences subreddit j. This effectively sets $a_{j,j}$ approximately equal to $-b_{1,j}$.

We fit VAR(1) models using ordinary least squares as implemented in the `vars` `R` package to predict the group size each week using over the history of each subreddit prior to November 4th 2019 (Pfaff, 2008). We hold out 24 weeks of data for forecast evaluation and fit our models on the remainder. To ensure that sufficient data is available for fitting the models, we exclude 946 subreddits and 89 clusters having less than 156 weeks of activity.

Characterizing ecological communities In Study B, we interpret the community matrix $\mathbf{\Phi}$ as a directed network of ecological interactions, a *competition-mutualism network* (Ives et al., 2003). Although the elements of $\mathbf{\Phi}$ correspond to direct associations between group sizes (Novak et al., 2016), ecological interactions can also be indirect. Consider 3 one-directional interactions between three groups (a, b, c) such that growth in a predicts decreased growth in b ($\phi_{a,b} < 0$), growth in b predicts decreased growth in c ($\phi_{b,c} < 0$), but a and c do not directly interact ($\phi_{a,c} \approx 0$).

This does not necessarily mean that groups A and C are independent. Rather, an exogenous increase in A predicts a decrease in B and thereby an eventual increase in C. Such indirect relationships are analyzed by using impulse response functions (IRFs) to interpret a VAR model (Box-Steffensmeier, 2014). In large VAR models containing many groups, the great number of parameters can mean that few specific elements of $\mathbf{\Phi}$ will be statistically significant, even as many weak direct relationships can combine into statistically significant IRFs (Canova, 2007).

Average ecological interaction $\overline{m}$ measures the extent to which an overall ecological community is mutualistic or competitive by taking the mean point estimate of the off-diagonal coefficients of $\mathbf{\Phi}$:

$$\overline{m} = \frac{1}{|M| - 1} \sum_{i \in M} \sum_{j \in M; j \neq i} \phi_{i,j} \tag{2.7}$$

if $\overline{m} > 0$ then mutualistic interactions within the ecological community are stronger than competitive ones, and if $\overline{m} < 0$ then competitive interactions are stronger then mutualistic ones.

Ecological interaction strength κ quantifies the overall strength of ecological interactions in an ecological community as the mean absolute value of the point estimates of the off-diagonal coefficients of $\mathbf{\Phi}$:

$$\kappa = \frac{1}{|M| - 1} \sum_{i \in M} \sum_{j \in M; j \neq i} \left| \phi_{i,j} \right| \tag{2.8}$$

where $\left| \phi_{i,j} \right|$ is the absolute value of the coefficient $\phi_{i,j}$.

Ecological communities of subreddits with overlapping users vary in both the overall strength of ecological interactions and in the overall degree of mutualism and competition between member groups. If an ecological community's average ecological interaction is positive, we say the ecological community is mutualistic. If it is negative, we say the ecological community is competitive. The average ecological interaction can be close to 0 in two ways. First, the ecological interaction strength can simply be low. Alternatively, the ecological community can have a mixture of competitive and mutualistic interactions that cancel one another out when averaged.

Impulse response functions (IRFs) of our VAR(1) models correspond to our visualizations of example competition-mutualism networks in §4.4. An IRF predicts how much each group's size would change in response to a sudden increase in the size of each other group (Verhoef & Morin, 2010):

$$\mathbf{\Theta_t} = \mathbf{\Theta_{t-1}} \mathbf{\Phi}, t = 1, 2, ... \tag{2.9}$$

where $\mathbf{\Theta_t}$ is the impulse response function at time t. $\mathbf{\Theta_0}$ is an M-by-M identity matrix so our impulses represent a log-unit increase of 1 to each group. $\mathbf{\Theta_t}$ is a matrix with elements $\theta^t_{i,j}$ corresponding to the response of group j to the impulse of group i. We draw an edge $i \rightarrow j$ in the competition-mutualism network if the 95% CI of $\theta^t_{i,j}$ does not include zero at any time $10 >= t > 0$. If $\theta^t_{i,j} > 0$, the edge indicates mutualism and if $\theta^t_{i,j} < 0$ the edge indicates competition.[2] We compute the IRFs with bootstrapped confidence intervals (CI) based on 1,000 samples using the `vars` `R` package.

Study C: Predicting growth

Average subreddit mutualism m_j is the independent variable for our test of H2 and measures the average influence of other subreddits in the ecological community on a given subreddit j, which we calculate by taking the mean of off-diagonal elements of row j of the community matrix:

$$m_j = \frac{1}{|M| - 1} \sum_{i \in M; i \neq j} \phi_{i,j} \tag{2.10}$$

where M is the set of subreddits in the ecological community and $|M|$ is the number of subreddits in M. We use the mean instead of the sum because different ecological communities have different numbers of subreddits.

[2] In higher-order VAR(p) models that use $p > 1$ past observations as predictors $\theta^t_{i,j}$ can be less than 0 for some t_a and greater than 0 for some t_b. However, this is not possible in the VAR(1) models we use.

Regression models for H2 We test H2 by using likelihood ratio tests to compare Model 1 and Model 2 which adds *average subreddit mutualism* (m_i) as a predictor. We also fit Model 3 which we compare to Model 2 to test if overlap density explains variation that average subreddit mutualism does not.

$$\text{Model 2} \qquad Y_i = B_0 + B_1 d_i + B_2 d_i^2 + B_3 m_i \tag{2.11}$$

$$\text{Model 3} \qquad Y_i = B_0 + B_3 m_i \tag{2.12}$$

where Y_i is the growth of subreddit i, d_i is its overlap density, m_i is its average subreddit mutualism, and B_0, B_1, B_2, and B_3 are regression coefficients.

Forecasting growth using ecological interactions To test H3, we evaluate whether modeling ecological interactions improves time series forecasting of future participation in online groups by comparing the model in Equation 4.6 to a baseline model with off-diagonal elements of $\mathbf{\Phi}$ fixed to `0`. This baseline model is equivalent to our VAR model, but excludes ecological interactions.

We use two forecasting metrics with differing assumptions: root-mean-square-error (RMSE) and the continuous ranked probability score (CRPS). RMSE is commonly used, non-parametric, and intuitive, but does not take differing scales of the predicted variable or forecast uncertainty into account. Thus, in our setting it may place excessive weight on the forecasts of larger subreddits where errors may have greater magnitude simply because the absolute magnitude of the variance is greater. By rewarding forecasts where the true value has high probability under the predictive distribution, the CRPS accounts for variance in the data and rewards forecasts for both accuracy and precision and is thus a "proper scoring rule" for evaluating probabilistic forecasts (Gneiting & Raftery, 2007). Our CRPS calculations assume that the predictive forecast distribution for each community is normal with standard deviations given by the 68.2% forecast confidence interval. We calculate CRPS using the `scoringRules` `R` package (Jordan et al., 2019).

2.4. Results

Study A: Density Dependence Theory

We test the classical prediction of density dependence theory as formulated in H1 using Model 1 which has first- and second-order terms for the effect of overlap density on growth. As described in §4.2, H1 hypothesizes that overlap density will have a curvilinear ∩-shaped (inverse-U-shaped) relationship with growth indicated by a positive first-order regression coefficient and a negative second-order coefficient.

As predicted, we observe a ∩-shaped relationship between overlap density and growth. Figure 4.1 plots the marginal effects of overlap density on growth for the median subreddit laid over the data on which the model is fit. Table 4.1 shows regression coefficients for Models 1-3. For about half of subreddits, increasing overlap density is associated with higher growth rates. The point where increasing density ceases to predict increasing growth and begins to predict decreasing growth is at the 49$^{\text{th}}$ percentile. Prototypical subreddits at this overlap density grew

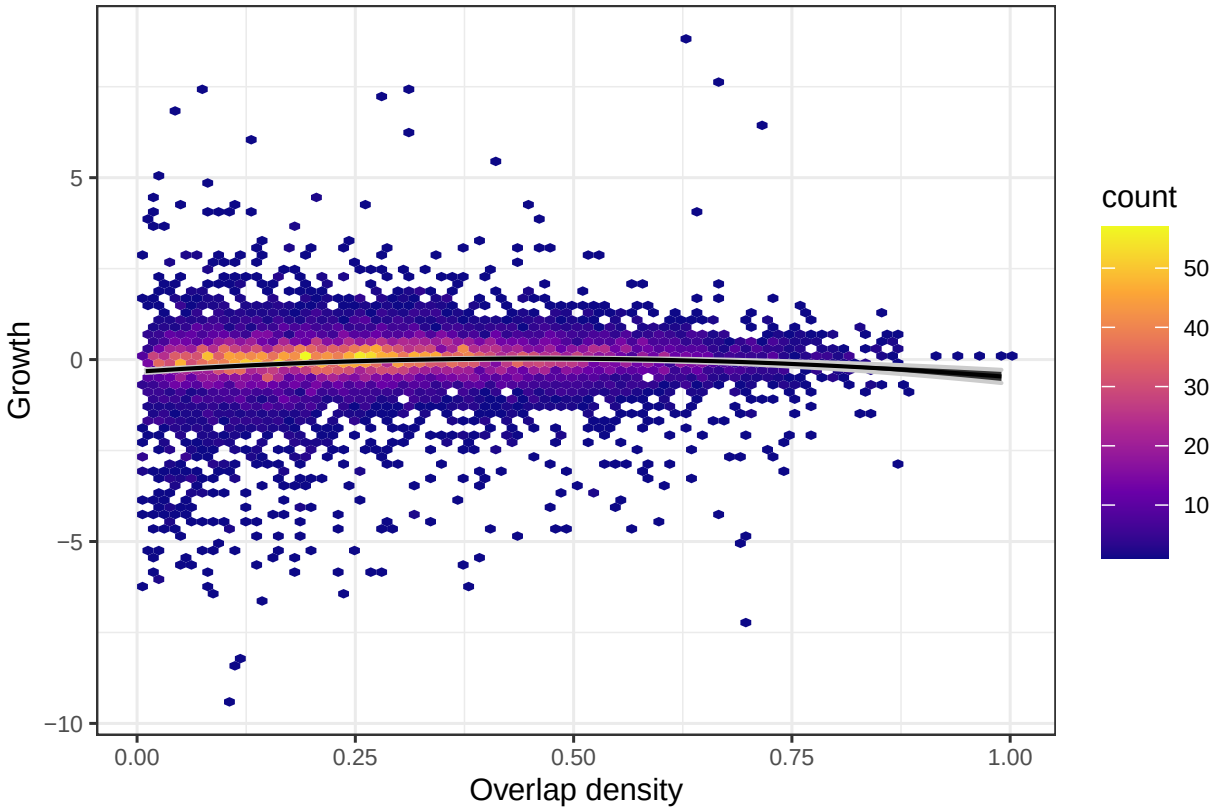


Figure 2.1: Relationship between density and growth. A 2D histogram of subreddits with overlap density (log-transformed) on the X-axis and the change in the logarithm of the number of distinct commenting users on the Y-axis. The black line shows the marginal effect of overlap density on growth as predicted by Model 2. The gray region shows the 95% confidence interval of the marginal effect.

	Model 1	Model 2	Model 3
Overlap density	1.50* (0.26)	1.50* (0.26)	
Overlap density^2	−2.08* (0.41)	−2.09* (0.41)	
Average subreddit commensalism		0.12 (0.26)	0.11 (0.26)
Constant	−0.23* (0.03)	−0.23* (0.04)	−0.04* (0.01)
Log Likelihood	-4970	-4970	-4986
Observations	4,090	4,090	4,090
Note:			*p< 0.01

Table 2.1: Loglinear regression predicting subreddit growth as a function of overlap density. The model supports the prediction of density dependence theory of a ∩-shaped relationship between overlap density and growth.

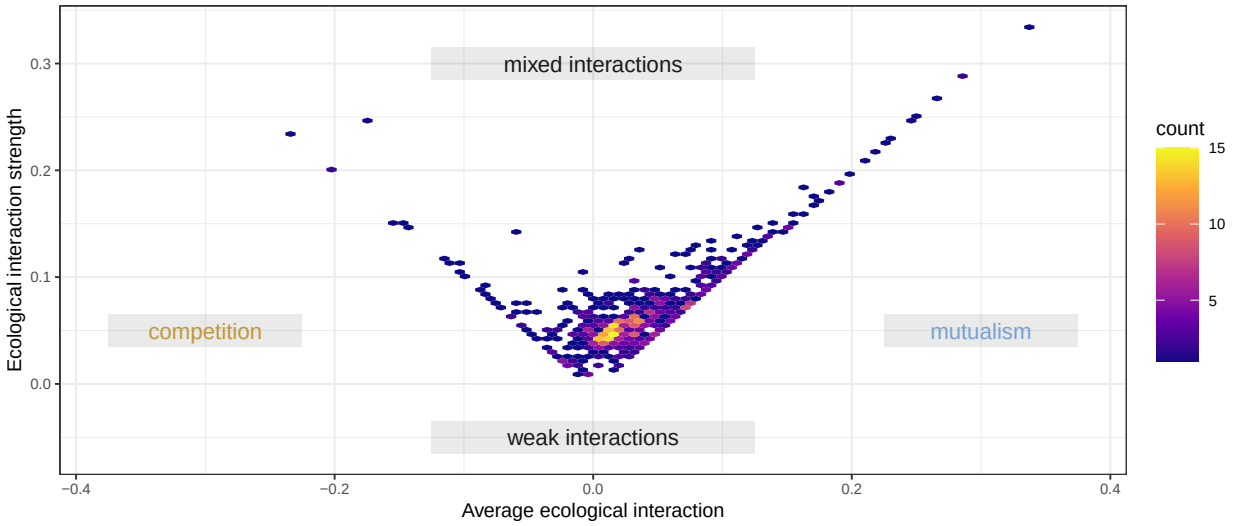


Figure 2.2: Two-dimensional histogram showing ecological communities on Reddit in our typology. The X-axis shows the overall degree of mutualism or competition in clusters of subreddits with high user overlap based on the average ecological interaction. The Y-axis shows the ecological interaction strength representing the overall magnitude of competition or mutualism.

slightly (95% CI:[0.001,0.06]). Yet subreddits at the lower and upper extremes of overlap density slightly declined on average. Typical groups at the 20th percentile of overlap density decline by 1.1 members (95% CI:[-1.1,-1.15]) and typical groups at the 80th percentile decline by 1.2 members (95% CI:[-1.1,-1.28]). While we find support for the classical theoretical prediction of a curvilinear, (∩-shaped) relationship between overlap density and growth, this does not imply that relationships between highly overlapping communities are more competitive.

Study B: Introducing Community Ecology

Figure 4.2 visualizes the distribution of average ecological interaction and ecological interaction strength over the 641 ecological communities we identify. We observe ecological communities characterized by strong forms of both mutualism and competition, others having mixtures of the two, and some with few significant ecological interactions. Mutualism is more common than competition, with the mean community having an average ecological interaction of 0.03 ($t = 14.5$, $p < 0.001$). We find that 524 clusters (81.7%) are mutualistic. Not only are most ecological communities mutualistic, but more mutualistic ecological communities have greater ecological interaction strength (Spearman's $\rho = 0.58$, $p < 0.001$). Therefore, our community ecology analysis suggests that among groups with similar users, mutualistic ecological interactions are more common than competitive ones.

Example ecological communities We present four case studies to illustrate our typology of ecological communities of online groups. Figure 4.2 shows that we find clusters of subreddits characterized by mutualism, competition, a mixture of mutualism and competition, and few ecological relationships at all. We select one case from each of these four types using our measures of average ecological interaction (§4.3) and ecological interaction strength (§4.3). To allow

for more interesting network structures, we draw our cases from the 367 large clusters having at least five subreddits.

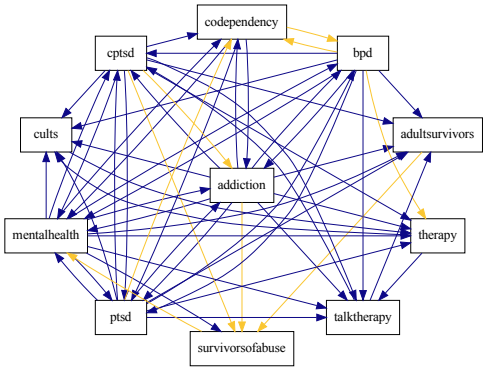
Figure 4.3, presents visualizations of competition-mutualism networks representing statistically significant impulse response functions as described in §4.3. During our analysis, we also examined the terms of the vector autoregression parameter $\mathbf{\Phi}$, the impulse response functions, and model fits and forecasts, all of which are available in our online supplement. We also visited each subreddit in the clusters and read their sidebars and top posts to support our brief qualitative descriptions.

Mutualism among mental health subreddits To find a case characterized by mutualism, we selected the top 37 large clusters with the greatest average ecological interaction. From these, we arbitrarily chose one interesting ecological community, the *mental health* cluster, which includes 11 subreddits for supporting people in struggles with mental health, addiction, and surviving abuse. Constitutive subreddits include those focused on specific mental health diagnoses like `r/bpd` (bipolar disorder) and `r/cptsd` (complex post traumatic stress disorder) while others like `r/survivorsofabuse` and `r/adultsurvivors` are support groups.

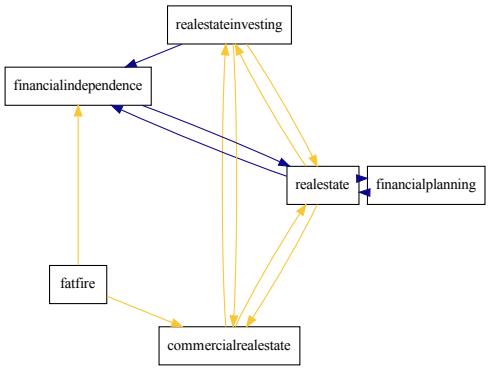
The interactions among these subreddits are dense and primarily mutualistic as shown in Figure 4.3a. There are a handful of competitive interactions like the reciprocal competition detected between `r/codedependence` and `r/bpd`. We also observe some interactions that are mutualistic in one direction and competitive in the other. For example, growth in `r/addiction` predicts increasing growth in `r/cptsd` even as that growth in `r/cptsd` predicts decreasing growth in `r/addiction`. This suggests a pattern in which `r/cptsd` siphons members from `r/addiction`. That said, the density of mutualistic interactions shown in Figure 4.3a suggests that different subreddits have complementary roles in this ecological community as people turn to different types of groups for help with interrelated problems. While attempting to explain why different online groups form mutualistic or competitive interactions is left to future research, the example of mental health subreddits shows how groups with related topics and overlapping participants can have mutualistic interactions where growth in one predicts growth in many of the rest.

Competition among real estate and finance subreddits To find competitive clusters, we selected from the 36 large clusters with the lowest average ecological interaction an ecological community that we label *finance*. Among the 6 subreddits in this cluster, `r/realestateinvesting`, `r/realestate` and `r/commercialrealestate` all deal in different aspects of the real estate industry, while `r/financialindependence` and `r/fatfire` (the acronym "fire" means "financial independence/retire early") are focused on building wealth and becoming financially independent and `r/financialplanning` is a general purpose subreddit for financial advice.

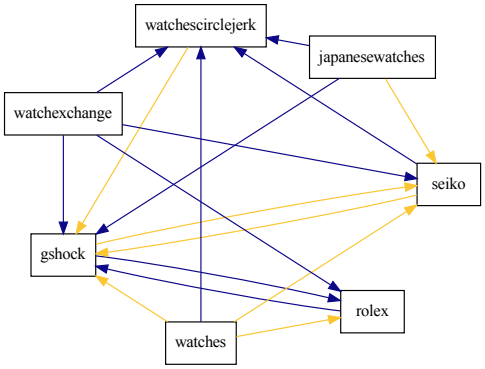
In contrast to the mental health ecological community, the finance cluster has mostly competitive ties as visualized in Figure 4.3b. The fact that even this cluster, among the most competitive in our data, contains a number of mutualistic ties reflects just how prevalent mutualism is among subreddits with high degrees of user overlap. That said, we detect three reciprocal competitive interactions among the three subreddits that focus on real estate. The edges from `r/fatfire` to `r/commercialrealestate` and `r/financialindependence` are competitive as



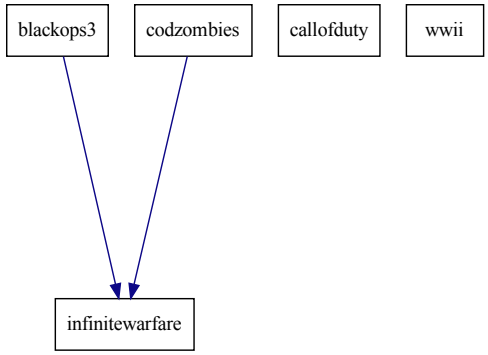
(a) The ecological community of subreddits for supporting mental health and survivors of abuse is dense with largely mutualistic interactions. Some interactions, like that between `r/mentalhealth` and `r/survivorsofabuse` are mutualistic in one direction but competitive in the other.



(b) The subreddits about real estate and finance are relatively competitive. We detect reciprocal competitive relationships among the real estate subreddits in the triad including `r/realestateinvesting`, `r/realestate` and `r/commercialrealestate`.



(c) Subreddits about watches are dense with both mutualistic and competitive interactions. There is a reciprocal competitive interaction between `r/gshock` and `r/seiko`, a reciprocal mutualistic interaction between `r/gshock` and `r/rolex` well as several unreciprocated mutualistic and competitive interactions.



(d) The ecological community of subreddits about Call of Duty video games is characterized by relatively sparse ecological interactions. We detect only two mutualistic interactions from `r/blackops3` to `r/infinitewarfare` from codzombies to `r/infinitewarfare`.

Figure 2.3: Network visualizations of commensal relationships in example ecological communities of subreddits with overlapping users. Yellow indicates competition and purple indicates mutualism.

well. Interestingly, all interactions between the general finance subreddits (r/financialplanning and r/financialindependence) and r/realestate are mutualistic.

Mixed interactions among timepiece subreddits Next, we turn to an example of an ecological community with low average ecological interaction but high ecological interaction strength. We first select the 36 large clusters with the average ecological interaction closest to 0. To find an ecological community with a mixture of mutualism and competition, we select from the 15 clusters with the greatest ecological interaction strength from within this group and chose the *timepiece* cluster containing 7 subreddits about watches.

As shown in Figure 4.3c, the ecological community of timepiece subreddits is dense with ecological interactions (although not as dense as the mental health subreddits). We observe both reciprocated mutualistic interactions, like that between r/rolex and r/gshock, and competitive interactions like that between r/gshock and r/seiko. We also observe numerous unreciprocated competitive and mutualistic relationships like the mutualism between r/watchexchange and r/watchcirclejerk[3] and the competition between r/japanesewatches and r/seiko. Though the average ecological interaction among these subreddits is near 0, our analysis reveals a complex ecological community with a mixture of competition and mutualism.

Sparse interactions among Call of Duty subreddits To find a case where ecological interactions are weak, we return to the group of the 36 large clusters with the average ecological interaction closest to 0 but select from the 15 clusters within this group with the lowest ecological interaction strength. From these, we chose the *Call of Duty* cluster containing five groups about the popular military first-person shooter series of video games.

The Call of Duty ecological community is sparse, having only two significant ecological interactions among its 5 member groups. This ecological community includes subreddits about different editions of the series such as r/blackops3, r/infinitewarfar and r/wwii as well as one about a popular spin-off zombie game r/codzombies and the more general r/callofduty subreddit. We find that growth in r/blackops3 or r/codzombies predicts growth in r/infinitewarfare and no other ecological interactions.

The timepiece and Call of Duty ecological communities illustrate how subreddits with overlapping users can have relatively strong or weak forms of ecological interdependence. Although both clusters are characterized by high degrees of user overlap and low average ecological interaction, the timepiece cluster has a dense competition-mutualism network while the call of duty network is sparse.

Study C: Predicting Growth

We now compare the environmental approach of population ecology with the relational approach of community ecology. In Study B, we presented examples of diverse ecological communities among subreddits with overlapping members. However, the presence of this diversity this does not mean that ecological interactions are related to the growth of online groups, the key outcome of previous ecological studies. We therefore hypothesized that ecological interactions will improve the predictive performance of a density dependence model in H2.

[3]The suffix is widely understood on Reddit to signify a jokey, meme, or satirical subreddit.

Ecological interactions do not improve growth prediction To test H2, we compare Model 1, our density dependence model having first- and second-order terms for overlap density, with Model 2, which also includes average subreddit mutualism (§4.3) as a predictor. We also examine Model 3, in which the only predictor is average subreddit mutualism. Table 4.1 shows regression coefficients for our models.

We do not observe a statistically significant association between average subreddit mutualism and growth ($B_3 = 0.12, SE = 0.26$). Moreover, a likelihood ratio test comparing Model 1 and Model 2 does not support H2 as Model 2 does not predict subreddit growth better than Model 1 ($\chi^2 = 0.23$, $p > 0.05$). Comparing Model 2 to Model 3 shows that overlap density explains variation that average subreddit mutualism does not ($\chi^2 = 33$, $p < 0.001$). Overlap density helps explain a group's future growth, but the overall degree of mutualism or competition a group faces in its ecological community does not.

Forecasting accuracy The likelihood ratio tests in §4.4 are limited because improvements in predictive performance (or lack thereof) may be due to unobserved factors predictive of growth that are correlated with average subreddit mutualism. We hypothesized in H3 that the inter-group dependencies in our VAR models can better forecast the size of subreddits compared to baseline time series models that do not account for ecological interactions. As described in §4.3, we test H3 by comparing two forecasting metrics: the root-mean-square-error (RMSE) and the continuous ranked probability score (CRPS).

VAR models including ecological interactions have forecasting performance superior to the baseline model in terms of both RMSE and CRPS. We evaluate the 24-week forecast performance for all subreddits which were assigned to clusters. The RMSE under the baseline model (0.84) is greater than the RMSE of the VAR models (0.75) and the CRPS of the baseline model (72,853) is also greater than the CRPS of the VAR models (72,669). This reflects a substantive improvement in forecast accuracy robust to the choice of the forecasting metric.

Our baseline model contains a constant term and a trend term for each group and therefore accounts for all time-invariant within-group variation. Because overlap density is a subreddit-level variable that does not vary over time, we know that the improvement in forecasting performance comes from modeling ecological interactions in ways not captured by overlap density.

2.5. Threats to Validity

Our work is subject to several important threats to validity that we cannot fully address. First, we study ecological communities on only one platform hosting online groups and our results may not generalize to other platforms or time periods. Additionally, while our community ecology approach assumes that ecological interactions drive dynamics in the size of groups over time and cause groups to grow or decline, drawing causal inference using our method would depend on several untestable assumptions. For example, our ability to infer causal relationships might be limited if groups we do not consider—including groups on other platforms—play a role in an ecological community. Regression estimates in Models 1-3 may be confounded by omitted variables and cannot support causal interpretation. Therefore, we refrain from claiming that the relationships we infer are causal.

The method we propose for identifying ecological interactions between online groups has limitations common to all time series analysis of observational data. Potential omitted variables might also include additional time lags of group size. Although we chose to use VAR(1) models with only 1 time lag, we hope future work can improve upon our approach and model more complex dynamics with additional lags. Like most other time series analysis, vector autoregression assumes that the error terms are stationary. This is difficult to evaluate empirically and may not be realistic (Canova, 2007). Future work might relax these assumptions using more complex models with time-varying parameters, state space models (Box-Steffensmeier, 2014), nonlinear time series models (Cenci et al., 2019; Kantz & Schreiber, 2003), or stationarity-enforcing priors (Heaps, 2020). Such approaches may require additional contextual knowledge and be difficult to scale to an analysis of hundreds of different ecological communities, but may prove fruitful in future work focusing on ecological communities of interest. Such models may also be useful in future work investigating how ecological interactions change over time.

Additional threats to validity stem from our use of algorithmic clustering to identify ecological communities. Organizational ecologists have rarely attempted to estimate the full community matrix for an entire population containing a large number of groups because of data and statistical limitations (e.g. Ruef, 2000; Sørensen, 2004). For instance, 100 million possible ecological interactions exist within a set of 10,000 communities. Attempting to infer them all raises considerable computational and statistical challenges. We chose to use a clustering analysis to explore the typical ecological communities on a platform.

While we choose clusters based on high degrees of user overlap and validate our clustering in terms of the silhouette coefficient and purity criteria, we might have obtained different results if we had clustered in a different way. Additionally, our efforts to obtain clusters with a high silhouette coefficient lead us to remove a large number of subreddits from our analysis. Thus, our results are not representative of Reddit overall, but only of those subreddits that were included in our analysis. Furthermore, clustering algorithms like the one we use may not have unique solutions and different initial conditions and hyperparameters might lead to different results. While these allow us to scale up our analysis, future work should use principled definitions of an ecological community based on qualitative contextual knowledge in focused studies of particular ecological communities.

2.6. Discussion

To introduce community ecology and compare it to population ecology, we presented three studies. In Study A, we found support for H1 showing—as predicted by density dependence theory—that overlap density has an ∩-shaped association with subreddit growth. Subreddits with moderate overlap density in our data declined less than subreddits with either very low or very high overlap density. According to population ecology theory, this suggests that high-density environments are competitive and less conducive to growth than medium-density environments.

Surprisingly, this contrasts with our results in Study B, where we studied the diversity of ecological communities using vector autoregression models of group size over time to infer networks of ecological interactions. We find ecological communities that are mutualistic or competitive, that mix the two, or that have few significant ecological interactions at all. Overall, however, ecological communities of subreddits are typically mutualistic and mutualistic inter-

actions are stronger on average than competitive ones. Although we find evidence of density dependence, density-dependent competition does not necessarily reflect typical relationships in ecological communities of highly overlapping subreddits.

Our results in Study C show that the size of the other members of an ecological community improves time series forecasts of participation in online groups. However, average subreddit mutualism did not help predict growth. This suggests that population ecology and community ecology offer complementary environmental and relational perspectives. Population ecology's focus on environmental factors such as niche and overlap density is useful for predicting growth, but does not provide a way to study networks of mutualism and competition. Community ecology unpacks density and provides insights about the specific relationships between groups. While modeling these interactions helps forecast participation levels in groups, the existence of these interactions may be independent of future growth. For example, if mutualistic relationships are common in declining ecological communities, that would explain our result for H2.

The complementary nature of the two ecologies is seen in the coincidence of our findings in Study A and Study B. Indeed, these results can help explain the puzzling set of empirical results about the relationship between overlap density and outcomes like growth, decline and survival (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). Studies of density dependence theory in social computing measure the density of an online group's niche in terms of its overlap in participants or topics. Our analysis clearly shows that resource overlaps between two groups might have little to do with whether they are mutualists or competitors. Instead, overlaps may simply reflect the hospitality of the environment to groups with overlapping topics or user bases. As a result, the differing environmental conditions of Wikis and Usenet groups might explain why user overlap was associated with the survival of wikis (Zhu, Kraut, et al., 2014) but with the decline of Usenet groups (X. Wang et al., 2012). Wikia was a young and growing platform during Zhu, Kraut, et al.'s (Zhu, Kraut, et al., 2014) data collection period when the growth of groups may have been limited by knowledge of how to build a wiki, and this knowledge was provided by overlapping experienced users. Usenet was in decline during X. Wang et al.'s (X. Wang et al., 2012) study period and this may have produced competitive environmental conditions as users became more scarce.

The widespread mutualism found in Study B resonates with long-held understandings of ecological interactions in evolutionary theory (Kropotkin, 1902/2012). Competition is unlikely to persist because it decreases survival. Because mutualism increases survival, it will be favored by natural selection (Armstrong & McGehee, 1980; Axelrod & Hamilton, 1981). Similarly, competition can be avoided if groups adopt specialized roles in their ecological community, a dynamic known as resource partitioning in organizational ecology (Carroll, 1985; Menge, 1972; Schoener, 1974). Resource partitioning theory suggests that the competition among real estate subreddits observed in Figure 4.3b may be due to a lack of specialization. If specialization does not emerge over time, such groups of competing subreddits may have decreased survival. By contrast, mental health support groups like those observed in Figure 4.3b appear to have distinctive purposes or roles. Future work to test such mechanisms in ecological communities of online groups may reveal ways that online groups complement or cooperate with each other.

Within large platforms for online groups, the great number of ecological communities that can be studied should make it possible for future work to apply methods from network science to construct and test generalizable theories about the roles of different types of resources, design

features of platforms, and governance institutions in these ecological interactions. Future work should also incorporate community ecology analysis in case studies of important topics such ecological communities engaged in peer production, political mobilization, misinformation, or mental health support.

Although we focused on online groups within a single platform, groups may use multiple platforms with distinctive affordances for different purposes (Fiesler & Dym, 2020; Kiene et al., 2019). Since the VAR method relies only on time series data to infer ecological interactions, it can be applied to study ecological communities spanning social media platforms. Community ecology can thus provide a bridge between quantitative studies of participation in online groups and theories of interconnected information ecologies (Nardi & O'Day, 1999). While we focus on relationships between groups sharing a platform, one can apply our concepts and methods to understand how interdependent systems of technologies and users give rise to higher levels of social organization on social media platforms (Aldrich & Ruef, 2006; Astley, 1985).

Implications for Design

In the final chapter of their book on *Building Successful Online Communities*, Kraut et al. (2012) advise managers of online groups to select an effective niche and beware of competition. However, these recommendations are based on little direct evidence from studies of online groups and offer almost no concrete steps that designer or group should take based on either piece of advice. Although further research into ecological interactions is needed before design principles can be derived, we provide a framework for online group managers to think about ecological constraints on group size. While intuition suggests that online group managers might seek out mutualistic relationships and avoid competitive ones, it is often not obvious whether another group with overlapping users is a competitor or mutualist. Our method provides a way for group managers to know.

Competitors have a negative impact on growth, but ecological theory suggests that specialization is an adaptive strategy in response to competition (Aldrich & Ruef, 2006; Carroll, 1985; Kraut et al., 2012; Powell et al., 2005). Using our method, group managers might identify competitors limiting the growth of their groups. With the knowledge of this analysis in hand, they might be able to escape a competitive dynamic by specializing. While competitive relationships are defined by how they decrease the size of groups, competition can also be important to the health of the broader ecological community. Exit to an alternative group can be an avenue for political change in response to grievances and poor governance (S. Frey & Sumner, 2019; Hirschman, 1970). The threat of competition with other groups may make expressions of voice more persuasive to moderators or platforms (Hirschman, 1970).

Groups looking to increase activity should desire to seek out mutualistic relationships, and we believe that designers of online platforms can help them do so. Features such as meta-groups, group search, recommendation engines, and practices like linking related groups may lower barriers between groups and support mutualism. However, it is not obvious to what extent particular features will support competition, mutualism, or both. Using our method, managers and designers can test features intended to support mutualism.

2.7. Conclusion

While explanations for the rise or decline of online groups often look to internal mechanisms, understanding the role of interdependence between online groups is increasingly important. While prior research has investigated competition and mutualism among online groups with overlapping users and topics using the population ecology framework (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014), this approach does not provide a way to infer competitive or mutualistic interactions among related groups. We introduce the community ecology framework as a complementary perspective to population ecology. By inferring competition-mutualism networks directly from time-series data, our community ecology approach helps resolve the empirical tensions raised by prior ecological work in social computing and reveal that most interactions within clusters of subreddits with highly overlapping users are mutualistic. Our methods provide a foundation for future work investigating related online groups.

Preface to Chapter 3

An important finding from Chapter 2 is that mutualism is much more common than competition among overlapping subreddits. This finding was also surprising because ecological theory and prior results in social computing suggest that greater niche overlaps result in stronger competition. Furthermore, theories of organizational ecology were insufficient for explaining the reasons why overlapping online communities exist in the first place. Therefore, the qualitative investigation presented in Chapter 3 provided important explanation and validation of the quantitative finding of widespread mutualism in terms of the experiences and understandings of active participants in overlapping subreddits. If the findings from Chapter 3 had been known in advance of Chapter 2's study, Chapter 2 would have been more likely to anticipate widespread mutualism and may have been designed to explain it.

Because Chapters 2, 3, and 4 are each written as stand-alone articles, some parts of the background section of Chapter 3, most notably the first 3 paragraphs of §3.2 makes some of the same points as the background section of Chapter 2. Also, the interview recruitment process uses an earlier version of clustering algorithm from Chapter 2 (before it was improved during a revise and resubmit process). The second paragraph of §3.5 summarizes the clustering procedure. Readers of Chapter 2 may quickly pass over the these paragraphs.

This chapter is a collaborative work with Charles Kiene, Isabella Brown, Laura (Alia) Levi, Nicole McGinnis, and Benjamin Mako Hill and is under review in Proceedings of the ACM on Human-Computer Interaction: Computer Supported Cooperative Work.

Chapter 3

No Community Can Do Everything: Why People Participate in Similar Online Communities

Large-scale quantitative analyses have shown that individuals frequently talk to each other about similar things in different online spaces. Why do these overlapping communities exist? We provide an answer grounded in the analysis of 20 interviews with active participants in clusters of highly related subreddits: within a broad topical area, there are a diversity of benefits an online community can confer. These include (a) specific information and discussions, (b) socialization with similar others, and (c) attention from the largest possible audience. A single community cannot meet all needs. Our findings suggest that topical areas within an online community platform tend to become populated by groups of specialized communities with diverse sizes, topical boundaries, and rules. Compared with any single community, such systems of overlapping communities are able to provide a greater range of benefits.

3.1. Introduction

Early work in social computing treated online communities as isolated units that could be understood without considering their members' participation in other online communities. As community hosting platforms such as Reddit and Facebook have grown in prominence, social computing scholars have sought to document and explore the connections between online communities (Datta & Adar, 2019; Hill & Shaw, 2019; Tan & Lee, 2015; Zhu, Chen, et al., 2014). This research has shown that online communities overlap with each other in terms of their memberships and topics in ways that have important consequences for a range of outcomes (Chandrasekharan et al., 2018; TeBlunthuis & Hill, 2021; X. Wang et al., 2012).

User and topic overlap is widespread—both within platforms and across them. For example, a range of studies have highlighted the fact that members frequently participate in multiple online groups. This occurs both serially as users migrate between communities over time (Lu et al., 2019; Tan, 2018; Tan & Lee, 2015) and concurrently as individuals belong to multiple groups at once (Hwang & Foote, 2021; X. Wang et al., 2012; Zhu, Kraut, et al., 2014). Many large platforms host distinct communities with similar topics and content (Datta et al., 2017; Zhu, Chen, et al., 2014). In at least one study, researchers have documented that overlaps in users and topics often coincide (Datta et al., 2017). In other words, members of online communities often simultaneously participate in overlapping conversations with overlapping groups of people in

different online spaces.

Why are the same individuals talking to each other about similar things in different online communities? Although social computing offers many theories of why individuals might want to participate in a community, almost all empirical work in social computing on user and topic overlap has used computational or quantitative analysis. As a result, we know very little about what overlaps mean to users. Critically, we also have very little in the way of empirical evidence that is able to speak to why communities overlap in the first place.

Our work seeks to complement existing quantitative research with a better qualitative understanding of intercommunity overlap and contribute to several streams of social computing scholarship. In particular, our work complements a series of social computing studies that have taken inspiration from ecological theory and shown that online groups' growth and survival are closely tied to activity in adjacent online spaces (TeBlunthuis & Hill, 2021; X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014).

We seek to answer our research question (in italics above) through an interview-based study of Reddit users with experience in overlapping communities. Using a dataset of posts and comments on Reddit, we identify clusters of communities on Reddit with highly overlapping users and topics and recruit a set of 20 participants from nine clusters. Drawing from a grounded theory analysis of interview transcripts, we develop an explanation of why many users simultaneously participate in communities with overlapping memberships and topics.

Our findings suggest that users seek three salient benefits from online groups: users want to (a) find specific types of content, discussions, and information; (b) connect with similar types of people; and (c) share content with the largest possible audience. Our work also suggests that these three benefits are frequently in conflict such that the more a community provides one of these benefits, the less able it may be to provide the other two. Because it is difficult for a single community to fully provide all three benefits, clusters of multiple overlapping communities are constructed to do so in aggregate.

3.2. Related Work

Although most research in online communities analyzes the internal factors driving online community success (Kraut et al., 2012), a growing literature studies communities related by overlaps in topic or membership (Datta et al., 2017; Tan & Lee, 2015; TeBlunthuis & Hill, 2021; Zhu, Kraut, et al., 2014). This work has found that concurrent engagement in multiple communities is common on large platforms that host online communities such as Reddit where individuals smoothly jump from community to community (Tan & Lee, 2015). With several exceptions (e.g., Fiesler & Dym, 2020; Hwang & Foote, 2021; Kiene et al., 2019; Zhao et al., 2016), this work typically takes the fact that communities overlap for granted and focuses on the consequences of overlap on outcomes such as the emergence and growth of communities (Butler & Wang, 2011; Zhu, Kraut, et al., 2014) and the diffusion of types of language such as hate speech (Chandrasekharan et al., 2018). None of this work provides insight into how communities come to overlap and why these overlaps persist.

Researchers have investigated intercommunity conflict and found that conflict is initiated by a very small proportion of online communities (Kumar et al., 2018). Other work has shown that content cross-posted to different communities contributes to the ongoing renegotiation of the topical boundaries (Butler & Wang, 2011). J. S. Zhang et al. (2021) have shown that

topical boundaries can also shift as similar communities attract users with different interests. In a related sense, Massanari (2017) has argued that toxic communities can influence the broader culture of a platform for online communities. As a result, banning problematic communities from a platform such as Reddit can reduce toxicity in adjacent communities that are not directly affected (Chandrasekharan et al., 2017; Ribeiro et al., 2021).

A number of studies on overlapping communities draw upon ecological theory (TeBlunthuis & Hill, 2021; X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). Ecological approaches in social computing theorize that overlaps between users and topics relate to competitive or mutualistic forces and drive outcomes such as growth and survival. For example, X. Wang et al. (2012) found that membership overlap reduced the growth rate of Usenet groups. Zhu, Chen, et al. (2014) found that participation rates often suffered if there was too little or too much overlap with other communities. Zhu, Kraut, et al. (2014) found that communities' survival was positively associated with membership overlap, especially with overlap with older communities. Recently, TeBlunthuis and Hill (2021) found that mutualism is common in clusters of overlapping subreddits.

Although these studies use statistical analysis to tell us about how communities relate to each other, they do not to speak to *how* participants understand the relationships between similar online communities or *why* they participate in overlapping communities. The exclusively quantitative nature of these accounts means that a range of potential explanations are possible.

Although we know of no qualitative examination focused directly on understanding why overlapping communities exist, there are a series of qualitative papers that point to potential answers. Fiesler and Dym (2020) describe the history of online fanfiction writing communities migrating across platforms in pursuit of hospitable infrastructure. Similarly, Zhao et al. (2016) describe how individuals use multiple social media platforms to meet varied and nuanced communication needs. Although their study is primarily quantitative, Zhu, Chen, et al. (2014) include quotes from interviews to support the emic validity of notions of competition and mutualism between groups in an enterprise social media system. Finally, Hwang and Foote's (Hwang & Foote, 2021) paper seeks to explain why individuals participate in persistently small online communities on Reddit and ends with a reflection that many small communities are sustainable only because they are "nested" within larger niches. All told, these findings suggest a rich social process by which participants in online communities purposefully construct and move between overlapping spaces.

However, the very small amount of qualitative evidence from participants in overlapping communities in the same platform means that we lack a strong sense of why members choose to participate in multiple communities simultaneously. Although ecological studies attempt to quantify competition and mutualism, we know little about how members understand the relationships between their communities or if these key ecological concepts have any emic resonance. Our work seeks to place ecological studies of online communities on firmer qualitative ground.

Reasons for Joining Online Communities

Decades of social computing research has sought to understand why people belong to particular online communities (Kraut et al., 2012). It has long been recognized that different people have different motivations and that a single individual may have multiple motivations that include

the social, informational, and material benefits users receive through their participation (Butler, 2001; Turner, 2005; Xigen Li, 2011). In terms of uses and gratifications theory, "users actively seek particular media with the goal of gratifying an existing need" (Lampe et al., 2010). Past research has shown that people seek online communities to collaborate on projects (Poor, 2014), to receive social support (Leimeister & Krcmar, 2005), to cooperate with friends (Turner, 2005), and, especially, to exchange information (Leavitt & Robinson, 2017; Liang, 2017; Muhtaseb & Frey, 2008; Ridings & Gefen, 2004). Other research focuses on the growth and decline of membership in online communities and surfaces motivations for why people choose not to participate (Cunha et al., 2019). Brandtzæg and Heim (2008) found that a lack of trust or low quality content can lead to declines in membership. Online communities may decline because leaders are resistant to change and unwelcoming to newcomers (Halfaker et al., 2013; Shaw & Hill, 2014; TeBlunthuis et al., 2018).

Although our findings are the result of an inductive process of bottom-up grounded theory analysis, the presentation of our findings relies on three existing concepts.

Finding specific content One of the most important features of online communities is their ability to enable the spread of useful knowledge and information (Faraj et al., 2016). By connecting individuals with specific information and skills that they desire, online communities match knowledge seekers with experts and foster collaboration on information goods (Benkler, 2006; Fiesler et al., 2017; Fulk et al., 1996; Lakhani & von Hippel, 2003). Research has often focused on the ways that individuals utilize diverse types of social computing systems to meet their specific information needs through systems such as Q&A sites (Adamic et al., 2008), synchronous chat systems (White et al., 2011), search engines (Morris et al., 2010a), social network sites (Morris et al., 2010b; Starbird, 2012), fanworks (Fiesler et al., 2017), and knowledge bases (Ackerman & Malone, 1990; Orlikowski, 1992).

Homophily A second need that online communities serve is to foster connections with similar others. The term *homophily*, "a tendency for friendships to form between those who are alike in a designated respect" (Lazarsfeld & Merton, 1954), describes the set of benefits people can only receive from others who share their identities, beliefs, interests, or culture (M. McPherson et al., 2001). In offline settings, homophily helps explain why tastes in cuisine, music, and other cultural preferences are often correlated (DellaPosta et al., 2015), why similar people tend to congregate, and what happens when they do (M. McPherson et al., 2001). Homophily on social networks may drive the emergence of online "echo chambers" as individuals seek online communities whose members share their beliefs (Dvir-Gvirsman, 2017; Grevet et al., 2014; Himelboim et al., 2016; T. J. Johnson et al., 2009).

Research has shown that people have greater degrees of trust in homophilous groups and are more likely to share content posted by homophilous others (Chang et al., 2014; Ma et al., 2019). Homophily has been described as an important feature of online fan communities (Fiesler et al., 2017; Hillman et al., 2014).

Finding the largest possible audience Research on online communities producing public information goods has found evidence that audience size motivates contributors (X. M. Zhang & Zhu, 2011). Additionally, numerous studies have shown that users of social networking sites

frequently consider the audience that their posts and messages may reach (Marwick & boyd, 2011; J. Zhang et al., 2020). As individuals on social media typically have little information about who sees their posts, they conceive of "imagined audiences" based on cues from visible activity (Bernstein et al., 2013) and target imagined audiences using deliberate strategies, such as using multiple platforms to reach distinct audiences, in order to control who sees or does not see their posts (Litt & Hargittai, 2016; Marwick & boyd, 2011; Zhao et al., 2016).

3.3. Study Design

To study overlapping membership in online communities, we conduct interviews with members of online communities hosted on Reddit, a social media platform for sharing, discussing, and rating news, media, and other content in user-created subcommunities called "subreddits." Individual users can participate in any of Reddit's millions of subreddit communities by posting "submissions" that might include a link to a news article, a question for discussion, an image, or text written by the submitter. Each submission has a corresponding threaded comments section. Users can also vote submissions and comments up or down as a form of distributed moderation and can give awards to comments and posts (Burtch et al., 2021; Lampe & Resnick, 2004).

Subreddit communities are managed by teams of volunteer content moderators tasked with curtailing abusive behavior and keeping conversation on topic (Matias, 2019; Seering et al., 2019). Subreddits exist covering an enormous range of topics (Fiesler et al., 2018), and Reddit has been the site of much research on overlapping online communities (Datta et al., 2017; Hessel et al., 2016; Tan, 2018; Tan & Lee, 2015; TeBlunthuis & Hill, 2021). Because the cost of creating and joining new communities on Reddit is very low, subreddits often overlap in both topic and membership. Users frequently create spinoff subreddit communities from larger and more established groups (Tan, 2018).

Participant Selection

To understand why people participate in overlapping communities, we set out to interview people who are active in highly related subreddits. Additional inclusion criteria were that users were adults (i.e., above the age of majority in their country) and able to participate in an interview in English.

Our participant selection process began by first choosing clusters of highly related groups. To do so, we built a web-based data visualization of a clustering algorithm derived from user overlap to identify groups of interest-based subreddits having similar users. To generate the visualization, we conducted a computational analysis of the Pushshift Reddit dump (Baumgartner et al., 2020), containing a nearly complete collection of Reddit comments made before April 2020. We selected the top 10,000 subreddits based on the number of comments in this data and excluded subreddits where a majority of submissions were flagged as not safe for work. Next, following an approach described in prior work (Datta et al., 2017), we constructed the measure of user similarity by taking the cosine similarities of TF-IDF vectors. Using this similarity measure, we ran affinity propagation clustering (B. J. Frey & Dueck, 2007) to group subreddits having overlapping users. We then built an HTML visualization of these clusters based on t-distributed stochastic neighbor embedding (t-SNE). We have included the visualization in our online supplement.

Although some aspects of our manual cluster selection process using this visualization were necessarily arbitrary, we tried to select clusters that were interest driven, involved primarily English language discussion, and were focused on content about which all members of the research team would be comfortable speaking. As a result, we did not select any clusters that were focused on sex or pornography, fringe or extreme politics, content specific to geographic regions, or topics that our group could not understand.

We sought out clusters that we hoped would result in individuals from a diverse range of ages, genders, and life experiences. Although we did not collect demographic information from our interviewees, our interviewees' presentation and descriptions of themselves suggested that these efforts were not entirely successful. Our pool of interviewees included young and middle-aged people; people of color; people from the United States, Canada, and Europe; people who did not speak English as a first language; and people who were non-male. That said, men were very likely over-represented in our pool of interviewees, perhaps even in relation to the disproportionate participation of men on Reddit (Amaya et al., 2021).

The clusters we selected each include 3–10 subreddits on the following topics: rock climbing, streetwear fashion, roller coasters, vintage audio, podcasting, painting, drag culture and performance, indie music, and dating for middle-aged adults. Information about each subreddit and cluster can be found in Table 3.1.

Using the Pushshift Reddit dataset, we identified candidate participants who were among the top 80% most frequent commenters within each cluster, who participated in multiple subreddits in the cluster, and who were active in the cluster during a period of at least 1 calendar year. We began recruiting a random sample of 50 candidates matching these criteria within each cluster by sending direct messages through Reddit. Interested potential recruits filled out a short online survey confirming that they were adults and able to participate in English language interviews. The survey also asked participants about their participation and familiarity with each of the subreddits in each cluster to verify that they were knowledgeable. At the beginning of each interview, we asked if there were any other subreddits related to those identified by the clustering algorithm. As a result, our conversations were not limited to the subreddits listed in Table 3.1.

We began by recruiting participants from the first three clusters listed in Table 3.1. We found ourselves reaching saturation within these clusters quickly. We also found that different clusters were surfacing quite different data. In response, we added additional clusters and recruited at least two participants from each until we reached global saturation. In some clusters, we did not reach saturation in two interviews. In these cases, we sent additional invitations and conducted additional interviews. In total, 20 participants were successfully recruited and interviewed by five members of the research team before we reached global saturation and ceased data collection. The characteristics of our interviewees are presented in Table 3.2.

All of our interviews were semistructured. Although we drew from a long series of open-ended questions about participation in different subreddits and the relationships between communities, we chose our questions based on what our subjects wanted to talk about. A copy of our interview protocol is included in our supplementary material. Interviews were 49 min long on average but varied substantially in length. We suggested conducting interviews over Zoom but offered participants their choice of communication channel. As a result, we conducted two interviews over the phone, one using Discord chat, and the rest over Zoom. Interviews were transcribed automatically using Zoom's built-in transcription and the otter.ai service and were then manually corrected by the authors. After each interview, participants were compensated

Subreddit	Cluster	Subscribers	Created
r/bouldering	Climbing	194,814	2009-10-28
r/climbharder	Climbing	117,288	2010-10-19
r/climbing	Climbing	935,621	2008-07-17
r/climbingcirclejerk	Climbing	45,032	2011-08-18
r/Drag	Drag	44,724	2011-01-15
r/Dragula	Drag	27,510	2016-11-03
r/rupaulsdragrace	Drag	440,329	2011-11-15
r/RPDR_UK	Drag	31,867	2019-02-07
r/SpoiledDragRace	Drag	69,027	2018-02-16
r/MsPaintsArtRace	Drag	61,292	2017-04-17
r/MGMT	Indie Music	17,744	2010-02-25
r/tameimpala	Indie Music	94,248	2011-10-30
r/kgatlw	Indie Music	59,191	2015-07-01
r/Indieheads	Indie Music	1,932,698	2013-12-24
r/datingoverthirty	Middle Age Dating	436,480	2014-11-04
r/DatingAfterThirty	Middle Age Dating	11,550	2018-03-09
r/datingoverforty	Middle Age Dating	52,522	2018-12-15
r/relationshipsover35	Middle Age Dating	14,916	2018-02-06
r/OilPainting	Painting	186,716	2011-09-22
r/Painting	Painting	280,865	2008-06-13
r/PourPainting	Painting	178,800	2017-07-28
r/Watercolor	Painting	269,882	2012-01-15
r/HappyTrees	Painting	53,362	2011-02-07
r/podcasts	Podcasting	1,995,693	2008-01-25
r/podcast	Podcasting	60,497	2009-01-02
r/podcasting	Podcasting	73,010	2010-09-17
r/audiodrama	Podcasting	129,102	2010-11-30
r/ska	Podcasting	34,397	2008-03-12
r/guessthecoaster	Rollercoasters	5,094	2017-06-30
r/rollercoasterjerk	Rollercoasters	12,378	2016-07-14
r/rollercoasters	Rollercoasters	66,652	2010-07-31
r/rct	Rollercoasters	55,275	2010-08-04
r/themeparkitect	Rollercoasters	13,536	2014-06-16
r/streetwear	Streetwear	2,678,745	2011-04-30
r/supremeclothing	Streetwear	154,797	2012-04-04
r/womensstreetwear	Streetwear	421,279	2016-04-25
r/bapeheads	Streetwear	19,672	2013-08-12
r/malefashion	Streetwear	207,843	2011-04-02
r/sadboys	Streetwear	74,932	2013-06-30
r/techwearclothing	Streetwear	94,675	2017-03-01
r/Vans	Streetwear	51,997	2011-07-01
r/cassetteculture	Vintage Audio	45,615	2011-05-25
r/typewriters	Vintage Audio	20,037	2010-10-25
r/vintageaudio	Vintage Audio	59,202	2011-09-18

Table 3.1: Clusters of subreddits from which we recruited participants, subscriber counts at the time of the study, and the creation date of each subreddit.

with a digital gift card for $20 USD through the Tango Card reward service[1]

[1] https://www.tangocard.com/

Qualitative Data Analysis

Our analysis followed Charmaz's (Charmaz, 2015) approach to grounded theory as closely as possible. We conducted coding and data collection in parallel. We generated over 950 codes, which we then grouped in an iterative axial coding process that generated 18 thematic memos. As we completed collecting data, we refined our codes and combined themes to identify answers to our following orienting research questions: Why are there so many similar online communities? And why not more? Although primarily inductive, our analysis was influenced by sensitizing concepts from prior work including our knowledge of scholarship on overlapping online communities described in §3.2 and the reasons that people participate in online communities summarized in §3.2. In analyzing our data, we noted that interviewees described their participation in multiple different subreddits and their preference for particular subreddits in terms of the inability of one community (often the "main" or "largest" community) to provide the desired benefits. This observation formed the basis of the grounded theory around which we organize our findings.

Participant ID	Cluster	Interview Length (min)
C1	Climbing	56
C2	Climbing	51
C3	Climbing	41
D1	Drag	51
D2	Drag	67
I1	Indie Music	71
I2	Indie Music	43
O1	Podcasting	30
O2	Podcasting	44
P1	Painting	58
P2	Painting	35
P3	Painting	40
P4	Painting	35
R1	Rollercoasters	24
R2	Rollercoasters	43
S1	Streetwear	79
S2	Streetwear	55
T1	Dating in Middle Age	63
T2	Dating in Middle Age	53
V1	Vintage Audio	34
V2	Vintage Audio	56

Table 3.2: List of anonymized participant IDs, the cluster from which we recruited them, and the length of their interview.

Ethical Considerations

Our study design was reviewed by the Institutional Review Board (IRB) at the University of Washington and was determined to be exempt. As part of the design of this study, we took several steps to protect the privacy of our research participants. Participants were fully briefed about the design of the study before being interviewed and were given documents concerning

the study and contact information for our IRB. Explicit consent was obtained from every participant.

Because this project involved collaboration with a relatively large team, we used the Keybase end-to-end encryption service for all discussion and data sharing. Finally, participants were anonymized so that no direct identifier was recorded in the process of data collection, and only anonymized pseudonyms (e.g., C1, P2, and V2, as show in Table 3.2) are published in this paper. We made several minor edits to quotes to obsure potentially identifying details.

3.4. Findings

Why do people participate in multiple online communities around the same topic? The answer that emerged from our grounded theory is that no one community can provide all the benefits that users want. At a high level, we find that people have multiple and diverse motivations for participation in online communities. In §3.4, we describe the types of benefits they seek organized into three categories: (a) engaging with specific types of content, (b) homophilous socialization, and (c) sharing content contributions with as large an audience as possible. In §4.2, we use data from our interviews to describe the tensions between these benefits. We also investigate how our interviewees understood competition and mutualism—key concepts from ecological studies in social computing—between overlapping communities. Our interviewees overwhelmingly found mutualism to be more consistent with their understandings of overlapping online communities than competition. Our contribution comes in the form of a theoretical framework, grounded in our data, that describes how the full benefits of participating in communities can only be satisfied by groups of communities.

Benefits Users Seek from Communities

Specific kinds of content Content on Reddit is organized into subreddits that define their own topical boundaries. These boundaries may be broad (e.g., news) or narrow (e.g., types of painting media). Moreover, subreddits that prohibit types of content or behavior generate niches for subreddits with different rules. Despite such forms of specialization, multiple communities often welcome the same content and encourage users to "cross-post" material.

A subreddit's topic—what it is about and what content should be posted—is often signified by its name. A climbing enthusiastic explains:

"I think the name itself [`r/climbharder`], kind of specifically points out that: this is not for people who climb hard. It's for people who climb and want to climb hard*er*. " (C1)

C1 describes how the purpose of a subreddit is tied to its name by emphasizing the adjectival suffix "-er" as indicative of the fact that the subreddit is not about achieving elite performance but about improving.

Similarly, a participant in subreddits about drag performance invokes Marshall McLuhan to describe how they know what content to post and where to post them:

> Let's say you were a drag artist and you wanted to show off something that you just created. You would have to go select which community you wanted to show it off in. And I guess among those, [`r/Drag`] would be the one to do that in. But if you're—if you're wanting to show off a piece of artwork or something that you made of a queen from Rupaul's drag race—and the best place to show that off would

> be to go to [r/rupaulsdragrace] and post it there. So it's [a] 'the medium is the message' kind of thing. ... You know where would get the most views [and] where would be the best place to post your content. (D1)

Like D1, our informants had deep knowledge of what kinds of specific content would be appropriate for each subreddit in their cluster.

Specialization also occurred as a form of regulatory arbitrage when one community had formal or informal rules about the kind of content that was allowed. In these cases, we would often hear about an adjacent community where breaking the rules is accepted, perhaps even the raison d'être. For example, r/rupaulsdragrace prohibits spoilers and information about the outcomes of a reality TV show. r/spoileddragrace is a community about the same show that allows spoilers.

This pattern is so widespread on Reddit that it is often signaled in subreddit naming conventions (Hessel et al., 2016). The "meta" prefix signals meta-discussions, often drama-centered, about another subreddit. The "jerk" suffix signals a space for memes, mockery, silliness, or other content unaccepted in the "main" subreddit. Both are commonly understood and were discussed at length by our interviewees. For example, among the Rollercoasters subreddits, R1 described the "jerk" subreddit as a "joke subreddit" where members of the main rollercoasters subreddit could make fun of themselves:

"I would definitely say r/rollercoasters and r/rollercoasterjerk are really deeply intertwined. It's usually all the same members and stuff because of the fact that the coaster 'jerk' is just meant to make fun of the main subreddit. It's just a joke subreddit." (R1)

"Jerk" subreddits were a common source of discussion among our participants.

Among the Climbing subreddits, the "main" subreddit about rock climbing (r/climbing) is welcoming to newcomers. C1 explained that members upvote posts by newcomers"to encourage more entrance into the sport." However, newcomer posts are often repetitive pictures of people climbing in gyms or videos of famous climbers. This annoys some experienced climbers. The "jerk" subreddit provides a backstage space where making fun of newcomers is permitted.

In addition to being divided by rules, interrelated subreddits can be structured as a ladder of "conceptual rungs" where one finds larger communities as one ascends the ladder. A participant in the subreddits on art and painting described this phenomenon as

" You go up through these conceptual rungs. ... When you go up from, say, r/OilPainting—like r/HappyTrees to r/OilPainting—it's a much bigger community. And then from r/OilPainting to Painting, which is even bigger. " (P2)

P2 explained that smaller subreddits such as /r/HappyTrees support learners and are generally more welcoming places. Although the quotation above suggests that the size of communities increases as one moves up conceptual rungs, the relationship between topical scope and size was more complicated. In some topical areas, subreddits with relatively specific topics have the largest and most active communities. For example, /r/rupaulsdragrace is the most active drag subreddit by a large margin, even though it focuses on a reality TV series that is part of the broader drag community covered by r/drag.

Although many specialized subreddits exist, people who want to share their work, ask a question, or have a specific discussion may not know the best place to post. Cross-posting—i.e., when someone posts the same content, questions, or messages in multiple communities—is widespread on Reddit. Cross-posting has sometimes been viewed negatively as a form of

attention grabbing (i.e., "karma whoring") (Poor, 2005). More often, however, we heard that cross-posting was acceptable and even encouraged to establish complementary conversations or find different audiences.

Multiple interviewees from the Climbing cluster, including C1, described how, when people ask for training advice in `r/climbing`, the largest subreddit about rock climbing, they will be advised to cross-post to `r/climbharder`:

"Somebody will post asking for advice in `r/climbing` and oftentimes, somebody will comment and be like, 'Hey, you know? You're welcome to ask this here, but you might get more and better responses at `r/climbharder`.'" (C1)

C1 explained that even though conversations about training often start in the main subreddit, they are not likely to gain traction because not everybody in the main community is interested in the more intensive aspects of climbing.

In sum, the ecosystem of subreddits about similar topics provides more opportunities for people to find specific desired discussions. People receive positive feedback and engagement when they post content that fits a subreddit's specific topic. That said, the subreddit where a particular piece of content will be best received is often not clear to the person posting it. Cross-posting provides multiple chances to start a desired discussion.

Homophily Online communities have long been recognized as a way to "find my people" by bringing together users who share things as diverse as a psychiatric diagnosis, enthusiasm for a hobby, or membership in a subculture or identity group. A member of the Middle Age Dating cluster of subreddits explains:

> [When I joined the ADHD Reddit sites], I feel like I found my people after all these years. ... If you don't have ADHD, and don't wonder what's going on other people's brains all the time, I think you just think that everybody thinks like you. And they don't. They don't. So if you're 30 and you're having a problem, you really just want to talk to other 30 somethings. (T2)

T2's description of having "found my people" and talking to other people like themselves invokes the idea of homophily: the desire to connect to others similar to oneself. Analytically distinct from finding personalized information in narrowly focused subreddits, homophily was frequently cited as an end in itself by our interviewees. Our interviewees sought to connect with "like-minded" people having similar interests, demographics, identities, tastes, and status.

Even though the identities of others in subreddit communities are largely invisible, participants can easily imagine the demography of the subreddit. A participant in the Drag cluster of subreddits described `r/Dragula`, a community of fans of a TV show featuring horror-infused drag styles, as follows:

" I think it would be a mostly LGBTQ audience. And not many straights. But if there are straights, they would be really open minded or edgy. Or, I don't know ... associated with that 'dark' aesthetic." (D2)

D2's thoughts on `r/Dragula` convey a clear sense of the audience the subreddit. Of course, the pseudonymous nature of Reddit obscures age, race, gender, and ethnicity. That said, Reddit users draw on stereotypes about fanbases and cues such as mentions of schools, selfie posts, linguistic markers, and cultural references to build clear models of the types of people in a subreddit.

In further unpacking these dimensions, D2 contrasts r/Dragula with the more mainstream subreddits about the show *Rupaul's Drag Race*:

> [As for subreddits about] the drag race (r/rupaulsdragrace), Drag Race UK (r/RPDR_-UK), and the spoiled drag race (r/SpoiledDragRace). ... Most of [the participants in these other groups] don't do drag. Most of them are, I think, white gay men, or straight women who see drag with a very narrow view of what drag is. Hegemonic? I don't know if that's the word, but they apply the same standards of beauty that are applied to women and men and artists and performers to this art form. (D2)

D2 conveyed both a strong sense of the demographics of different drag subreddits and a strong sense of identification with r/Dragula, which they described as less toxic, more inclusive, and more creative, in part because its membership has a greater concentration of LGBTQ and non-White people who are less interested in conforming to hegemonic beauty standards.

Subreddits divide broad topical areas such as drag, art, and fashion into subgroups of people occupying strata of status hierarchies associated with identity, expertise, and class. For example, in the Climbing cluster of subreddits, rock climbing ability confers status and separates beginners from advanced athletes. We found that these two groups concentrate their participation in different subreddits. Across the clusters, we found that experts sought out fellow experts with whom to share knowledge, offer reflections, and give advice grounded in shared extensive experience.

Our Streetware interviewees reported that subreddits about fashion are split along lines that are associated with the price and status of the clothes being discussed:

"The kind of person, the Platonic ideal poster or user of something like r/streetwear, is probably more open-minded, maybe, in terms of what they think is cool, what they think is worth wearing. Whereas, you know, ... the r/malefashion snob is a snob." (S1)

Even though users of r/streetwear share and discuss men's fashion, r/malefashion, which focuses on higher-status and more expensive styles, looks down on their casual and youthful styles. S1 is a member of the r/streetwear subreddit. Although their groups are "chill" and "supportive," higher-status groups are "snobby." It is clear that S1 feels unwelcome and out of place in the higher-status group.

Similarly, our interviewees described status hierarchies in Painting subreddits related to skill level and medium. P4 described how they were invited to cross-post their work from r/Watercolor to r/Artoilpainting, a smaller subreddit that seems to have a complicated relationship with watercolor. Although watercolor submissions are allowed, and, in this instance, encouraged, both the subreddit's name and the similarity between its visual tag for watercolor submissions with the downvote button suggest that oil is the preferred medium in this community. In this way, the division of topical spaces into spheres of similar status and identity allows members to find groups that exclude both those who look down on them and those who they look down upon.

Although "finding your people" is satisfying in itself, it can also be a foundation for a wide range of other kinds of benefits. For example, a homophilous community leads to conversations that can promote trust. Trust has many benefits such as building confidence in the advice and information shared within a community. In some communities we studied, this trust enabled buying, selling, and trading of material goods.

V2, one of our interviewees from the Vintage Audio cluster, described a community of record collectors on Reddit that acted as a market for buying, selling, and trading records. They preferred this subreddit to other online markets such as Ebay because the community holds members accountable for honest transacting and because of the intrinsic reward that comes from sharing records with a fellow community member:

> Because it's a group of people that are like-minded, … your feet are kind of held to the fire a little bit more about actually being realistic with the condition [of the material you are selling]. Whereas, [when you buy] vinyl at the used record shop, sometimes you feel like someone's trying to pull one over on you … I feel like because it is a community, sometimes you can get some kind of better deal …
> I found other people that share the hobby that I like. So I almost, definitely, feel like they're friends in a little way. And so I want to, if I'm ever selling, I'm going out of my way to make sure that whatever I'm doing, everything I'm doing, is above board. (V2)

V2 was very enthusiastic about the "marketplace wrapped in a community" for vinyl records. According to V2, both buyers and sellers of records benefit from transacting within a community of like-minded hobbyists. Because the community holds sellers accountable, the community promotes honest representation of merchandise. Being part of a like-minded community where members feel friendship with each other gives sellers a reason to be honest, and even to discount their wares, because they get "some kind of better deal."

In sum, our interviewees turned to specific subreddits to find people who share their interests, tastes, problems, and identities. Our participants described subreddits in terms of demographics and identity groups as well as styles, subgenres, or categories related to social status such as wealth, expertise, and beauty standards. They used these categories to place themselves within the constellation of related subreddits they participated in. Members of subreddits who are "finding their people" benefit each other by acting as communities as well as building trust and feelings of friendship. Over time, these feelings can provide further benefits such as the ability to more safely engage in buying and selling.

Finding the largest possible audience A third type of benefit derives from the number of members in a subreddit. All our interviewees were keenly aware of the fact that a post reaching one of the top positions on a larger subreddit would receive the attention of a vast audience. They described this attention as emotionally thrilling and otherwise beneficial. For artists and influencers, large audiences brought material rewards. For learners, a large audience's collective knowledge could bring hard-to-find answers and advice.

That said, our interviewees explained that larger subreddits do not necessarily provide a larger audience because posts in larger subreddits are more likely to be ignored or missed in the torrent of other content. Although posting in a smaller subreddit might increase the chances of finding an audience at all, subreddits that were too small were described as unattractive because they would not attract many posts or replies. Interviewees responded by choosing where to post strategically.

Although the competition for the top spots on the front page of large subreddits can be fierce, this competition can make recognition from a large subreddit extremely gratifying:

> Likes are just kind of fake: fake social currency. But yeah, when you get a charge out of it, yeah, I love it. Most of the time, painting is a really busy sub. I mean, like, in any given hour, the new page is already replaced. . . .
> If you can get something that gets a hold there and stays on the front page for a little while, [if] it gets up in even the top five, I've had a handful do that. That's kind of cool. (P2)

P2 describes the thrill of reaching top positions in `r/painting` with posts of their paintings. Even though they are dismissive of likes on Reddit, they desire the attention their work gets from the subreddit. It sends traffic to their websites, raises their artistic profile, and helps them sell their art. Although these material incentives are important, part of the thrill comes from knowing that a given subreddit is competitive. Smaller subreddits are simply unable to provide these benefits.

However, posting in a large subreddit means the risk of being ignored:

"I think there's this weird bell curve where the community needs to be big enough where people want to post content. But it can't get too big where people are drowning each other out for attention." (S2)

S2 was among several of our interviewees who described an ideal "middle ground" for subreddit size. In general, we heard that people were less interested in posting content in very small subreddits that do not provide an audience. Thus, competition over the largest audiences drives people to smaller subreddits where they can reliably find an audience. I2 from our Indie Music cluster explained:

> Usually `r/Indieheads` is the way to reach more people if you want to. Just like if you wanted to do even more, you'd probably do it on `r/music`. . . . Say a small indie band decided to do an AMA they would probably want to do it on `r/Indieheads`. Because if they did it on `r/music`, it would get drowned out and nobody would see it because there's so many posts. In `r/Indieheads` it would get a decent bit of attention, I think. In the band subreddit, it would probably get a lot of attention too. But `r/Indieheads` seems like the best middle ground for that kind of thing. (I2)

I2 explained that when the psych-rock band *King Gizzard and the Wizard Lizard* wanted to engage with an audience on Reddit, they had a choice whether to post in the smaller "band subreddit" dedicated to them, the very large `r/music`, or the medium-sized `r/Indieheads`. Although posting in the band subreddit would have surely provided an audience, they chose `r/Indieheads`, which was large but where there was still little risk that their post would be drowned out.

Our interviewees repeatedly described how finding an audience for one's content is a clear motivation for posting in larger subreddits. However, we also heard that competition for attention in the largest subreddits leads people to try to find an audience in smaller subreddits. In the smallest subreddits, posting may not seem worthwhile at all. This trade-off between finding a large audience and being ignored suggests that posting in subreddits of intermediate size can be the most reliable way to reach a sizable audience.

Tensions Between the Benefits

The findings in the previous sections imply a clear reason that so many overlapping subreddits exist. When one subreddit prohibits a certain type of content or conversation, an adjacent group can form that allows it. When an identity group is marginalized in one subreddit, members of that group may form a subreddit of their own. When getting attention in a large subreddit is too difficult, a smaller subreddit becomes attractive. Using data from our interviewees, we describe each of the three possible tensions that exist between the three benefits: (1) subreddits where one finds a large audience are less able to provide specific types of content; (2) communities with large audiences are rarely able to provide a community of similar others; (3) some valuable types of discussion and information are found only in diverse groups of people. As we discuss in §3.5, taken together, these tensions form a "trilemma"—i.e., a choice with three mutually incompatable options—between our interviewees' desires for specific content, homophily, and finding audiences. A single community might provide two of these benefits, but almost never all three.

Larger audiences create background noise In §3.4, we described how subreddits are structured according to distinctions between different types of content. Breaking topical areas into subreddits of varying levels of granularity makes finding specific content easier because doing so reduces the need to sift through unrelated material in a large and broad subreddit. Our interviewees often expressed that larger subreddits are simply not the best places for enthusiasts to have discussions:

> I see this background noise problem building [in] `r/climbing`, the main climbing community, [which] has just become less and less and less interesting and less relevant as it's gotten bigger. That's not really a problem. Right? That's probably has more to do with my interest level and how long I've been on it. And my experience level with climbing. I'm just a little bit more crusty about it, you know? (C2)

C2 describes losing interest in the primary subreddit about climbing as it grew because of the interviewee's specific interest in particular types of climbing content (i.e., material associated with being "crusty" or experienced). C2 recognizes that when `r/climbing` experienced growth, the larger volume of posts by newcomers to the sport created a "background noise problem" that made it difficult for established climbers to find discussions of interest.

Similarly, smaller subreddits can be incredibly valuable to those looking for highly specialized information. Even though they may have very low levels of activity, they can provide a way to learn about rare forms of expertise. A participant in our Vintage Audio cluster explained how they might seek out advice on building a reel-to-reel audio setup:

> If you're at [`r/ReelToReel`]. Everybody is hyper into them. Whereas there's probably overlap with somebody in `r/vintageaudio` ... If I'm like trying to rebuild my reel-to-reel player, I want to talk to ... the most knowledgeable person particularly about building reel-to-reel ... So I know that who I'm talking to is hyper specific to the knowledge I want. (V2)

Invoking `r/ReelToReel`, V2 describes a highly niche subreddit about archaic audio tape equipment with only 3,200 subscribers and a handful of posts each day. V2 is simply not looking to find a large audience. Instead, they want access to the "most knowledgeable person" with

specific expertise because access to this expertise makes it possible for them to consider doing their own reel-to-reel projects.

Although the r/ReelToReel community overlaps with the larger and more general r/vintageaudio, the latter does not provide the ability to connect with a small group of expert enthusiasts in an old-fashioned technology.

Similarly, when someone wants a podcast recommendation tailored to their personal tastes, asking in a larger subreddit is not likely to prove as fruitful as it is within a smaller one. O2, a participant in the Podcasting cluster explained:

> So I think for like r/audiodrama, I would probably write a longer post, and probably get a bit more into like, my personal tastes. Like I would comment about, 'oh, I really love the acting in this one, is there anything similar?' Open up a bit more about what I do and don't like. Whereas I think in podcasts, it probably would be more direct. I'd ask a specific question ... more to the point, more factual, probably just more almost transactional. (O2)

Although the larger r/podcasts subreddit is a popular place to promote podcasts on Reddit, O2 explains that they prefer asking for recommendations in the smaller r/audiodrama where they find others willing to take their personal tastes into account. Our interviewees did not advance a "smaller is better" argument. O2 explains that they still engage in larger subreddits but use a more direct and transactional approach to information exchange when they do. Similarly, large art communities provide opportunities to find a large audience, but someone can find more substantive feedback to improve their skills, by posting in a smaller subreddit organized specifically for this purpose.

Interviewees described the most general interest-based subreddits such as r/podcasts, r/painting, and r/climbing as more accessible and welcoming to newcomers and as reaching a larger audience: all things they valued. They also described these larger groups as having a high volume of low-effort posts or comments. Our interviewees explained that although they play a useful role in an information ecosystem, the largest subreddits in a topical area are rarely the best places to look for information or advice. They explained that small subreddits can effectively play host to content, information, and discussion that larger subreddits cannot.

Homophily is more difficult in larger groups Because they have less background noise, smaller subreddits are more likely to provide better opportunities to connect with people who share one's distinctive interests, tastes, and identity. Smaller subreddits are also better places to find a community because they provide opportunities to have repeated encounters with recognizable others, off-topic discussions, and personal interactions. P4 explained:

"Obviously, I want as many people to see my stuff as possible, especially [since I am] trying to establish myself. But at the same time, I do want to build a relationship with any sort of community that I can." (P4)

P4 explained that they participated in multiple communities because they have two goals as an artist. First, they want to find an audience for their artwork to establish their career. Second, they want to build a community with others who share their craft. They felt that they needed to turn to multiple subreddits to satisfy both needs.

Although larger subreddits provide a large potential audience, smaller subreddits were described as being friendlier. Another interviewee from our Painting cluster explains that this is because of how people act differently in large and small subreddits:

> I live in the middle of nowhere. And every so often, before the pandemic, I would visit the [large city several hours away]. Now I found there were very polite people, both in [the city] and in [my rural area]. But the tone by which people carried themselves changes in their environment: that's kind of one of the big changing factors. So, in the city, people are in a rush, they're about their business. We don't really have time to chat. ... The big subreddits might seem unfriendly [but] it's not that so much. Individual members are impolite or unfriendly. But it's almost as though people carry themselves differently when we're in different subreddits. (P3)

In their extended metaphor, P3 explained that large subreddits are like big cities full of busy people who do not "have time to chat." Evocatively, they described people as behaving differently in large and small subreddits. The very same people who are rude in large subreddits might be friendly in smaller subreddits where people have repeated encounters with one another and have a stronger sense of knowing each other. In another quote from the same cluster, P2 described how the small subreddit for Bob Ross-inspired painters, `r/HappyTrees`, stands out from the larger art subreddits because people know one another and it does not feel anonymous. The tight-knit nature of this community contributes to its utility as a source for feedback.

Tension between finding specific content and homophily A third tension described by our interviewees is that between the desire to connect with similar others and the desire for forms of discussion, content, and feedback that can only be found in diverse groups inclusive of dissimilar others. Our interviewees described a range of situations when they sought out dissimilar others. For example, they described beginners seeking to learn from experts and outsiders seeking to learn about other cultures. They also described how subreddits instituted rules to limit or organize content that also interfered with unstructured and off-topic discussions that helped with community building.

For example, although multiple subreddits with overlapping users discuss the same episodes of the TV series *Rupaul's Drag Race*, they have different understandings of events in the show depending on their national identities. D1 explained:

" The discussions played out differently on different subreddits. In the Drag Race UK sub there's a lot more understanding about [a British drag queen] in particular, about where they come from ... In America we don't understand how that person is from Worcestershire. " (D1)

D1 explained that the cultural background of one of the drag queens was a subject of discussion in `r/RPDR_UK`, the UK drag race subreddit, while the main subreddit, `r/rupaulsdragrace`, was "dominated by the American viewpoint."

Our interviewees described a number of subreddits focused on discussing broad topics from a specific national or regional culture context. These cultural communities within a topical area provide a homophilous space for sharing distinctive cultural knowledge and sensibilities. The wrinkle is that even for our American interviewee D1, the `r/RPDR_UK` subreddit provided an opportunity to enhance their own experience and appreciation of the show by observing and learning from members of another culture. In examples such as these, our interviewees

explained that communities where like-minded people can share their distinctive appreciation show could provide a source of knowledge for outsiders.

Similarly, Painting participant P2 explained that a group that has a mixture of experts and beginners provides a better learning environment than does a group of beginners alone:

> If you can find a small group, with a small core of people who are particularly skilled, they sort of energize the group as a whole. `r/HappyTrees`, even though it's kind of a beginner subreddit, there's some people that posts there that are like, you know, Bob Ross instructors, or they've been doing this for years. And they've mastered that sort of ... "happy trees" thing. (P2)

P2 explains that part of what makes `r/HappyTrees` great is that it connects learners to experts. A homogenous subreddit of only beginners or experts would not provide the same opportunities.

To stay focused on specific types of content, subreddit moderators will frequently employ strict rules and heavy-handed moderation. Our respondents explained that smaller subreddits can get by with fewer rules and lighter moderation because they have fewer behavior problems and are less attractive to toxic outsiders. They are also more able to self-police using Reddit's voting system and through direct interpersonal sanctions such as admonition. In the words of one of the Vintage Audio participants,

"In Reddit, the more users you get, the more strict the rules, and the more strict the moderation. Just to prevent problems. " (V2)

V2 continued and explained that when a subreddit is small enough that you can "wrap your hands around" and is built around a "like-minded" group, it can develop and enforce shared behavioral norms that substitute for formal rules and rigid enforcement regimes. V2 explained that the processes of creating spaces for specific types of information got in the way of building community.

Similarly, one of our interviewees described `r/Indieheads`'s rules limiting how often one can post, requiring specific titles and tags, and prohibiting types of user-generated content. Although these rules help maintain a high-quality feed, they also prevent sharing of more personal and relatable forms of content such as amateur performances and chit-chat. As a result, subreddits that make rules to ensure that posts are on-topic frequently have adjacent "-jerk" subreddits that provide an outlet for jokes and memes and act as places where off-topic discussions can thrive.

Interviewee's Understandings of Competition and Mutualism

Except for a small qualitative subpart of a single paper (Zhu, Chen, et al., 2014), prior ecological studies in social computing have relied on concepts such as competition and mutualism but have provided limited evidence that such concepts are salient to participants. As part of our interviews, we asked our interviewees if they perceived relationships between the communities they participated in to be competitive or mutualistic. In some cases, interviewees imagined hypothetical scenarios where competition might emerge from the perspective of subreddit moderators. For example, a participant in Climbing said:

"I guess if you put your Reddit [moderator status] on your resume or something, and you want to be a moderator of a larger community, you could try to get users from other communities. But I haven't seen or experienced competition." (C1)

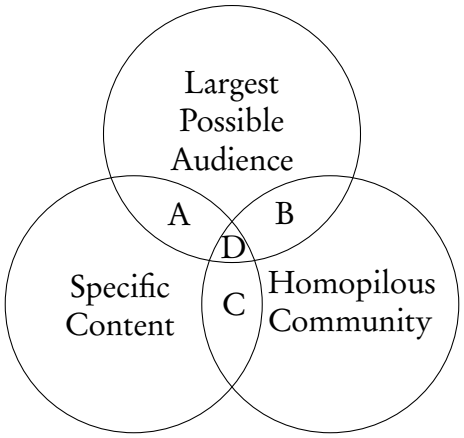


Figure 3.1: Venn diagram illustrating the specificity-homophily-audience "trilemma."

Although we asked nearly every interviewee about competition, only one interviewee (S2) described an actual instance of conflict or direct competition. In nearly every other interview, our subjects found our suggestion that subreddits might be in competition to be surprising and strange.

However, the idea that communities are complementary and mutualistic was much more intuitive. One Vintage Audio participant explained the relationship between subreddits:

"Yeah, the overlapping. … They each have their own niche. … They get big enough to have super critical mass of people. Then they'll have a reason to exist. And then they'll sort of fit into the ecosystem of different communities." (V2)

Consonant with this description of subreddits in unproblematic coexistence, our interviewees repeatedly suggested that there were not meaningful structural or technical limitations on the number of subreddits a user can join and this reduced the possibility of competition, if it did not eliminate it altogether.

3.5. Discussion

The tensions between the benefits that our interviewees sought can be thought of as forming a "trilemma" between finding specific content, homophily, and finding as large an audience as possible. This three-way dilemma captures the fact that the more a subreddit succeeds in providing any one of these benefits, the less able it will be able to provide the others. A portfolio of overlapping communities solves this problem by providing all three types of benefits.

Figure 3.1 visualizes the theorized trilemma. Each of the benefits described in §3.4 is reflected in large circles. Each of the tensions described in §3.4 is reflected in the overlapping areas in the figure. Area A contains communities that provide the largest possible audience and specific content but are unlikely to provide homophily to community members. Subreddits that provide large audiences face "the background noise problem" as a large volume of submissions makes it difficult for people to find the specific content they care about. Area B contains communities that offer both large audiences and homophily but that will struggle to provide specific content. For example, an American interested in learning about international drag culture finds the need to search beyond `r/rupaulsdragrace`. Area C contains communities that provide specialized

content and a homophilous community but that may not attract large audiences. Although not everyone who desires a specific type of content may be similar to those who produce the content, smaller subreddits can often provide both desired content and opportunities to socialize with similar others. However, as the size of the audience increases, subreddits encounter the background noise problem and acquire a "big city" air of unfriendliness.

Connections to Prior Research

Finding specific content Our findings are consonant with prior work that the primary benefits provided by online communities stem from their power to connect people to novel and hard-to-find sources of information (Benkler, 2006; Campbell et al., 2016; Fiesler et al., 2017; von Hippel, 2016). Our study adds to this work and complements recent findings of Hwang and Foote (2021) by describing how nested and overlapping online communities are useful for information seeking and managing one's information exposure. Individuals often desire multiple types of content within a general subject area such as spoiled and spoiler-free discussions. Even when a relatively obscure community such as `r/vintageaudio` exists, an even more specialized community such as `r/ReelToReel` may provide access to an even more specialized set of experts.

Finding homophilous community Prior work has recognized the importance of homophily in motivating and structuring participation in online communities (Chang et al., 2014; Cunha et al., 2019; Grevet et al., 2014). Contributing to this line of research, we identified a number of types of homophily that drive an individual's decisions to participate. These included hobbies, expertise, age, national culture, identity, and status. Homophily was in tension with the need for specific content in that differences among many of these dimensions were valuable for finding information.

Our results suggest that participants in online communities face trade-offs between homophily and information novelty. These may be similar in structure to the trade-offs between short and long ties observed in contexts such as work groups (Ruef et al., 2003) and social networks (Granovetter, 1973; Grevet et al., 2014). One advantage of joining a group of overlapping online communities is that it can help find information that would be unavailable in homophilous groups.

Finding the largest possible audience Much social computing research points to the benefits of large audiences and large communities (Kraut et al., 2012). Our work adds more evidence to back up those claims. More relevant, perhaps, are recent counterclaims about the benefits of smallness. Hwang and Foote (2021) presents an interview study with members of small Reddit communities. Although our results about the tensions between large audience size and other benefits are fully in line with Hwang and Foote's findings, our starting assumptions and ultimate takeaways are quite different. Hwang and Foote seek to understand why people participate in persistently small communities and conclude that smallness offers a range of benefits. Our results suggest that individuals seek out benefits that happen to be incompatible with largeness and participate in portfolios of communities that, because of the trilemma we described, will almost certainly include small ones. Although we believe that Hwang and Foote's (Hwang &

Foote, 2021) emphasis on smallness might draw focus to a side effect instead of the cause, we believe that the findings in our two papers are largely complementary.

Although users may desire large audiences, large online communities often require additional structure to maintain order(Gillespie, 2018; Kiene et al., 2019; Kiene et al., 2016). Kiene et al. (2016) describes how a massive influx of newcomers presents difficulties that can be managed by appointing additional moderators, increasing norm enforcement, and limiting the frequency of posts. Lin et al. (2017) find that such interventions help subreddits maintain comment quality and stay on topic during massive influxes of growth. Our sense is that these changes ensure the availability of specific content, in part, because of the growth-limiting effects of rules and enforcement (Halfaker et al., 2013; TeBlunthuis et al., 2018). We see this as yet more evidence in favor of our theory.

Implications for Ecological Studies in Social Computing

The quote by V2 in §3.4 can be read as a kind of summary of resource partitioning theory (RPT), a strand of ecological research in organizational science that focuses on explaining specialization (Carroll, 1985). Although RPT has not been deeply examined in prior social computing work, our findings suggest that it may be able to explain the widespread occurrence of overlapping communities. RPT proposes that the reason that small specialized organizations coexist with large generalist organizations is that generalists are constrained in their ability to meet distinctive needs in niche markets (Carroll & Swaminathan, 2000; Swaminathan, 2001). In V2's terms, the "ecosystem of different communities" is constructed by a process in which those that "have a reason to exist" and are specialized to "have their own niche" will achieve "critical mass."

Our grounded theory suggests that the trade-offs in the capacity of an online community to provide different types of benefits that people seek from online communities give rise to new niches. On the basis of our findings and our understanding of RPT, we hypothesize the following process to describe how systems of overlapping communities develop:

When a new topical area grows, the bulk of activity will happen in a generalist community. New members joining that community may seek and find the perceived benefits described in §3.4 (i.e., specific kinds of content, homophily, and the largest possible audience). If a topic area, such as art, is sufficiently general, initial membership growth occurs as the community attracts new and existing users interested in both general and more specific types of content.

As growth continues, membership in the generalist community becomes heterogeneous with lower levels of homophily (e.g., amateur and professional artists) and more specific interests (e.g., painters and photographers) and types of engagement desired (e.g., attention from an audience or critique). At this point, the trade-offs we discuss in §4.2 related to size become relevant. Finding information related to a specialized subtopic and homophilous socializing grows difficult.

If, as with Reddit, creating new communities is low cost, a community specialized in a subtopic can emerge. This specialized community will likely not attract as large an audience as the generalist community. However, those most interested in the specific subtopic will join it to escape what our interviewees describe as "background noise" in the larger generalist community. Similarly, those seeking personal interaction or social bonding with other community members will be more likely to find them in the specialized community. A similar process occurs in the formation of spaces having different rules or purposes (such as "jerk" spaces). The cycle will then begin anew as subreddits repartition a subtopic such as `r/painting` into subspecialists such

as r/oilpainting and r/watercolor. Although some of our interviewees described parts of this process, the model we have narrated is an untested theory. We leave it to future work to establish its empirical validity.

Implications for Design

By allowing users to create multiple communities with similar or identical topics, platforms can host ecosystems of online communities capable of providing a larger range of benefits to a larger range of users. Some platforms, such as Stack Exchange, prohibit new communities from overlapping with existing communities (Fu & Stvilia, 2016). Our findings suggest that such rules limit the range of the benefits the platform's communities can confer.

Existing designs for online community platforms such as Reddit are at best "first-order approximations" of an ideal solution in that a "sociotechnical gap" remains between these designs and the goal of a platform that meets every person's every need (Ackerman, 2000). Our interviewees partly filled this gap with personalized bespoke solutions in the form of their handpicked portfolios of communities. Improved designs for multi-community discovery and engagement can better support users in knitting together portfolios of communities.

Many Reddit users make heavy use of the aggregated streaming feeds r/all and r/popular, which surface highly upvoted posts from across Reddit. Our interviewees described these feeds as most often featuring content from subreddits that are already extremely popular. Furthermore, Reddit's system for recommending subreddits often returned irrelevant suggestions. Suggesting communities in as many cells in Figure 3.1 as possible could help users build their portfolios of communities. Because increased visibility may create stress and labor for communities and moderators (Kiene et al., 2016), recommendations should target those potential members likely to be positive contributors.

Although some of our interviewees used the "multireddit" feature for making a custom feed of subreddits, they described this feature as cumbersome and overwhelming. A design alternative is to formalize or even automate the types of informal social practices our interviewees described such as cross-community linking and cross-posting. For example, a subreddit such as r/vintageaudio might configure an auto-moderator to detect posts about reel-to-reel equipment and recommend cross-posting to r/reeltoreel. A discussion-focused subreddit might routinely invite productive contributors to discussions in the related "main" subreddit. Because intercommunity interactions can give rise to conflict, individual communities should have control of how such practices are implemented. New tools for collaboration between moderation teams may enable the institution of policies encouraging productive concurrent participation in overlapping communities.

Limitations

Our study has limitations common to all interview-based studies. Our findings derive from in-depth conversations with relatively few of the people who were highly active participants in the handful of clusters of communities in our sample. Although our study was designed to achieve analytic saturation within each cluster and to cover a wide range of types of topics discussed on Reddit, additional interviews across a wider range of communities might uncover new types of specialization. Additionally, our interviewees were among the most active members of the

clusters, and their experiences may differ from those of peripheral members. Similarly, we cannot speak to the experiences of those who participated in only one community within a cluster.

Our interview data were collected at one point in time and cannot speak to how the dynamics we describe played out over time or how new communities were created and emerged. Relatedly, although we find that overlapping communities tend to provide different benefits to members, we did not set out to interview community founders and thus cannot speak to the reasons that communities were created (Foote et al., 2017).

Furthermore, our study focuses only on the Reddit platform. Reddit has distinctive affordances for voting, moderation, and multicommunity engagement that might shape the construction and use of overlapping communities. Although Reddit is among the most popular online community platforms. Our findings may not describe relationships between overlapping communities on other platforms, or between one platform and another. Different platforms likely have different strengths or weaknesses for building communities that provide some types benefits but not others. At the same time, cross-platform engagement may involve frictions related to the use of multiple identities and sociotechnical systems. Future research should investigate how people use portfolios that include communities on multiple platforms.

3.6. Conclusion

Why are the same people talking to each other about similar things in different online communities? We answer this question by developing a theory grounded in the analysis of 20 interviews with members of highly related communities on Reddit. Our answer suggests that people turn to online communities in search of multiple benefits—specific kinds of content and discussion, socialization in a homophilous community, and attention from the largest possible audience. We argue that although structures such as the topic, rules, and size of a community might improve the degree to which it provides one of these benefits, they will necessarily detract from its ability to provide others. Multiple communities having a range of structures exist to provide the full range of benefits. No community can do everything.

Preface to Chapter 4

As was the case with Chapter 3, Chapter 4 is written as a stand-alone article building upon Chapter 3. It repeats some of the same motivating points in the first paragraph of §4.1, and the first two paragraphs of §4.2.

This study also reuses the clustering procedure from Chapter 2, but on a larger dataset. The first three paragraphs of §4.3 describe the clustering procedure. Those who have read Chapter 2 may quickly pass over these paragraphs, noting that the sample size, dimensionality of LSI, and the number of clusters are different from Chapter 2.

Chapter 4

Dynamics of Ecological Adaptation in Online Communities

We introduce a method for inferring competitive and mutualistic interactions between online groups from time series participation data based on the theoretical framework of community ecology. Platforms often host multiple online groups with highly overlapping topics and members. How can researchers and designers understand how interactions between related groups affect measures of group health? Inspired by population ecology, prior social computing research has studied competition and mutualism among related groups by correlating group size with degrees of overlap in content and membership. The resulting body of evidence is puzzling as overlaps seem sometimes to help and other times to hurt. We suggest that this confusion results from aggregating intergroup relationships into an overall environmental effect instead of focusing on networks of competition and mutualism among groups as our approach does. We compare population and community ecology analyses of online community growth by analyzing clusters of subreddits with high user overlap but varying degrees of competition and mutualism.

4.1. Introduction

Although the fact is frequently ignored in social computing scholarship, online groups do not exist in isolation.[1] Indeed, although studying interdependence between online groups is different and complex (Hill & Shaw, 2019), research in social computing has sought to quantify how online groups share users or topics (Datta et al., 2017; Del Tredici & Fernández, 2018; Hessel et al., 2016; Tan & Lee, 2015), and how such interactions relate to outcomes like the emergence of new groups (Tan, 2018), contributions to peer-produced knowledge (Vincent et al., 2018), and the spread of hate speech (Chandrasekharan et al., 2017). Although this work has demonstrated that intergroup interactions matter very little intergroup research has tackled questions of group success—i.e., why some online groups succeed in maintaining active and long-lived participation while most do not. Can intergroup relationships explain whether online groups will grow or decline?

Studies in social computing have drawn from organizational ecology to answer this question (Resnick et al., 2012; X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). Inspired by the ecological study of biological systems, organizational ecology is an influential body of theory in sociology that studies competition and mutualism among human

[1] We use the term "online group" instead of "online community" to help avoid confusion with our term "community ecology" which plays an important conceptual and analytic role in our paper.

organizations Although ecological studies of firms and social movements have developed a clear and established body of theory with strong empirical support (Baum & Shipilov, 2006), similar studies of online groups have yielded inconsistent results that differ both from one context to another and from theoretical predictions. For example, wikis whose memberships overlap with other wikis survived longer (Zhu, Chen, et al., 2014), but Usenet groups with overlapping memberships failed more quickly (X. Wang et al., 2012).

We argue that these confusing results are the result of a conflation of concepts and measures from two distinct strands of theory in organizational ecology: *population ecology* and *community ecology*. Both define competition as a form of interdependence that *decreases* growth and mutualism as one that *increases* growth. However, population ecology focuses on modeling the how overlapping resources among groups affect their subsequent growth, decline, or survival (Astley, 1985; Baum & Shipilov, 2006; Dobrev et al., 2001). It does not attempt to directly study competitive and mutualistic interactions. On the other hand, community ecology recognizes that groups often exist within "ecological communities," or clusters of highly related entities, and provides an approach for inferring competitive and mutualistic interactions among these. Although the stated goal of ecological research in social computing has been to understand how groups influence each others' ability to sustain participation, ecological research in social computing has relied exclusively on concepts and measures from population ecology. This paper seeks to explain the puzzling set of findings in ecological social computing research by introducing community ecology.

We do so in a three-part empirical study using a dataset drawn from the 10,000 communities on Reddit with the most contributors to analyze 641 clusters of online groups with overlapping participants. In Study A, we conduct the most important type of population ecology analysis, a test of what is called density dependence theory, and find support for the theory. This analysis suggests that high degrees of user overlap are associated with competition. In Study B, we introduce our method for community ecology analysis that infers networks of competitive and mutualistic interactions by using clustering analysis and vector autoregression (VAR) models of group size over time (Canova, 2007; Ives et al., 2003; Sims, 1980). We illustrate the method in four case studies and present a large-scale computational analysis showing that mutualistic interactions are far more common than competitive ones. Finally, in Study C, we bring Study A and Study B together to compare population ecology and community ecology by extending the density dependence model from Study A with a variable accounting for competition and mutualism. While we find that adding this variable does not help predict growth, including ecological interactions in our VAR models improves time series forecasting.

We discuss how these findings illuminate the differences between population ecology and community ecology and show how the two perspectives are complementary. While Study A suggests that competition is strongest when user overlap is high, Study B finds widespread mutualism among groups with overlapping membership. Although these findings might seem contradictory, they reflect how population ecology studies overlapping resources related to favorable or unfavorable environmental conditions, while community ecology studies competitive and mutualistic interactions playing out in local networks of specific groups. By demonstrating that mutualistic and competitive interactions within clusters of highly related groups are important—and by describing how to measure them—this paper lays the groundwork for future research to investigate and design for interdependence between online groups that supports their growth and success.

4.2. Related Work

Online groups are important sites for social support (De Choudhury & De, 2014), entertainment (Ducheneaut et al., 2006), information sharing (Benkler, 2006), and political mobilization of disinformation campaigns and protest movements (Benkler et al., 2013; Choudhury et al., 2016; Krafft & Donovan, 2020). Although an online group's ability to achieve its goals depends on attracting and retaining contributors, few develop a sizable group of participants (Benkler, 2006; P. DiMaggio et al., 2001; S. L. Johnson et al., 2014; Koh et al., 2007; Kraut & Fiore, 2014). Many attempts to explain the success and growth of online groups look to properties of individual groups like characteristics of founders (Kraut & Fiore, 2014), language use (Danescu-Niculescu-Mizil et al., 2013), turnover (Dabbish et al., 2012), and designs for regulating behavior (Halfaker et al., 2013; TeBlunthuis et al., 2018).

Recent research suggests that interdependence among online groups is also important to explain success and failure (Cunha et al., 2019; Kairam et al., 2012; Tan, 2018; Tan & Lee, 2015). For example, banning hate subreddits reduced hate speech in related subreddits (Chandrasekharan et al., 2017). In a very different context, there is evidence that Reddit and Stack Overflow receive substantial benefits from activity on Wikipedia (Vincent et al., 2018). Our work contributes to this literature by providing a new conceptual lens and statistical method for studying competition and mutualism between online groups.

Online Groups Depend on Resources

Like prior ecological research in social computing and information systems, we build on resource dependence theory (RDT) (Butler, 2001; X. Wang et al., 2012). Butler (2001) introduces RDT to argue that growth in online groups is driven by positive feedback as participants contribute resources such as content, information, attention, or social interactions, which motivate further contributions by subsequent participants. That said, online groups do not grow forever and RDT explains that growth is self-limiting because costs of participation increase in larger groups (Butler, 2001; Butler et al., 2014).

Ecological approaches recognize that interrelated online groups may share resources with one another in ways that constrain their growth and survival. *Rival* resources like participants' time, attention, and efforts raise the possibility of competition because they become unavailable to others when used by one group (Benkler, 2006; Kubiszewski et al., 2010; Ostrom & Ostrom, 1977; Romer, 1990). RDT suggests that declines in online participation can be explained in terms of competition over important rival resources (X. Wang et al., 2012).

On the other hand, online groups also rely on *nonrival* resources. They can even produce connective and communal public goods like opportunities to communicate or collections of information (Fulk et al., 1996) which can be "antirival" when their usefulness increases as a result of others using them (Kubiszewski et al., 2010; Weber, 2000). For example, the usefulness of a communication network increases as more people join it (Fulk et al., 1996; Katz & Shapiro, 1985). Similarly, the usefulness of an information good can increase as more people come to know, refer to, and depend upon it (Kubiszewski et al., 2010; Weber, 2000). If multiple online groups help build the same connective or communal public goods, they may form mutualistic interactions where contributions to one group may "spill over" and motivate participation in mutualist groups (Zhu, Kraut, et al., 2014). Ecological approaches seek to understand how

different types of resources will limit or promote growth.

Population Ecology, Density Dependence and Overlapping Resources

While this paper focuses on the ecological study of online groups, other social computing and HCI scholars have used the term "ecology" (and related concepts like "ecoystem" and "environment") to denote an assemblage of sites, devices, or platforms (Nardi & O'Day, 1999; Y. Wang et al., 2015). We use the term more narrowly to refer to conceptual and mathematical models of ecological dynamics. In particular, our work builds on a tradition rooted in *organizational ecology*. First developed in the late 1970s by sociologists studying interactions between firms, organizational ecology was inspired by, and has drawn closely from, ecological studies in biology (Hannan & Freeman, 1977).

Because online groups bear similarities to traditional organizations, organizational ecology provides a compelling theoretical framework for understanding interdependence among online groups. It has inspired at least three high-quality empirical studies of how resources shared by online groups shared shape their growth, decline, or survival (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). These studies draw from the *population ecology* strand of organizational ecology that studies ecological dynamics within a population of groups. In organizational ecology, populations have been defined as sets of organizations sharing an organizational industry or business model (Hannan & Freeman, 1989). In social computing, populations have been defined as online groups sharing a given social media platform (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014).

While population ecology involves several distinct theoretical propositions, *density dependence theory* (DDT) is perhaps the most prominent and is the subject of all three prior ecological studies of online groups (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). DDT models competitive or mutualistic forces in a population of groups as a function of *density* which, in the earliest and most influential studies of DDT, is simply the size of the population. In this way, DDT assumes that every group in the population is facing the same competitive and mutualistic pressures (Aldrich & Ruef, 2006). However, online groups sharing a platform have diverse topics (Kairam et al., 2012), norms (Chandrasekharan et al., 2018; Fiesler et al., 2018), and user bases (Tan & Lee, 2015). Because groups sharing few resources are unlikely to be strongly interdependent, ecological studies of online groups have modeled density dependence based on the concept of *overlap density* (Baum & Shipilov, 2006; Dobrev et al., 2001; X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). Rather than the number of groups that exist in a population, overlap density measures the extent to which an one group's members or topics overlap with all other groups'. Overlap density thus characterizes a group's *niche* or local *resource environment* defined by its distinctive topic and membership.

DDT proposes a model for the growth of organizational populations that has a similar structure to Butler (2001) RDT model for the growth of online groups. In DDT, mutualism is the engine of positive feedback driving population growth. Organizational ecologists show how successful organizations in an emerging industry develop nonrival resources like the legitimacy of a business model or industrial know-how that attract new organizations to enter the market (Carroll & Hannan, 1989; Hannan & Freeman, 1989). Similarly, a population of online groups, such as those sharing a platform, may grow in size as their platform gains in popularity, as estab-

lished groups spin off new ones, and as useful knowledge develops that can be shared between groups (Tan, 2018; Zhu, Kraut, et al., 2014).

In RDT, growth of online groups is self-limiting because of the challenges in managing large groups (Butler, 2001). In DDT, competition among population members over rival resources limits growth (Hannan & Freeman, 1989). DDT thus proposes a trade-off in which low density reflects limited opportunities for mutualistic contributions of nonrival resources like legitimacy, connectivity, and knowledge, but high density reflects competition over rival resources. Therefore, DDT predicts that the relationship between density and positive outcomes like growth or survival is ∩-shaped (inverse-U-shaped) (Baum & Shipilov, 2006; Carroll & Hannan, 1989).

Tests of DDT in populations of online groups yield inconsistent results. In X. Wang et al. (2012), user overlap in Usenet newsgroups is associated with decreasing numbers of participants. Similarly, TeBlunthuis et al. (2020) find that topical overlaps between online petitions are negatively associated with participation. By contrast, Zhu, Kraut, et al. (2014) find that membership overlap is positively associated with increasing survival of new Wikia wikis. Only Zhu, Chen, et al. (2014) find support for the ∩-shaped relationship predicted by DDT in an enterprise social media platform.

In Study A, we provide a test of DDT using data from Reddit. The classical logic of DDT appears reasonable in the context of Reddit because low overlap density is likely to reflect an impoverished environment lacking in non-rival resources like skills and knowledge of experienced users, while a group with high overlap is likely to face competition over its members (Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014): ***(H1)*** *The relationship between overlap density and the growth of online groups is ∩-shaped (inverse-U-shaped).*

DDT proposes that very high levels of density will decrease growth because of increasing forces of competition within a niche. However, to conclude that groups with the greatest membership overlap are likely competitors would be to commit a well-known statistical fallacy (Piantadosi et al., 1988; Robinson, 1950). The density of a group's environment suggests that it faces competition or mutualism, but it does not tell us which overlapping communities are competitors and which are mutualists. Community ecology overcomes this limitation of DDT.

Introducing Community Ecology

Perhaps the most natural way to understand the distinction between population ecology and community ecology is in where they believe ecological dynamics like competition and mutualism play out (Astley, 1985). While population ecology locates competition and mutualism within an environmental niche, community ecology locates competition and mutualism in networks of interdependent groups called *ecological communities* (Aldrich & Ruef, 2006). In organizational ecology, this can mean studying interactions between different organizational populations (e.g. J. M. McPherson, 1983; Sørensen, 2004), or networks of interactions between organizations (e.g Margolin et al., 2012; Powell et al., 2005). While varying conceptions of community ecology are found in the organizational ecology literature (J. H. Freeman & Audia, 2006), the approach we describe is identical in structure to that taken by Aldrich and Ruef (2006) and Hawley (1986).

Community ecology focuses on *ecological interactions* (Aldrich & Ruef, 2006). Ecological interactions can be mutualistic when one group has a positive influence on the second such that growth in the first group leads to growth in the second. They can also be competitive if one

group has a negative effect on the second such that growth in the first group leads to decline in the second. Ecological interactions can be reciprocated if mutualism (or competition) from one group to another group is returned in kind. An ecological interaction can also be mutualistic in one direction and competitive in the other. The competitive or mutualistic interactions in an ecological community are quantified by the *community matrix*, a central analytical object in community ecology in both biology and organization science (Aldrich & Ruef, 2006; Novak et al., 2016; Verhoef & Morin, 2010).

In Study B, we demonstrate community ecology by inferring networks of ecological interactions in ecological communities on Reddit. Because our understanding of community ecology theory does not suggest hypotheses about what we will find, we conduct an exploratory data analysis to determine whether mutualism or competition among subreddits is more common on Reddit and present case studies illustrating the types of ecological communities we identify.

Predicting Growth

In Study C we build upon our analyses from Study A and Study B by testing whether community ecology can explain the growth and decline of online groups in ways that population ecology can not. We do this by analyzing in two different ways whether accounting for ecological interactions helps predict future group sizes. In general, competition for overlapping resources will have no effect on group growth if something besides the overlapping resource limits growth (Verhoef & Morin, 2010). For example, two wikis might share a large number of contributors (they have high user overlap), but their growth might be limited by a lack of core contributors who perform important administrative tasks like policy making and software administration (Zhu, Kraut, et al., 2014). Community ecology relaxes the assumption that competition and mutualism are caused by user overlap density and instead seeks to infer these relationships from data. We test the importance of this conceptual shift for predicting growth by testing two hypotheses. The first uses a model comparison approach to test if adding a measure of ecological interactions to the density dependence model in Study A improves prediction of growth: (***H2)*** *A model with ecological interactions and density dependence predicts growth in online groups better than density dependence alone.*

Support for H2 may be a relatively low bar for assessing whether ecological interactions are important factors shaping the growth of online groups because of confounding moderator or mediator variables related to the occurrence of ecological interactions. Therefore, we also use a time series forecasting approach to test whether modeling ecological interactions is useful for making time series forecasts of participation in online groups: (***H3)*** *The addition of ecological interactions to a baseline time series model improves the forecasting performance.* While this does not directly compare population ecology and community ecology, it validates that ecological interactions are important.

4.3. Materials & Methods

Data

Our data are drawn from the publicly available Pushshift archive of Reddit submissions and comments which we obtained from December 5th 2005 to April 13th 2020 Baumgartner et al.

(2020). Within this dataset, we limit our analysis to submissions and comments from the 10,000 subreddits with the highest number of comments. There are 702 subreddits larger than the smallest subreddit included in our dataset having a majority of submissions marked "NSFW," which typically indicates pornographic material. As others have done in large-scale studies of Reddit (e.g., Datta et al., 2017), we exclude these subreddits to avoid asking members of our research team to inspect clusters including pornography. The top 10,000 subreddits provide a sufficiently large number of ecological communities for our statistical analysis.

Study A: Density Dependence Theory

User overlap $o_{i,j}$ quantifies the degree to which two subreddits (i and j) share users. Zhu, Kraut, et al. (2014) and X. Wang et al. (2012) both measure user overlap between two groups by counting the number of users contributing to both groups at least once and exclude users who appear in more than 10 groups. In our preliminary analysis, we found that this measure led to similarity measures and clusters with poor face validity. These issues may have stemmed from how Reddit users often peripherally participate in many groups while participating heavily in few (Hamilton et al., 2017; Tan & Lee, 2015; J. Zhang et al., 2017). Therefore, our measure of user overlap follows Datta et al. (2017) by using the number of comments each user makes in each pair of groups.

To measure user overlap between subreddits, we first build user frequency vectors by counting the number of times each user comments in each subreddit. We prevent giving undue weight to subreddits with higher overall activity levels by normalizing the comment counts for each subreddit by the maximum number of comments by a single author in the subreddit:

$$f_{u,j} = \frac{n_{\mathrm{u,j}}}{max_{v \in \mathrm{J}} n_{v,j}} \tag{4.1}$$

where $n_{u,j}$, the user frequency, is the number of times that user u authors a comment in subreddit j.

This results in a user frequency vector F_j for each subreddit that is sparse and high-dimensional, having one element for each user account that comments in any subreddit in our dataset. Next, we use LSA to reduce the dimensionality of the user frequency vectors. LSA is based on the singular value decomposition and is common in natural language processing and information retrieval. LSA preserves subreddit similarities while removing noise and dealing with sparsity (Dumais, 2004):

$$\begin{aligned} \mathbf{F} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ \widetilde{F}_j &= \mathbf{U_k}^T F_j \end{aligned} \tag{4.2}$$

$\mathbf{F}$ is the matrix where columns are author frequency vectors F_j and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is its singular value decomposition. Truncating the singular value decomposition to use only the first k left-singular vectors gives $\mathbf{U_k}$. Left-multiplying a subreddit's author frequency vector by $\mathbf{U_k}$ transforms the high-dimensional author frequencies into $\widetilde{F}_j$, their approximation in the k-dimensional space.

We then obtain our measure of *user overlap* by taking the cosine similarities between the resulting vectors for a pair of subreddits:

$$o_{i,j} = \frac{\widetilde{F}_j \cdot \widetilde{F}_i}{\|\widetilde{F}_i\|\|\widetilde{F}_j\|} \tag{4.3}$$

where $\|\widetilde{F}_i\| = \sqrt{\sum_{x=1}^{k} \widetilde{f_{x,i}}^2}$ is the euclidean norm of the transformed user frequencies for subreddit i.

Growth is the dependent variable in our density dependence model testing H1 and is also used in our test of H2 as part of Study B. Growth is measured as the change in the (log-transformed) size of a subreddit over the final 24 weeks of our data, from to November 4th 2019 to April 13th 2020.

Overlap density d_i is the normalized average user overlap for a given subreddit. It is the independent variable in our density dependence model testing H1:

$$\begin{aligned} d_i^* &= \frac{1}{|S| - 1} \sum_{j \in R; j \neq i} \mathrm{o}_{i,j} \\ d_i &= \frac{d_i^*}{\max_j d_j^*} \end{aligned} \tag{4.4}$$

where S is the set of groups in our dataset.

Regression model for H1 To test H1, we fit Model 1 which has first and second-order terms for overlap density to allow for a curvilinear relationship between *overlap density* and *growth*.

$$\text{Model 1} \qquad Y_i = B_0 + B_1 d_i + B_2 d_i^2 \tag{4.5}$$

where Y_i is the growth of subreddit i and d_i is its overlap density.

Study B: Introducing Community Ecology

Clustering to identify ecological communities Analyzing networks of ecological interactions is the key difference between community ecology and population ecology. To identify ecological communities of related subreddits, we use a clustering procedure based on the user overlap measure described above in §4.3. We selected a clustering model using grid search to obtain a high silhouette coefficient (Rousseeuw, 1987). The silhouette coefficient captures the degree to which a clustering creates groups of subreddits with high within-cluster similarity.

Our description of our measure for user overlap in §4.3 does not explain how we choose the number of LSA dimensions k. To do so, we ran the affinity propagation (B. J. Frey & Dueck, 2007), HDBSCAN (McInnes et al., 2017) and *k*-means clustering algorithms and selected the algorithm, hyperparameters, and LSA dimensions k that resulted in the clustering with a high silhouette coefficient having less than 5,000 isolated subreddits, and at least 50 clusters. We limit

the number of isolated subreddits because some choices of hyperparamters for the HDBSCAN algorithm could improve the silhouette coefficient, but at the cost of greatly increasing numbers of isolated subreddits. Choosing a relatively high limit to the number of isolates helps ensure that our clusters contain highly related communities. We chose an HDBSCAN clustering with 731 clusters, 4964 isolated subreddits, $k = 600$ LSI dimensions, and a silhouette score of 0.48. We exclude the isolated subreddits from our analysis. More details about our clustering selection process are found in the online supplement.

We evaluate the external validity of the chosen clustering using the purity evaluation criterion (Manning et al., 2018) To do so, an undergraduate research assistant examined a random sample of 100 clusters including 744 subreddits. By visiting the subreddits and using her own judgment, the assistant flagged subreddits that did not seem like a good fit for their assigned cluster. Using these labels and excluding 25 subreddits that have been deleted, made private, or banned, we calculated the purity of our clustering as 0.92. This means that we believe that 92% of subreddits belong to their assigned cluster.

Group size is the dependent variable of the models we use to infer ecological interactions. Measured as the number of distinct commenting users in a subreddit each week, group size quantifies the number of people who participate in a subreddit over time. Typical of social media participation data, group size is highly skewed. Therefore, we transform it by adding 1 and taking the natural logarithm.

Inferring ecological interactions using Vector Auto Regression The community matrix $\mathbf{\Phi}$ of ecological interactions can be inferred from time series data using vector autoregression models (VAR models). VAR models are a workhorse in biological ecology because VAR(1) models (i.e., VAR models with a single autoregressive term) have a close relationship with the Gompertz of population growth which is widely used in ecology (Ives et al., 2003). Even in the presence of unmodeled nonlinearities, VAR(1) models can reliably identify competition or mutualism in empirically realistic scenarios (Certain et al., 2018). VAR models also been widely adopted in the social sciences, particularly in political science and in macroeconomics (Box-Steffensmeier, 2014).

VAR(1) models can be intuitively understood as a generalization of auto-regressive AR(1) models in time series analysis. But while AR(1) models predict the state of a single time series as a function of its previous value, VAR(1) models simultaneously predict multiple time series as a function of the values of every other variable in the system (Canova, 2007; Ives et al., 2003):

$$Y_t = B_0 + B_1 t + \sum_{k \in K} A_k x_{k,t} + \sum_{j \in M} \Phi_j y_{j,t-1} + \epsilon_t \tag{4.6}$$

where Y_t is a vector containing the sizes of a set of online groups (M) at time t. B_0 is the vector of intercept terms and B_1 is the vector of linear time trends ($b_{1,j}$) for each community (j). Φ_j represents the influence of $y_{j,t-1}$, the size of the j^{th} online group at time $t-1$ on Y_t. Φ_j is a column of $\mathbf{\Phi}$, a matrix of coefficients in which the diagonal elements correspond to intrinsic growth rates (marginal to the trend) for each online group and the off-diagonal elements are intergroup influences, and ϵ_t is the vector of error terms

Additional time-dependent predictors ($x_{k,t}$) can be included in the vectors X_k with coefficients a_k. Because subreddits are created at different times, growth trends must begin only after the subreddit is created. We use X_k to introduce a counter-trend during the period prior to the creation of subreddits so that each group's growth trend begins in the period the group is created. For each group j created at time t_j^0 we fill X_j with the sequence $[1, 2, 3, \dots, t_j^0 - 1, 0, 0, 0, \dots]$. In other words, X_j adds a counter-trend only during the period prior to the first comment in subreddit j. We fix the elements $a_{j,i}$ of A_j equal to `0` unless $i = j$, so the counter trend only influences subreddit j. This effectively sets $a_{j,j}$ approximately equal to $-b_{1,j}$.

We fit VAR(1) models using ordinary least squares as implemented in the `vars` `R` package to predict the group size each week using over the history of each subreddit prior to November 4$^{\text{th}}$ 2019 (Pfaff, 2008). We hold out 24 weeks of data for forecast evaluation and fit our models on the remainder. To ensure that sufficient data is available for fitting the models, we exclude 946 subreddits and 89 clusters having less than 156 weeks of activity.

Characterizing ecological communities In Study B, we interpret the community matrix $\mathbf{\Phi}$ as a directed network of ecological interactions, a *competition-mutualism network* (Ives et al., 2003). Although the elements of $\mathbf{\Phi}$ correspond to direct associations between group sizes (Novak et al., 2016), ecological interactions can also be indirect. Consider 3 one-directional interactions between three groups (a, b, c) such that growth in a predicts decreased growth in b ($\phi_{a,b} < 0$), growth in b predicts decreased growth in c ($\phi_{b,c} < 0$), but a and c do not directly interact ($\phi_{a,c} \approx 0$).

This does not necessarily mean that groups A and C are independent. Rather, an exogenous increase in A predicts a decrease in B and thereby an eventual increase in C. Such indirect relationships are analyzed by using impulse response functions (IRFs) to interpret a VAR model (Box-Steffensmeier, 2014). In large VAR models containing many groups, the great number of parameters can mean that few specific elements of $\mathbf{\Phi}$ will be statistically significant, even as many weak direct relationships can combine into statistically significant IRFs (Canova, 2007).

Average ecological interaction $\overline{m}$ measures the extent to which an overall ecological community is mutualistic or competitive by taking the mean point estimate of the off-diagonal coefficients of $\mathbf{\Phi}$:

$$\overline{m} = \frac{1}{|M| - 1} \sum_{i \in M} \sum_{j \in M; j \neq i} \phi_{i,j} \tag{4.7}$$

if $\overline{m} > 0$ then mutualistic interactions within the ecological community are stronger than competitive ones, and if $\overline{m} < 0$ then competitive interactions are stronger then mutualistic ones.

Ecological interaction strength κ quantifies the overall strength of ecological interactions in an ecological community as the mean absolute value of the point estimates of the off-diagonal coefficients of $\mathbf{\Phi}$:

$$\kappa = \frac{1}{|M| - 1} \sum_{i \in M} \sum_{j \in M; j \neq i} \left| \phi_{i,j} \right| \tag{4.8}$$

where $\left|\phi_{i,j}\right|$ is the absolute value of the coefficient $\phi_{i,j}$.

Ecological communities of subreddits with overlapping users vary in both the overall strength of ecological interactions and in the overall degree of mutualism and competition between member groups. If an ecological community's average ecological interaction is positive, we say the ecological community is mutualistic. If it is negative, we say the ecological community is competitive. The average ecological interaction can be close to 0 in two ways. First, the ecological interaction strength can simply be low. Alternatively, the ecological community can have a mixture of competitive and mutualistic interactions that cancel one another out when averaged.

Impulse response functions (IRFs) of our VAR(1) models correspond to our visualizations of example competition-mutualism networks in §4.4. An IRF predicts how much each group's size would change in response to a sudden increase in the size of each other group (Verhoef & Morin, 2010):

$$\mathbf{\Theta_t} = \mathbf{\Theta_{t-1}}\mathbf{\Phi}, t = 1, 2, ... \tag{4.9}$$

where $\mathbf{\Theta_t}$ is the impulse response function at time t. $\mathbf{\Theta_0}$ is an M-by-M identity matrix so our impulses represent a log-unit increase of 1 to each group. $\mathbf{\Theta_t}$ is a matrix with elements $\theta^t_{i,j}$ corresponding to the response of group j to the impulse of group i. We draw an edge $i \rightarrow j$ in the competition-mutualism network if the 95% CI of $\theta^t_{i,j}$ does not include zero at any time $10 >= t > 0$. If $\theta^t_{i,j} > 0$, the edge indicates mutualism and if $\theta^t_{i,j} < 0$ the edge indicates competition.[2] We compute the IRFs with bootstrapped confidence intervals (CI) based on 1,000 samples using the `vars` `R` package.

Study C: Predicting growth

Average subreddit mutualism m_j is the independent variable for our test of H2 and measures the average influence of other subreddits in the ecological community on a given subreddit j, which we calculate by taking the mean of off-diagonal elements of row j of the community matrix:

$$m_j = \frac{1}{|M| - 1} \sum_{i \in M; i \neq j} \phi_{i,j} \tag{4.10}$$

where M is the set of subreddits in the ecological community and $|M|$ is the number of subreddits in M. We use the mean instead of the sum because different ecological communities have different numbers of subreddits.

Regression models for H2 We test H2 by using likelihood ratio tests to compare Model 1 and Model 2 which adds *average subreddit mutualism* (m_i) as a predictor. We also fit Model 3 which we compare to Model 2 to test if overlap density explains variation that average subreddit mutualism does not.

[2] In higher-order VAR(p) models that use $p > 1$ past observations as predictors $\theta^t_{i,j}$ can be less than 0 for some t_a and greater than 0 for some t_b. However, this is not possible in the VAR(1) models we use.

$$\text{Model 2} \qquad Y_i = B_0 + B_1 d_i + B_2 d_i^2 + B_3 m_i \tag{4.11}$$

$$\text{Model 3} \qquad Y_i = B_0 + B_3 m_i \tag{4.12}$$

where Y_i is the growth of subreddit i, d_i is its overlap density, m_i is its average subreddit mutualism, and B_0, B_1, B_2, and B_3 are regression coefficients.

Forecasting growth using ecological interactions To test H3, we evaluate whether modeling ecological interactions improves time series forecasting of future participation in online groups by comparing the model in Equation 4.6 to a baseline model with off-diagonal elements of $\mathbf{\Phi}$ fixed to 0. This baseline model is equivalent to our VAR model, but excludes ecological interactions.

We use two forecasting metrics with differing assumptions: root-mean-square-error (RMSE) and the continuous ranked probability score (CRPS). RMSE is commonly used, non-parametric, and intuitive, but does not take differing scales of the predicted variable or forecast uncertainty into account. Thus, in our setting it may place excessive weight on the forecasts of larger subreddits where errors may have greater magnitude simply because the absolute magnitude of the variance is greater. By rewarding forecasts where the true value has high probability under the predictive distribution, the CRPS accounts for variance in the data and rewards forecasts for both accuracy and precision and is thus a "proper scoring rule" for evaluating probabilistic forecasts (Gneiting & Raftery, 2007). Our CRPS calculations assume that the predictive forecast distribution for each community is normal with standard deviations given by the 68.2% forecast confidence interval. We calculate CRPS using the `scoringRules` `R` package (Jordan et al., 2019).

4.4. Results

Study A: Density Dependence Theory

We test the classical prediction of density dependence theory as formulated in H1 using Model 1 which has first- and second-order terms for the effect of overlap density on growth. As described in §4.2, H1 hypothesizes that overlap density will have a curvilinear ∩-shaped (inverse-U-shaped) relationship with growth indicated by a positive first-order regression coefficient and a negative second-order coefficient.

As predicted, we observe a ∩-shaped relationship between overlap density and growth. Figure 4.1 plots the marginal effects of overlap density on growth for the median subreddit laid over the data on which the model is fit. Table 4.1 shows regression coefficients for Models 1-3. For about half of subreddits, increasing overlap density is associated with higher growth rates. The point where increasing density ceases to predict increasing growth and begins to predict decreasing growth is at the 49$^{\text{th}}$ percentile. Prototypical subreddits at this overlap density grew slightly (95% CI:[0.001,0.06]). Yet subreddits at the lower and upper extremes of overlap density slightly declined on average. Typical groups at the 20$^{\text{th}}$ percentile of overlap density decline by 1.1 members (95% CI:[-1.1,-1.15]) and typical groups at the 80$^{\text{th}}$ percentile decline by 1.2 members (95% CI:[-1.1,-1.28]). While we find support for the classical theoretical prediction of

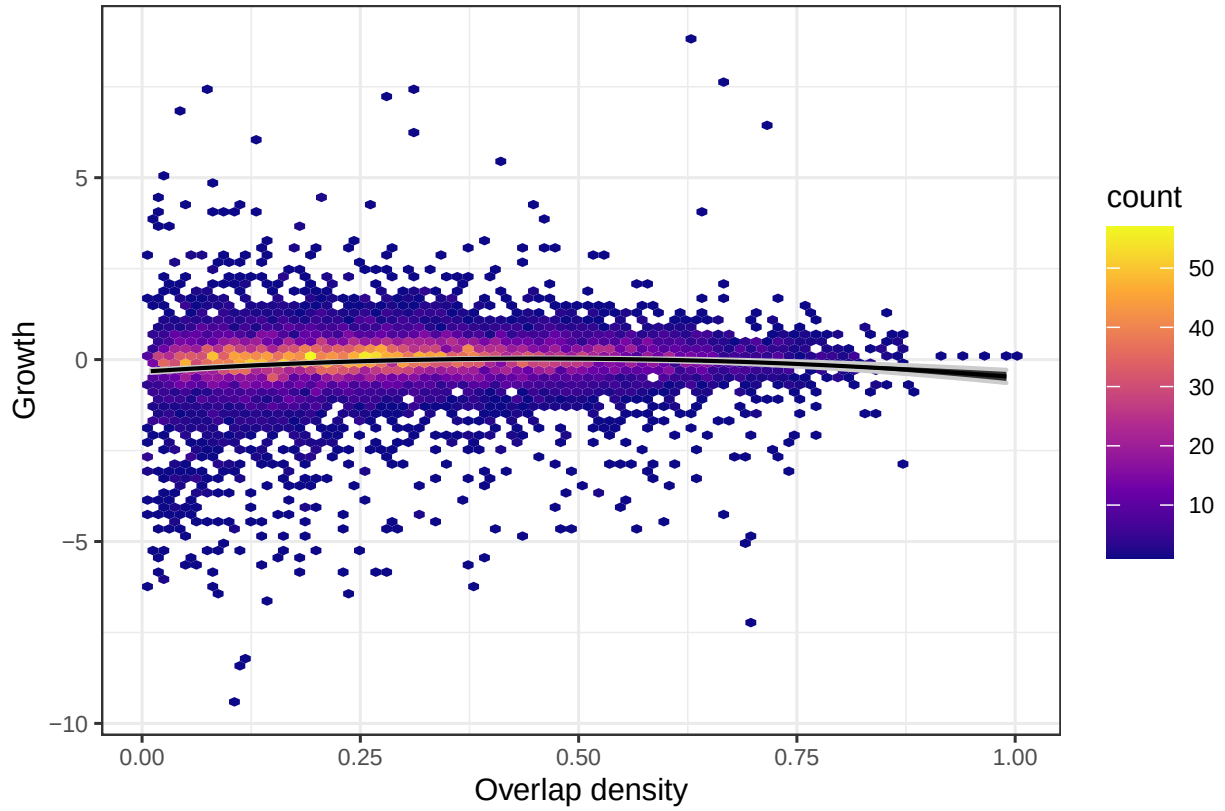


Figure 4.1: Relationship between density and growth. A 2D histogram of subreddits with overlap density (log-transformed) on the X-axis and the change in the logarithm of the number of distinct commenting users on the Y-axis. The black line shows the marginal effect of overlap density on growth as predicted by Model 2. The gray region shows the 95% confidence interval of the marginal effect.

	Model 1	Model 2	Model 3
Overlap density	1.50* (0.26)	1.50* (0.26)	
Overlap density^2	−2.08* (0.41)	−2.09* (0.41)	
Average subreddit commensalism		0.12 (0.26)	0.11 (0.26)
Constant	−0.23* (0.03)	−0.23* (0.04)	−0.04* (0.01)
Log Likelihood	-4970	-4970	-4986
Observations	4,090	4,090	4,090
Note:			*p< 0.01

Table 4.1: Loglinear regression predicting subreddit growth as a function of overlap density. The model supports the prediction of density dependence theory of a ∩-shaped relationship between overlap density and growth.

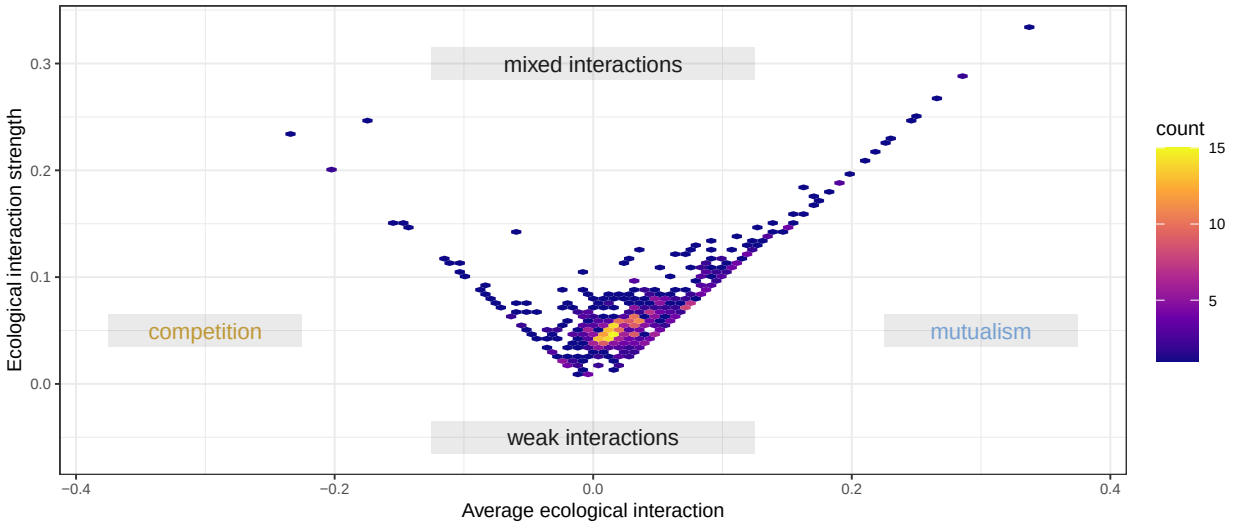


Figure 4.2: Two-dimensional histogram showing ecological communities on Reddit in our typology. The X-axis shows the overall degree of mutualism or competition in clusters of subreddits with high user overlap based on the average ecological interaction. The Y-axis shows the ecological interaction strength representing the overall magnitude of competition or mutualism.

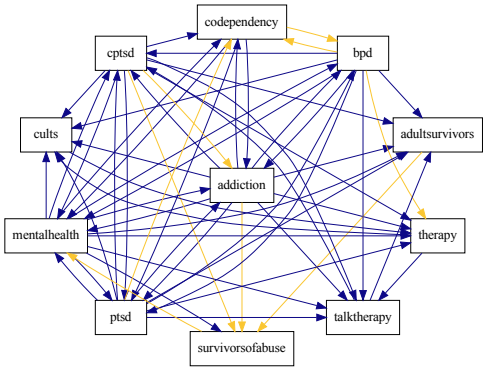
a curvilinear, (∩-shaped) relationship between overlap density and growth, this does not imply that relationships between highly overlapping communities are more competitive.

Study B: Introducing Community Ecology

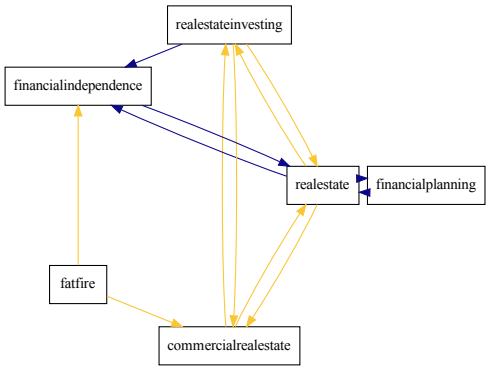
Figure 4.2 visualizes the distribution of average ecological interaction and ecological interaction strength over the 641 ecological communities we identify. We observe ecological communities characterized by strong forms of both mutualism and competition, others having mixtures of the two, and some with few significant ecological interactions. Mutualism is more common than competition, with the mean community having an average ecological interaction of 0.03 ($t = 14.5$, $p < 0.001$). We find that 524 clusters (81.7%) are mutualistic. Not only are most ecological communities mutualistic, but more mutualistic ecological communities have greater ecological interaction strength (Spearman's $\rho = 0.58$, $p < 0.001$). Therefore, our community ecology analysis suggests that among groups with similar users, mutualistic ecological interactions are more common than competitive ones.

Example ecological communities We present four case studies to illustrate our typology of ecological communities of online groups. Figure 4.2 shows that we find clusters of subreddits characterized by mutualism, competition, a mixture of mutualism and competition, and few ecological relationships at all. We select one case from each of these four types using our measures of average ecological interaction (§4.3) and ecological interaction strength (§4.3). To allow for more interesting network structures, we draw our cases from the 367 large clusters having at least five subreddits.

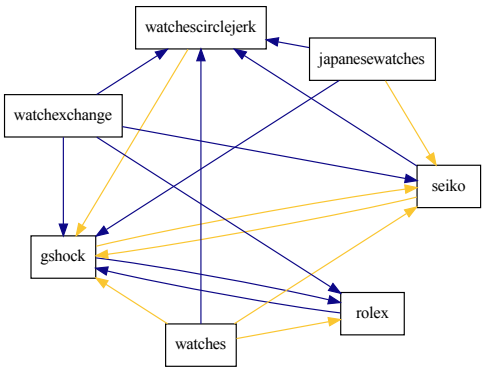
Figure 4.3, presents visualizations of competition-mutualism networks representing statistically significant impulse response functions as described in §4.3. During our analysis, we also



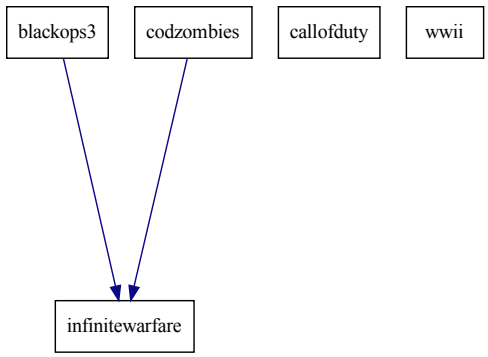
(a) The ecological community of subreddits for supporting mental health and survivors of abuse is dense with largely mutualistic interactions. Some interactions, like that between `r/mentalhealth` and `r/survivorsofabuse` are mutualistic in one direction but competitive in the other.



(b) The subreddits about real estate and finance are relatively competitive. We detect reciprocal competitive relationships among the real estate subreddits in the triad including `r/realestateinvesting`, `r/realestate` and `r/commercialrealestate`.



(c) Subreddits about watches are dense with both mutualistic and competitive interactions. There is a reciprocal competitive interaction between `r/gshock` and `r/seiko`, a reciprocal mutualistic interaction between `r/gshock` and `r/rolex` well as several unreciprocated mutualistic and competitive interactions.



(d) The ecological community of subreddits about Call of Duty video games is characterized by relatively sparse ecological interactions. We detect only two mutualistic interactions from `r/blackops3` to `r/infinitewarfare` from codzombies to `r/infinitewarfare`.

Figure 4.3: Network visualizations of commensal relationships in example ecological communities of subreddits with overlapping users. Yellow indicates competition and purple indicates mutualism.

examined the terms of the vector autoregression parameter Φ, the impulse response functions, and model fits and forecasts, all of which are available in our online supplement. We also visited each subreddit in the clusters and read their sidebars and top posts to support our brief qualitative descriptions.

Mutualism among mental health subreddits To find a case characterized by mutualism, we selected the top 37 large clusters with the greatest average ecological interaction. From these, we arbitrarily chose one interesting ecological community, the *mental health* cluster, which includes 11 subreddits for supporting people in struggles with mental health, addiction, and surviving abuse. Constitutive subreddits include those focused on specific mental health diagnoses like `r/bpd` (bipolar disorder) and `r/cptsd` (complex post traumatic stress disorder) while others like `r/survivorsofabuse` and `r/adultsurvivors` are support groups.

The interactions among these subreddits are dense and primarily mutualistic as shown in Figure 4.3a. There are a handful of competitive interactions like the reciprocal competition detected between `r/codedependence` and `r/bpd`. We also observe some interactions that are mutualistic in one direction and competitive in the other. For example, growth in `r/addiction` predicts increasing growth in `r/cptsd` even as that growth in `r/cptsd` predicts decreasing growth in `r/addiction`. This suggests a pattern in which `r/cptsd` siphons members from `r/addiction`. That said, the density of mutualistic interactions shown in Figure 4.3a suggests that different subreddits have complementary roles in this ecological community as people turn to different types of groups for help with interrelated problems. While attempting to explain why different online groups form mutualistic or competitive interactions is left to future research, the example of mental health subreddits shows how groups with related topics and overlapping participants can have mutualistic interactions where growth in one predicts growth in many of the rest.

Competition among real estate and finance subreddits To find competitive clusters, we selected from the 36 large clusters with the lowest average ecological interaction an ecological community that we label *finance*. Among the 6 subreddits in this cluster, `r/realestateinvesting`, `r/realestate` and `r/commercialrealestate` all deal in different aspects of the real estate industry, while `r/financialindependence` and `r/fatfire` (the acronym "fire" means "financial independence/retire early") are focused on building wealth and becoming financially independent and `r/financialplanning` is a general purpose subreddit for financial advice.

In contrast to the mental health ecological community, the finance cluster has mostly competitive ties as visualized in Figure 4.3b. The fact that even this cluster, among the most competitive in our data, contains a number of mutualistic ties reflects just how prevalent mutualism is among subreddits with high degrees of user overlap. That said, we detect three reciprocal competitive interactions among the three subreddits that focus on real estate. The edges from `r/fatfire` to `r/commercialrealestate` and `r/financialindependence` are competitive as well. Interestingly, all interactions between the general finance subreddits (`r/financialplanning` and `r/financialindependence`) and `r/realestate` are mutualistic.

Mixed interactions among timepiece subreddits Next, we turn to an example of an ecological community with low average ecological interaction but high ecological interaction strength. We

first select the 36 large clusters with the average ecological interaction closest to 0. To find an ecological community with a mixture of mutualism and competition, we select from the 15 clusters with the greatest ecological interaction strength from within this group and chose the *timepiece* cluster containing 7 subreddits about watches.

As shown in Figure 4.3c, the ecological community of timepiece subreddits is dense with ecological interactions (although not as dense as the mental health subreddits). We observe both reciprocated mutualistic interactions, like that between r/rolex and r/gshock, and competitive interactions like that between r/gshock and r/seiko. We also observe numerous unreciprocated competitive and mutualistic relationships like the mutualism between r/watchexchange and r/watchcirclejerk[3] and the competition between r/japanesewatches and r/seiko. Though the average ecological interaction among these subreddits is near 0, our analysis reveals a complex ecological community with a mixture of competition and mutualism.

Sparse interactions among Call of Duty subreddits To find a case where ecological interactions are weak, we return to the group of the 36 large clusters with the average ecological interaction closest to 0 but select from the 15 clusters within this group with the lowest ecological interaction strength. From these, we chose the *Call of Duty* cluster containing five groups about the popular military first-person shooter series of video games.

The Call of Duty ecological community is sparse, having only two significant ecological interactions among its 5 member groups. This ecological community includes subreddits about different editions of the series such as r/blackops3, r/infinitewarfar and r/wwii as well as one about a popular spin-off zombie game r/codzombies and the more general r/callofduty subreddit. We find that growth in r/blackops3 or r/codzombies predicts growth in r/infinitewarfare and no other ecological interactions.

The timepiece and Call of Duty ecological communities illustrate how subreddits with overlapping users can have relatively strong or weak forms of ecological interdependence. Although both clusters are characterized by high degrees of user overlap and low average ecological interaction, the timepiece cluster has a dense competition-mutualism network while the call of duty network is sparse.

Study C: Predicting Growth

We now compare the environmental approach of population ecology with the relational approach of community ecology. In Study B, we presented examples of diverse ecological communities among subreddits with overlapping members. However, the presence of this diversity this does not mean that ecological interactions are related to the growth of online groups, the key outcome of previous ecological studies. We therefore hypothesized that ecological interactions will improve the predictive performance of a density dependence model in H2.

Ecological interactions do not improve growth prediction To test H2, we compare Model 1, our density dependence model having first- and second-order terms for overlap density, with Model 2, which also includes average subreddit mutualism (§4.3) as a predictor. We also examine

[3]The suffix is widely understood on Reddit to signify a jokey, meme, or satirical subreddit.

Model 3, in which the only predictor is average subreddit mutualism. Table 4.1 shows regression coefficients for our models.

We do not observe a statistically significant association between average subreddit mutualism and growth ($B_3 = 0.12, SE = 0.26$). Moreover, a likelihood ratio test comparing Model 1 and Model 2 does not support H2 as Model 2 does not predict subreddit growth better than Model 1 ($\chi^2 = 0.23$, $p > 0.05$). Comparing Model 2 to Model 3 shows that overlap density explains variation that average subreddit mutualism does not ($\chi^2 = 33$, $p < 0.001$). Overlap density helps explain a group's future growth, but the overall degree of mutualism or competition a group faces in its ecological community does not.

Forecasting accuracy The likelihood ratio tests in §4.4 are limited because improvements in predictive performance (or lack thereof) may be due to unobserved factors predictive of growth that are correlated with average subreddit mutualism. We hypothesized in H3 that the inter-group dependencies in our VAR models can better forecast the size of subreddits compared to baseline time series models that do not account for ecological interactions. As described in §4.3, we test H3 by comparing two forecasting metrics: the root-mean-square-error (RMSE) and the continuous ranked probability score (CRPS).

VAR models including ecological interactions have forecasting performance superior to the baseline model in terms of both RMSE and CRPS. We evaluate the 24-week forecast performance for all subreddits which were assigned to clusters. The RMSE under the baseline model (0.84) is greater than the RMSE of the VAR models (0.75) and the CRPS of the baseline model (72,853) is also greater than the CRPS of the VAR models (72,669). This reflects a substantive improvement in forecast accuracy robust to the choice of the forecasting metric.

Our baseline model contains a constant term and a trend term for each group and therefore accounts for all time-invariant within-group variation. Because overlap density is a subreddit-level variable that does not vary over time, we know that the improvement in forecasting performance comes from modeling ecological interactions in ways not captured by overlap density.

4.5. Threats to Validity

Our work is subject to several important threats to validity that we cannot fully address. First, we study ecological communities on only one platform hosting online groups and our results may not generalize to other platforms or time periods. Additionally, while our community ecology approach assumes that ecological interactions drive dynamics in the size of groups over time and cause groups to grow or decline, drawing causal inference using our method would depend on several untestable assumptions. For example, our ability to infer causal relationships might be limited if groups we do not consider—including groups on other platforms—play a role in an ecological community. Regression estimates in Models 1-3 may be confounded by omitted variables and cannot support causal interpretation. Therefore, we refrain from claiming that the relationships we infer are causal.

The method we propose for identifying ecological interactions between online groups has limitations common to all time series analysis of observational data. Potential omitted variables might also include additional time lags of group size. Although we chose to use VAR(1) models with only 1 time lag, we hope future work can improve upon our approach and model more

complex dynamics with additional lags. Like most other time series analysis, vector autoregression assumes that the error terms are stationary. This is difficult to evaluate empirically and may not be realistic (Canova, 2007). Future work might relax these assumptions using more complex models with time-varying parameters, state space models (Box-Steffensmeier, 2014), nonlinear time series models (Cenci et al., 2019; Kantz & Schreiber, 2003), or stationarity-enforcing priors (Heaps, 2020). Such approaches may require additional contextual knowledge and be difficult to scale to an analysis of hundreds of different ecological communities, but may prove fruitful in future work focusing on ecological communities of interest. Such models may also be useful in future work investigating how ecological interactions change over time.

Additional threats to validity stem from our use of algorithmic clustering to identify ecological communities. Organizational ecologists have rarely attempted to estimate the full community matrix for an entire population containing a large number of groups because of data and statistical limitations (e.g. Ruef, 2000; Sørensen, 2004). For instance, 100 million possible ecological interactions exist within a set of 10,000 communities. Attempting to infer them all raises considerable computational and statistical challenges. We chose to use a clustering analysis to explore the typical ecological communities on a platform.

While we choose clusters based on high degrees of user overlap and validate our clustering in terms of the silhouette coefficient and purity criteria, we might have obtained different results if we had clustered in a different way. Additionally, our efforts to obtain clusters with a high silhouette coefficient lead us to remove a large number of subreddits from our analysis. Thus, our results are not representative of Reddit overall, but only of those subreddits that were included in our analysis. Furthermore, clustering algorithms like the one we use may not have unique solutions and different initial conditions and hyperparameters might lead to different results. While these allow us to scale up our analysis, future work should use principled definitions of an ecological community based on qualitative contextual knowledge in focused studies of particular ecological communities.

4.6. Discussion

To introduce community ecology and compare it to population ecology, we presented three studies. In Study A, we found support for H1 showing—as predicted by density dependence theory—that overlap density has an ∩-shaped association with subreddit growth. Subreddits with moderate overlap density in our data declined less than subreddits with either very low or very high overlap density. According to population ecology theory, this suggests that high-density environments are competitive and less conducive to growth than medium-density environments.

Surprisingly, this contrasts with our results in Study B, where we studied the diversity of ecological communities using vector autoregression models of group size over time to infer networks of ecological interactions. We find ecological communities that are mutualistic or competitive, that mix the two, or that have few significant ecological interactions at all. Overall, however, ecological communities of subreddits are typically mutualistic and mutualistic interactions are stronger on average than competitive ones. Although we find evidence of density dependence, density-dependent competition does not necessarily reflect typical relationships in ecological communities of highly overlapping subreddits.

Our results in Study C show that the size of the other members of an ecological community improves time series forecasts of participation in online groups. However, average subreddit mutualism did not help predict growth. This suggests that population ecology and community ecology offer complementary environmental and relational perspectives. Population ecology's focus on environmental factors such as niche and overlap density is useful for predicting growth, but does not provide a way to study networks of mutualism and competition. Community ecology unpacks density and provides insights about the specific relationships between groups. While modeling these interactions helps forecast participation levels in groups, the existence of these interactions may be independent of future growth. For example, if mutualistic relationships are common in declining ecological communities, that would explain our result for H2.

The complementary nature of the two ecologies is seen in the coincidence of our findings in Study A and Study B. Indeed, these results can help explain the puzzling set of empirical results about the relationship between overlap density and outcomes like growth, decline and survival (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014). Studies of density dependence theory in social computing measure the density of an online group's niche in terms of its overlap in participants or topics. Our analysis clearly shows that resource overlaps between two groups might have little to do with whether they are mutualists or competitors. Instead, overlaps may simply reflect the hospitality of the environment to groups with overlapping topics or user bases. As a result, the differing environmental conditions of Wikis and Usenet groups might explain why user overlap was associated with the survival of wikis (Zhu, Kraut, et al., 2014) but with the decline of Usenet groups (X. Wang et al., 2012). Wikia was a young and growing platform during Zhu, Kraut, et al.'s (Zhu, Kraut, et al., 2014) data collection period when the growth of groups may have been limited by knowledge of how to build a wiki, and this knowledge was provided by overlapping experienced users. Usenet was in decline during X. Wang et al.'s (X. Wang et al., 2012) study period and this may have produced competitive environmental conditions as users became more scarce.

The widespread mutualism found in Study B resonates with long-held understandings of ecological interactions in evolutionary theory (Kropotkin, 1902/2012). Competition is unlikely to persist because it decreases survival. Because mutualism increases survival, it will be favored by natural selection (Armstrong & McGehee, 1980; Axelrod & Hamilton, 1981). Similarly, competition can be avoided if groups adopt specialized roles in their ecological community, a dynamic known as resource partitioning in organizational ecology (Carroll, 1985; Menge, 1972; Schoener, 1974). Resource partitioning theory suggests that the competition among real estate subreddits observed in Figure 4.3b may be due to a lack of specialization. If specialization does not emerge over time, such groups of competing subreddits may have decreased survival. By contrast, mental health support groups like those observed in Figure 4.3b appear to have distinctive purposes or roles. Future work to test such mechanisms in ecological communities of online groups may reveal ways that online groups complement or cooperate with each other.

Within large platforms for online groups, the great number of ecological communities that can be studied should make it possible for future work to apply methods from network science to construct and test generalizable theories about the roles of different types of resources, design features of platforms, and governance institutions in these ecological interactions. Future work should also incorporate community ecology analysis in case studies of important topics such ecological communities engaged in peer production, political mobilization, misinformation, or

mental health support.

Although we focused on online groups within a single platform, groups may use multiple platforms with distinctive affordances for different purposes (Fiesler & Dym, 2020; Kiene et al., 2019). Since the VAR method relies only on time series data to infer ecological interactions, it can be applied to study ecological communities spanning social media platforms. Community ecology can thus provide a bridge between quantitative studies of participation in online groups and theories of interconnected information ecologies (Nardi & O'Day, 1999). While we focus on relationships between groups sharing a platform, one can apply our concepts and methods to understand how interdependent systems of technologies and users give rise to higher levels of social organization on social media platforms (Aldrich & Ruef, 2006; Astley, 1985).

Implications for Design

In the final chapter of their book on *Building Successful Online Communities*, Kraut et al. (2012) advise managers of online groups to select an effective niche and beware of competition. However, these recommendations are based on little direct evidence from studies of online groups and offer almost no concrete steps that designer or group should take based on either piece of advice. Although further research into ecological interactions is needed before design principles can be derived, we provide a framework for online group managers to think about ecological constraints on group size. While intuition suggests that online group managers might seek out mutualistic relationships and avoid competitive ones, it is often not obvious whether another group with overlapping users is a competitor or mutualist. Our method provides a way for group managers to know.

Competitors have a negative impact on growth, but ecological theory suggests that specialization is an adaptive strategy in response to competition (Aldrich & Ruef, 2006; Carroll, 1985; Kraut et al., 2012; Powell et al., 2005). Using our method, group managers might identify competitors limiting the growth of their groups. With the knowledge of this analysis in hand, they might be able to escape a competitive dynamic by specializing. While competitive relationships are defined by how they decrease the size of groups, competition can also be important to the health of the broader ecological community. Exit to an alternative group can be an avenue for political change in response to grievances and poor governance (S. Frey & Sumner, 2019; Hirschman, 1970). The threat of competition with other groups may make expressions of voice more persuasive to moderators or platforms (Hirschman, 1970).

Groups looking to increase activity should desire to seek out mutualistic relationships, and we believe that designers of online platforms can help them do so. Features such as meta-groups, group search, recommendation engines, and practices like linking related groups may lower barriers between groups and support mutualism. However, it is not obvious to what extent particular features will support competition, mutualism, or both. Using our method, managers and designers can test features intended to support mutualism.

4.7. Conclusion

While explanations for the rise or decline of online groups often look to internal mechanisms, understanding the role of interdependence between online groups is increasingly important. While prior research has investigated competition and mutualism among online groups with

overlapping users and topics using the population ecology framework (X. Wang et al., 2012; Zhu, Chen, et al., 2014; Zhu, Kraut, et al., 2014), this approach does not provide a way to infer competitive or mutualistic interactions among related groups. We introduce the community ecology framework as a complementary perspective to population ecology. By inferring competition-mutualism networks directly from time-series data, our community ecology approach helps resolve the empirical tensions raised by prior ecological work in social computing and reveal that most interactions within clusters of subreddits with highly overlapping users are mutualistic. Our methods provide a foundation for future work investigating related online groups.

Chapter 5

Future Directions in the Ecology of Online Communities

Chapter 1 says that "the project of this dissertation is to begin reconstructing organizational ecology in the relatively theory-poor but data-rich context of online communities." By focusing on understanding the relationships between related online communities in ecological terms of competition and mutualism and in the emic language of members of overlapping communities, the preceding work seeks to build an empirical foundation to build new ecological theory. It has found qualitative and quantitative evidence that overlapping online communities often fill distinctive niches by providing complementary benefits to their users. Competitive dynamics also occur, and can be strong, but do not last as long. Although competition and mutualism play a role in their growth and survival, communities may not adapt to promote mutualism and avoid competition. Rather it seems likely that the "principle of competitive exclusion" takes hold in some other way, perhaps through a selection process in which communities normally must provide complementary benefits to existing ones in order to take off.

Of course, these claims are limited by the empirical tools that were used to support them. Inferences about competition and mutualism are based upon time series models with fundamentally untestable assumptions. By fitting a far greater number of models than I could carefully specify I have taken an unabashed "big data" approach. To make confident claims about any particular competitive or mutualistic relationship between two subreddits I would have to conduct a relatively exacting model selection and comparison procedure based on additional contextual knowledge of the communities' histories. The large scale of this analysis supports the general findings enumerated above assuming that any model misspecifications have not introduced errors in a systematic and misleading way. The fact that both the linear and nonlinear time series analyses and the active community members all seem to agree that mutualism is more common than competition provides some reassurance. It seems quite unlikely that all three will mislead in similar ways.

The reconstruction project is still beginning, but at this stage we can propose preliminary answers to some key theoretical questions: (1) How do people construct systems of overlapping online communities? (2) What types of "resources" are most important for mediating ecological interactions? (3) How do ecological interactions relate to broader dynamics such as the growth of a platform or the popularity of a broader topic? and (4) How do barriers between different platforms affect cross-platform ecological relationships?

Question (1) is fundamental to an ecological explanation for the development of online communities. A preliminary answer is that *people construct systems of overlapping online communities as new online communities find distinctive niches in the neighborhood of existing communities relatively early in their development*. Chapter 4 finds evidence that systems of overlapping online communities are not constructed through an adaptation process and suggests a selection process as an alternative. It should be noted that selection and adaptation are not mutually exclusive and the systems of overlapping communities may develop through a hybrid process. Chapter 3 suggests that a large majority of active online communities each have a distinctive ecological niche. It seems likely that successful online communities are often quickly find a niche early in their development.

Prior research and Chapter 2 both see users and topics as related to rival or non-rival resources that make competition and mutualism between online communities possible. However, Chapter 2 finds that user and topic overlap densities are very weakly correlated with online community growth suggesting that user and content overlaps are not very close analogs for the kinds of resource overlaps considered by organizational ecologists, such as the technological range of a firm's outputs (Dobrev et al., 2003). Based on findings from Chapter 3, a preliminary answer to question (2) is that *online communities' ecological niches are a product of content categories, audiences, and social capital*. These dimensions of an online community's niche might be difficult to precisely measure, but they can be described in theory.

Content categories are socially constructed classes such as "memes," "Q&A," "news," "commentary," "art," "documentation" and "discussion." Online communities often specialize in a subset of possible content categories. Specialization in a set of content categories might be achieved formally through rules or definitions of topical scope or informally, through the community's size, or the preferences and behaviors of its members. The empirical work so far considers topics measured through semantic similarity or language models. Content categories are likely to be correlated with such measures, but the measures are unlikely to faithfully capture important aspects of content categories like differences in medium, genre and form.

The notion of social capital and audience as distinct aspects of a niche disentangles the concept of "user." Social capital refers to the benefits that come from interpersonal interaction and sense-making with a homophilous or tight-knit community (Ackerman et al., 2013). Measures of group size or user overlap may be correlated with social capital, but they do little to distinguish a user who comments as a member crowd-like audience from a user who seeks social bonds and interactions with fellow members of their identity group or enthusiasts in their hobby.

Question (3) is an important part of an ecological explanation for the rise and decline of platforms in terms of the communities they host. A preliminary answer is that *ecological interactions and the rise or decline of a topical area drive one another in feedback process*. Chapter 2 suggests that growing platforms may be more likely to have mutualistic dynamics as they have an increasing number of potential niches for online communities of varying sizes and scopes. At the same time, mutualistic interactions among overlapping communities are likely to drive the rise of a platform as mutualists enrich niches in their neighborhoods. In a similar way, competition and the decline of topical area might reinforce each other if out-migration of users interested in the topic induces competition over the remaining users and this accelerates the communities' declines prompting further out-migration.

Question (4) considers the ecological consequences of how different social media platforms divide related online communities such as the Wikis and subreddits about the same topic. A pre-

liminary answer is that *barriers between platforms limit both mutualistic and competitive dynamics* because of how they limit the sharing of users or content across platforms. However, when non-rival resources such as information and community building know-how are transferable across platforms communities on platforms designed to provide different types of benefits are likely to be mutualists. For example, subreddits and Wikis about similar topics are probably mutualists because wikis are designed primarily for developing and sharing encyclopedic information and subreddits often focus on socialization and discussion.

Now I will sketch several possible directions for near-future work in this research program. Some of these potential projects seek to develop more complete answers for the key theoretical questions and others will bridge ecological analysis to specific practical problems. My hope is that empirical support and theoretical development will soon be sufficiently advanced to inform the design of present and future online community ecosystems and to understand the successes and limitations of peer production.

5.1. Ecological Relationships Between Platforms

A significant limitation of my empirical studies has been that they focus only on interactions among communities within a single large platform. However, online communities often overlap across platforms (Kiene et al., 2019) and cross-platform interactions are likely to be important (Vincent et al., 2018). For example, Reddit's growth enormously increased in 2010 when users of rival site Digg.com migrated *en mass*, suggesting that during this period subreddits and Digg sections were in competition ("Digg," 2021). In chapter 6 of *Building Successful Online Communities*, Resnick et al. (2012) recommend that new online communities "carve out a useful and defendable (sic) niche in the ecology of competing communities." They base this recommendation upon virtually no evidence taken from studies of online communities or organizational ecology but rather by following intuitions drawn from economics and assuming that online communities may find themselves in "winner-take-all" situations. Although they recommend specialization as a strategy for avoiding competition, they also suggest "lock-in" features like having different user interfaces and making it so identities cannot be shared between communities.

At issue is how Resnick et al. (2012) attempt to simultaneously adopt the perspectives of two different types of actors whose interests are often unaligned. Commercial platforms need to generate private revenues and seem to better fit the classical models of organizational ecology that have niche overlaps as highly correlated with competition. A commercial platform may find mutualism between cross-platform communities a nuisance and may find the "lock-in" features unequivocally beneficial. However, building a successful online community is not the same as building a platform that hosts online communities. My ecological studies of relationships between communities suggest that mutualism is widespread among actually existing online communities within a platform. In my conversations with members of overlapping communities, I learned that they often benefit from overlapping communities on different platforms. Therefore it seems likely that that communities on commercial platforms that are both sufficiently "open" and sufficiently differentiated will also be mutualistic, even if the platforms compete with each other over revenues. If so, this points to the promise of designs that support resource sharing across such platforms.

Knowledge about inter-platform ecological dynamics is only beginning to be created. Nagaraj and Piezunka (2021) have found that open source knowledge projects like open street map

are hurt by competition with proprietary alternatives. Cross-platform studies of digital traces face difficulties because it is not generally possible to associate user accounts on different platforms. However, the time-series models I have used only depend on finding related communities and therefore enable studying ecological interactions without tracking users across platforms. I am developing a new dataset of related subreddits, Fandom.com wikis, and Wikipedia articles to investigate ecological interactions between related communities on different platforms.

5.2. Selecting Niche Width

Choosing a scope is an important design decision for organizations and for online communities. As I found in Chapter 3, broad and narrow scopes are associated with trade-offs in the types of benefits that a community can provide. The choice of scope, or the choice of how a community will specialize, may also have implications for the community's short and long run survival. According to theories of organizational ecology, the choice of scope may affect a community's competitive and mutualistic dynamics and its ability to weather changes in a turbulent environment.

Resource partitioning theory, discussed briefly in sections of Chapters 2, 3 and 4, provides a framework for understanding how specialization relates to competition. It proposes that larger generalists can coexist with specialists because large generalists are not optimally efficient at all of their activities, leaving opportunities for specialists to out-compete them in narrow niches (Carroll, 1985). Findings from Chapter 2 suggest that one prediction of resource partitioning theory seems to obtain in groups of overlapping online communities. This is that they often have a "main" community which is a large generalist and people participate in the specialist communities in order to obtain distinctive benefits not easily obtained in the main community (Baum & Shipilov, 2006).

A related theory fragment of organizational ecology, niche width theory (Dobrev et al., 2001; J. Freeman & Hannan, 1983), proposes that specialists are less able to survive during periods of rapid change. Large generalists may have advantages in changing environments because their diversity of interests which spreads out risk, their experience transferring knowledge between different parts of their organization, and their slack resources can all help them absorb negative outcomes (Dobrev et al., 2003). As discussed in Chapter 4, online communities may inhabit unstable environments where sudden events, ongoing trends, and abrupt policy changes can all affect participation (Ratkiewicz et al., 2010).

An example illustrates how environmental change can threaten the success of specialists. During the Trump administration, a number of anti-trump subreddits were organized around specific controversies (e.g., `r/the_meuller`, `r/marchagainsttrump`, `r/keep_track`, `r/russialago`). `r/the_mueller` was a subreddit about the Special Counsel's investigation into Russian election interference. the number of posts in these subreddits declined following the end of the investigation. However, this subreddit has survived by successfully adapted and now has several posts a day critical of Trump but not specifically about the Meuller investigation. Yet a similar subreddit, `r/russialago` has declined to a much lower activity level (a few posts a week) but remains focused on Russian interference. By comparison, the number of posts in the generalist (but still left-leaning) `r/politics` has remained relatively stable. Niche width theory would predict that shifting to more general types of anti-Trump content may expose `r/the_-`

mueller to greater competition with other political subreddits. However, if it had not adapted it might have little reason to exist after the end of Mueller's investigation.

Theories of online community specialization can be empirically testable with better quantification of the ways that overlapping communities are different from one another. These include features of content like choice of medium (text, images, video, links), content sources (what websites are they linking to?), types of participants with varying roles and styles of participation, and structures like policies, size and moderation. Niche width theory additionally requires measuring environmental changes that may threaten the survival of communities. Observable events corresponding to interesting environmental variation may include crisis events, elections and the release or cancellation of entertainment products. Comparing the growth, performance, and ecological dynamics of overlapping communities during times of high or low change can test these theories and point toward design principles for online community scoping that account for the trade-offs in different types of specialization.

5.3. Ecological Implications for Production and Performance

So far, the ecology of online communities has focused on understanding competition and mutualism among overlapping online communities. An important limitation of this work has been to conceptualize competition and mutualism as dynamics related to the growth of online communities. This follows biologists and organizational ecologists, but not all online communities have to grow in order to provide their intended benefits (Foote et al., 2017). An important step forward this research program will be to relate interdependence between online communities to outcomes besides growth that may be more directly connected to the value of the public goods that communities produce.

Quantifying the value of public information goods produced by online communities is a major methodological and theoretical challenge. Much of the field of economics depends on the assumption that the utility of a good can be measured by its price. Price is a valuable measure of value in economic theory because it is set by market mechanisms that align supply and demand. Online communities are thought to be able to produce public goods because they can lower transaction costs (Benkler, 2002). Negotiating a price in these settings is simply not worth it. A price will reintroduce transaction costs and undercut the pro-social motivations people have for contributing.

Of course, this does not mean the public goods online communities produce are worthless. Estimates of the cost of replacing by paying editors a market rate placed its value between 6 and 10 billion dollars in 2013 (Band & Gerafi, 2013). However, without a price mechanism, supply and demand may become "misaligned." The quality of Wikipedia articles is uneven and the most popular content is often not the highest quality (Gorbatai, 2011; Warncke-Wang et al., 2015). In classical economic theories, goods will be produced to meet the demand, but in peer production the size of an audience seems only weakly related to the level of production. Explaining when online communities will produce high quality public goods like Wikipedia articles (Arazy et al., 2019; Arazy & Nov, 2010; Asthana & Halfaker, 2018) or open source software (Champion & Hill, 2021) is thus important to understanding the successes and failures of peer production.

Critical mass theory offers to explain the conditions for successful collective action in public goods production and can also be synthesized with ecology (Marwell & Oliver, 1993). Many CSCW systems appear to require a critical mass of users to start or sustain their usefulness

(Ackerman, 2000). The most important device in the theory is the *production function*, which maps an individual's contributions to the value they get from contributing. The theory proposes that the shape of the production function is determined by the collective action problem that a group faces in producing the good. If a production function is *accelerating* (*decelerating*) then a contribution increases (decreases) the payoff of the next contribution. The rational actors in a group each have their own production function and together these determine the level of the good that they will produced. Some prior research applies this theory to Wikipedia, but does not attempt to measure value of contribution or operationalize the theory's propositions about the relationship between production functions and collective action (Raban et al., 2010; Solomon & Wash, 2014). Analyzing critical mass theory in the context of communal public goods production can also be an important theoretical contribution to communication theory (Fulk et al., 1996).

To illustrate, consider a hypothetical example of the construction of an online community for building a collaborative knowledge base, such as Wikidata. This can be cast as a collective action problem because the project can provide a wide range of benefits to a potentially large group of people, but no individual can provide the full range benefits alone (Fulk et al., 1996; Marwell & Oliver, 1993). Say a single individual, the community's founder who is an expert engineer and researcher, attempts to bootstrap the community by providing an initial design and implementation for the novel system, a small number of entries and by making efforts to publicize the community. The founder hopes that others to join and contribute to constructing a valuable resource.

During this period in the community's development, the *critical mass* consists of just the founder, who is motivated and capable of in the hopes that others will see these contributions and subsequently make their own. The founder has a large and unique set of resources enabling them pay the *start-up costs* involved in founding the community when no one else would. After these start-up costs are paid, others can make much more granular contributions like adding entries to the knowledge base. The founder hopes that others will perceive expected benefits from contributing that exceed the costs of contributing. In theoretical terms, the founder hopes that the others' production functions are accelerating and paying the start up costs will move the others' production functions into a favorable region where they will contribute.

Ecology has important implications for critical mass theory because important aspects of the collective action problem that influence the production function are related to the composition of the group and prior work suggests that individuals with varying experiences are important to online community growth (Kairam et al., 2012). Heterogeneous groups are thought to be conducive to collective action because they are more likely to contain individuals who can contribute different things like start up costs or rare pieces of information Fulk et al., 1996.

Returning to the example of a collaborative knowledge base, it is important to recognize that many contributions will involve *articulation work* activities like documenting, answering questions, naming, and interpreting that are required to make the knowledge base work in practice (Schmidt & Bannon, 1992; Suchman, 1996). Even though contributions of articulation work might not directly add new features or data to the knowledge base, they can be important to accelerating community members' production functions. A heterogeneous community may be more likely to include members who are skilled at articulation work that benefits other members. On the other hand, If different subgroups of a large community have sufficiently different application areas some articulation work might be specific to each subgroup. For example,

biologists might make and document biology-specific norms for the collaborative knowledge base, but this would not be useful to physicists. Thus individuals' production functions might depend most strongly on the other members of their subgroup when subgroup-specific articulation work is a limiting factor.

I am starting work to find out how production functions help explain when online communities achieve critical mass and produce quality outputs and if relationships among communities influence the shape of production functions in ways that make collective action easier or more difficult in different conditions. Measuring production functions requires the ability to precisely quantify the quality or value of individual contributions. As a step in this direction, I have developed an improved measurement of Wikipedia article quality in research accepted for publication and included in Appendix A. Prior article quality measures have been based on machine learning models that do not provide a continuous measure amenable to statistical analysis and that were miscalibrated for units of analysis like articles or projects. Research using these measures has got around these problems by adopting an assumption that article quality levels on Wikipedia are "evenly spaced" from one another. I use a method that relaxes this assumption, provides evidence that it is unfounded, and improves the accuracy of the models.

I have also done some methodological work on the "demand side" to understand how audiences use Wikipedia content. Most prior work has been limited to measuring page views. In Appendix B, I study the amount of time spent reading articles by Wikipedia visitors and find that readers in the Global South remain on pages for longer, especially in the last page view in a session. Although the measure used in that study may not be available for use in the future, this work has prepared me for the time when better reading time data is available. It will be interesting to see if the audience for an article relates to critical mass dynamics.

5.4. Ecology and the Diffusion of Technologies for Community Governance

Future ecological research can also look at the role of ecological dynamics in the emergence and diffusion of novel artifacts, technologies, information and ideas. Overlapping technology use in particular is a potential mechanism for specialization and mutualism. I have previously suggested that sharing a host platform may not be sufficient for defining an organizational form because communities have considerable flexibility in making their own rules and configuring their own custom technology. If sufficiently strong patterns are found in the sets of rules or technologies that communities adopt, these might justify treating communities sharing such structures as organizational forms or at least a potentially important kind of niche overlap.

When online communities share technologies, this can create important forms of interdependence and collaborative innovation on tools is potentially an important type of mutualism. For example Chandrasekharan et al. (2019) developed a system called "Cross Mod" for subreddits to collaborate on customizable machine learning models for monitoring misbehavior. Smaller communities pooling data about rule violations can potentially build more accurate models than single communities can. Technologies like Cross Mod allow communities to select which other communities they wish to import data from and therefore are most useful when communities are institutionally compatible. This suggests that sharing governance technologies may be a good proxy for an organizational form.

However, as I found in Appendix C, my study of algorithmic flagging tools on Wikipedia, machine learning tools for predicting misbehavior may reproduce the biases of community mod-

erators. They can also improve the fairness of moderator judgments if moderators use the models instead of other biased social signals to find potential misbehavior. Additional risks may arise when algorithmic tools are shared by overlapping communities. The learned norms and and standards of behavior from one community may not be appropriate in other communities. If shared flagging algorithms can more easily implement norms that are more widely held, the diffusion of an algorithm that makes regulating behavior easier and more predictable might mediate the diffusion of the norm.

The method I developed for the study in Appendix C provides a way to assess the consequences of a machine learning classifier without intervening in a community. Future work at the intersection of ecology and online community governance might use this method in a study of the relationships between the performance of algorithms for enforcing different rules, the diffusion of the rules, and the growth and survival of communities having the rules.

5.5. Microfoundations for Ecological Macrodynamics

Predominant approaches in HCI and social computing and popular conceptions of social media platforms most often emphasize the role of managers of platforms in building online communities. However, platforms have only a limited control over the ways that users build communities. Furthermore, platforms struggle to maintain participants who may migrate to competing platforms. Communities and their organizers can engage in collective action to protest platform's governance and design decisions (Matias, 2016). Online communities also form intermediate structures over which platforms have limited influence such as the widespread clusters of highly overlapping communities I identify in Chapter 2. An important goal of the ecology of online communities is to understand how patterns of action within individual communities are co-constitutive with the cultures and institutions of platforms. This goal faces a key type of puzzle in social science: to account for how "micro-level" individual actors give rise to "macro-level" organizations, institutions, online communities, and cultures even as individuals are situated within these very structures.

Micro-macro puzzles are not only found in the constitution of individual persons and the social structures they inhabit. Organizational ecology takes up a different kind of micro-macro puzzle at the level of reciprocal dependence between organizations and the organizational fields or industries they comprise. The performance of an individual organization depends on ecological dynamics in its organizational field, but the organization itself contributes to these very dynamics. Initial work in organizational ecology avoided this reciprocal causation by minimizing the action of individual organizations. Structural inertia constrained the agency of organizational actors, and external institutions, competition, and legitimacy constrained organizational performance.

At first, organizational ecologists did not deny that factors internal to organizations matter to organizational performance. Yet they argued that *ceteris paribus*, the chances of an organization's survival depend on environmental conditions and on mutualistic and competitive pressures (Hannan & Freeman, 1989). Later on, organizational ecology began accounting for rational adaptation and failure of individual organizations (Baum & Shipilov, 2006). Recently they have incorporated the role of human cognition and social learning into their conceptualizations (Hannan, 2019), but as far as I am aware, empirical analyses have not stretched all the way from individual persons to inter-organizational dynamics.

Online communities provide a distinctive opportunity to connect individual behaviors to outcomes at the community and ecological levels thanks to the finely grained behavioral data that made possible the analyses in Chapters 2 and 4. However, all of the measures used in these projects have aggregated the behavior of many individuals into measures of overlap or group size. I have not shown how the ways that individuals navigate among overlapping online communities give rise to the ecological dynamics I find. Aware of this limitation, I initially proposed constructing an agent-based model to theorize the micro-mechanisms of ecological dynamics. Along the way, I found that talking to individuals provided a more valuable micro-level account of how and why people participate in overlapping online communities.

These interviews surfaced a conceptual model of a process by which new communities in a topical area spin-off specialists. An important direction for future research will be to operationalize and test this model with data. This future work should look for inspiration from measures of individual behavior introduced in recent research in HCI and social computing (Tan, 2018; Tan & Lee, 2015; J. S. Zhang et al., 2021). Specifically, (Tan, 2018) provide a method to associate newly created subreddits with prior subreddits whose users join the new subreddit and measure the language use of individuals to characterize their similarity to the other members of the community. Also, Waller and Anderson (2019) quantify users of online communities as generalists and specialists based on their activity styles using embedding methods.

5.6. Focused Case Studies

Finally, in order for ecological research in online communities to be useful to publics and practitioners, it will be important to conduct focused case studies of practical and popular interest. Studies of the ecology of political communities, communities tying to make sense of the pandemic, "meme stock" and cryptocurrency communities, and pop culture fandom communities are all promising candidates. A future project should investigate one or more cases in a mixed-methods study combining carefully constructed time series models for inferring ecological relationships and qualitative data in the form of grounded narrative accounts or interviews.

In conclusion, my research set out to understand interdependence among online communities through the lens of organizational ecology. It has questioned the how well foundational assumptions of organizational ecology apply to online communities and set out to validate basic assumptions like when online communities will form competitive or mutualistic relationships. It has provided new methods for studying competition and mutualism among online communities and shown that mutualistic relationships are more common than competitive ones because they last longer. Although the question of how groups of mutualistic online communities are constructed remains open, selection process theories provide a starting point for future investigation. Many applications of ecological theories and methods to important questions about the emergence, performance, and design of online communities are promising.

References

Ackerman, M. S., & Malone, T. W. (1990). Answer Garden: A Tool for Growing Organizational Memory. *Proceedings of the ACM SIGOIS and IEEE CS TC-OA Conference on Office Information Systems*, 31–39.

Ackerman, M. S. (2000). The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human–Computer Interaction*, *15*(2-3), 179–203 _eprint: https://doi.org/10.1207/S15327051HCI1523_5.

Ackerman, M. S., Dachtera, J., Pipek, V., & Wulf, V. (2013). Sharing Knowledge and Expertise: The CSCW View of Knowledge Management. *Computer Supported Cooperative Work (CSCW)*, *22*(4-6), 531–573.

Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: Everyone knows something. *Proceedings of the 17th International Conference on World Wide Web*, 665–674.

Aldrich, H., & Ruef, M. (2006). *Organizations Evolving* (2nd ed.). SAGE Publications.

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data. *Social Science Computer Review*, *39*(5), 943–960.

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

Arazy, O., Lindberg, A., Rezaei, M., & Samorani, M. (2019). The evolutionary trajectories of peer-produced artifacts: Group composition, the trajectories' exploration, and the quality of artifacts. *MIS Quarterly*.

Arazy, O., & Nov, O. (2010). Determinants of wikipedia quality: The roles of global and local contribution inequality. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 233–236.

Arazy, O., Ortega, F., Nov, O., Yeo, L., & Balila, A. (2015). Functional roles and career paths in Wikipedia. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1092–1105.

Armstrong, R. A., & McGehee, R. (1980). Competitive Exclusion. *The American Naturalist*, *115*(2), 151–170.

Asthana, S., & Halfaker, A. (2018). With Few Eyes, All Hoaxes Are Deep. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), 21:1–21:18.

Astley, W. G. (1985). The Two Ecologies: Population and Community Perspectives on Organizational Evolution. *Administrative Science Quarterly*, *30*(2), 224–241.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396.

Band, J., & Gerafi, J. (2013). *Wikipedia's Economic Value* (SSRN Scholarly Paper). Social Science Research Network. Rochester, NY.

Barnett, W. P., & Carroll, G. R. (1987). Competition and mutualism among early telephone companies. *Administrative Science Quarterly*, *32*(3), 400–421.

Baum, J. A. C., & Shipilov, A. V. (2006). Ecological approaches to organizations. *Sage Handbook for Organization Studies* (pp. 55–110). Sage.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 830–839.

Benkler, Y. (2002). Coase's penguin, or, Linux and 'The nature of the firm'. *The Yale Law Journal*, *112*(3), 369.

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

Benkler, Y., Roberts, H., Faris, R., Solow-Niederman, A., & Etling, B. (2013). *Social Mobilization and the Networked Public Sphere: Mapping the SOPA-PIPA Debate* (SSRN Scholarly Paper No. ID 2295953). Social Science Research Network. Rochester, NY.

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

Bernstein, M. S., Bakshy, E., Burke, M., & Karrer, B. (2013, April 27). Quantifying the invisible audience in social networks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 21–30). Association for Computing Machinery.

Bimber, B. A., Flanagin, A. J., & Stohl, C. (2012). *Collective action in organizations: Interaction and engagement in an era of technological change*. Cambridge University Press.

Box-Steffensmeier, J. M. (2014). *Time series analysis for the social sciences* OCLC: 879601718.

Brandtzæg, P. B., & Heim, J. (2008). User Loyalty and Online Communities: Why Members of Online Communities are not Faithful. *Proceedings of the 2nd International Conference on INtelligent TEchnologies for Interactive enterTAINment*.

Bryant, S. L., Forte, A., & Bruckman, A. (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.

Burtch, G., He, Q., Hong, Y., & Lee, D. (2021). How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit. *Management Science*.

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

Butler, B. S., Bateman, P. J., Gray, P. H., & Diamant, E. I. (2014). An attraction-selection-attrition theory of online community size and resilience. *MIS Q.*, *38*(3), 699–728.

Butler, B. S., & Wang, X. (2011). The cross-purposes of cross-posting: Boundary reshaping behavior in online discussion communities. *Information Systems Research*, *23*, 993–1010.

Campbell, J., Aragon, C., Davis, K., Evans, S., Evans, A., & Randall, D. (2016). Thousands of Positive Reviews: Distributed Mentoring in Online Fan Communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 691–704.

Canova, F. (2007). VAR Models. *Methods for Applied Macroeconomic Research* (pp. 111–164). Princeton University Press.

Carroll, G. R. (1985). Concentration and specialization: Dynamics of niche width in populations of organizations. *American Journal of Sociology*, *90*(6), 1262–1283.

Carroll, G. R., & Hannan, M. T. (1989). Density dependence in the evolution of populations of newspaper organizations. *American Sociological Review*, *54*(4), 524.

Carroll, G. R., & Swaminathan, A. (2000). Why the microbrewery movement? Organizational dynamics of resource partitioning in the U.S. brewing industry. *American Journal of Sociology*, *106*(3), 715–762.

Cenci, S., Sugihara, G., & Saavedra, S. (2019). Regularized S-map for inference and forecasting with noisy ecological time series. *Methods in Ecology and Evolution*, *10*(5), 650–660 _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13150.

Certain, G., Barraquand, F., & Gårdmark, A. (2018). How do MAR(1) models cope with hidden nonlinearities in ecological dynamics? *Methods in Ecology and Evolution*, *9*(9), 1975–1995.

Champion, K., & Hill, B. M. (2021). Underproduction: An approach for measuring risk in open source software. *IEEE International Conference on Software Analysis, Evolution and Reengineering*.

Chandrasekharan, E., Gandhi, C., Mustelier, M. W., & Gilbert, E. (2019). Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–30.

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

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.*, *2*, 32:1–32:25.

Chang, S., Kumar, V., Gilbert, E., & Terveen, L. G. (2014). Specialization, homophily, and gender in a social curation site: Findings from pinterest. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 674–686.

Charmaz, K. (2015). *Constructing grounded theory: A practical guide through qualitative analysis* (2nd ed.). SAGE.

Choudhury, M. D., Jhaver, S., Sugar, B., & Weber, I. (2016). Social Media Participation in an Activist Movement for Racial Equality. *Tenth International AAAI Conference on Web and Social Media*.

Cunha, T., Jurgens, D., Tan, C., & Romero, D. (2019). Are all successful communities alike? Characterizing and predicting the success of online communities. *The World Wide Web Conference*, 318–328.

Dabbish, L., Farzan, R., Kraut, R., & Postmes, T. (2012). Fresh faces in the crowd: Turnover, identity, and commitment in online groups. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 245–248.

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

Datta, S., & Adar, E. (2019). Extracting Inter-Community Conflicts in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 146–157.

Datta, S., Phelan, C., & Adar, E. (2017). Identifying Misaligned Inter-Group Links and Communities. *Proceedings of the ACM on Human-Computer Interaction*, *1*, 37:1–37:23.

De Choudhury, M., & De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 71–80.

Del Tredici, M., & Fernández, R. (2018, June 15). *Semantic Variation in Online Communities of Practice*. arXiv: 1806.05847 [cs].

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why Do Liberals Drink Lattes? *American Journal of Sociology*, *120*(5), 1473–1511.

Digg. (2021, August 26). In *Wikipedia*
Page Version ID: 1040737272.

DiMaggio, P., Hargittai, E., Neuman, W. R., & Robinson, J. P. (2001). Social implications of the Internet. *Annual Review of Sociology*, *27*(1), 307–336.

DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, *48*(2), 147–160.

Dimmick, J., & Rothenbuhler, E. (1984). The Theory of the Niche: Quantifying Competition Among Media Industries. *Journal of Communication*, *34*(1), 103–119.

Dobrev, S. D., Kim, T.-Y., & Carroll, G. R. (2003). Shifting Gears, Shifting Niches: Organizational Inertia and Change in the Evolution of the U.S. Automobile Industry, 1885-1981. *Organization Science*, *14*(3), 264–282.

Dobrev, S. D., Kim, T.-Y., & Hannan, M. T. (2001). Dynamics of niche width and resource partitioning. *American Journal of Sociology*, *106*(5), 1299–1337.

Ducheneaut, N., Yee, N., Nickell, E., & Moore, R. J. (2006). "Alone together?": Exploring the social dynamics of massively multiplayer online games. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 407–416.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 188–230
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440380105.

Dvir-Gvirsman, S. (2017). Media audience homophily: Partisan websites, audience identity and polarization processes. *New Media & Society*, *19*(7), 1072–1091.

Faraj, S., von Krogh, G., Monteiro, E., & Lakhani, K. R. (2016). Online community as space for knowledge flows. *Information Systems Research*, *27*(4), 668–684.

Fiesler, C., & Dym, B. (2020). Moving Across Lands: Online Platform Migration in Fandom Communities. *Proc. ACM Hum.-Comput. Interact*, *4*, 042:1–042:25.

Fiesler, C., Jiang, J. A., McCann, J., Frye, K., & Brubaker, J. R. (2018). Reddit rules! Characterizing an ecosystem of governance. *Proceedings of the International AAAI Conference on Web and Social Media*, 72–81.

Fiesler, C., Morrison, S., Shapiro, R. B., & Bruckman, A. S. (2017). Growing Their Own: Legitimate Peripheral Participation for Computational Learning in an Online Fandom Community. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1375–1386.

Foote, J. (2019). *The formation and growth of collaborative online organizations* (PhD dissertation). Northwestern University. Evanston, IL.

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

Freeman, J., & Hannan, M. T. (1983). Niche width and the dynamics of organizational populations. *American Journal of Sociology*, *88*(6), 1116–1145.

Freeman, J. H., & Audia, P. G. (2006). Community ecology and the sociology of organizations. *Annual Review of Sociology*, *32*, 145–169.

Frey, B. J., & Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, *315*(5814), 972–976.

Frey, S., & Sumner, R. W. (2019). Emergence of integrated institutions in a large population of self-governing communities. *PLOS ONE*, *14*(7), e0216335.

Fu, H., & Stvilia, B. (2016). Knowledge curation discussions and activity dynamics in a short lived social Q amp;A community. *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, 203–204.

Fulk, J., Flanagin, A. J., Kalman, M. E., Monge, P. R., & Ryan, T. (1996). Connective and communal public goods in interactive communication systems. *Communication Theory*, *6*(1), 60–87.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press
OCLC: on1005113962.

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378
_eprint: https://doi.org/10.1198/016214506000001437.

Gorbatai, A. D. (2011). Exploring Underproduction in Wikipedia. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 205–206.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Grevet, C., Terveen, L. G., & Gilbert, E. (2014). Managing political differences in social media. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1400–1408.

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

Hamilton, W. L., Zhang, J., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017, May 24). *Loyalty in online communities*. arXiv: 1703.03386 [cs].

Hannan, M. T. (2019). *Concepts and categories: Foundations for sociological and cultural analysis*
OCLC: 1083703599.

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

Hannan, M. T., & Freeman, J. (1984). Structural inertia and organizational change. *American Sociological Review*, *49*(2), 149.

Hannan, M. T., & Freeman, J. (1989). *Organizational ecology* (1st ed.). Harvard University Press.

Hannan, M. T., Pólos, L., & Carroll, G. (2007). *Logics of organization theory: Audiences, codes, and ecologies*. Princeton University Press
OCLC: 646517503.

Hawley, A. H. (1986). *Human ecology: A theoretical essay*. University of Chicago Press
OCLC: 993363851.

Heaps, S. E. (2020, April 20). *Enforcing stationarity through the prior in vector autoregressions* (1). arXiv: 2004.09455 [stat].

Hessel, J., Tan, C., & Lee, L. (2016). Science, askscience, and badscience: On the coexistence of highly related communities. *Tenth International AAAI Conference on Web and Social Media*, 11.

Hill, B. M., & Shaw, A. (2019, September). Studying populations of online communities. In B. Foucault Welles & S. González-Bailón (Eds.), *The Oxford Handbook of Networked Communication* (pp. 173–193). Oxford University Press.

Hillman, S., Procyk, J., & Neustaedter, C. (2014). 'alksjdf;Lksfd': Tumblr and the fandom user experience. *Proceedings of the 2014 Conference on Designing Interactive Systems*, 775–784.

Himelboim, I., Sweetser, K. D., Tinkham, S. F., Cameron, K., Danelo, M., & West, K. (2016). Valence-based homophily on Twitter: Network Analysis of Emotions and Political Talk in the 2012 Presidential Election. *New Media & Society*, *18*(7), 1382–1400.

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

Hwang, S., & Foote, J. D. (2021). Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction*.

Ives, A. R., Dennis, B., Cottingham, K. L., & Carpenter, S. R. (2003). Estimating Community Stability and Ecological Interactions from Time-Series Data. *Ecological Monographs*, *73*(2), 301–330.

Jarvenpaa, S. L., & Leidner, D. E. (1998). Communication and trust in global virtual teams. *Journal of Computer-Mediated Communication*, *3*(4), 0–0.

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

Johnson, T. J., Bichard, S. L., & Zhang, W. (2009). Communication Communities or "CyberGhettos?": A Path Analysis Model Examining Factors that Explain Selective Exposure to Blogs. *Journal of Computer-Mediated Communication*, *15*(1), 60–82.

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software*, *90*(1), 1–37.

Kairam, S. R., Wang, D. J., & Leskovec, J. (2012). The life and death of online groups: Predicting group growth and longevity. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, 673–682.

Kantz, H., & Schreiber, T. (2003). *Nonlinear Time Series Analysis* (2nd ed.). Cambridge University Press.

Katz, M. L., & Shapiro, C. (1985). Network Externalities, Competition, and Compatibility. *The American Economic Review*, *75*(3), 424–440.

Keegan, B., Lev, S., & Arazy, O. (2016). Analyzing Organizational Routines in Online Knowledge Collaborations: A Case for Sequence Analysis in CSCW. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1065–1079.

Kiene, C., Jiang, J. ", & Hill, B. M. (2019). Technological frames and user innovation: Exploring technological change in community moderation teams. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 44:1–44:23.

Kiene, C., Monroy-Hernández, A., & Hill, B. M. (2016). Surviving an "Eternal September": How an online community managed a surge of newcomers. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1152–1156.

Kiene, C., Shaw, A., & Hill, B. M. (2018). Managing organizational culture in online group mergers. *Proc. ACM Hum.-Comput. Interact.*, *2*, 89:1-89–21.

Koh, J., Kim, Y.-G., Butler, B., & Bock, G.-W. (2007). Encouraging participation in virtual communities. *Communications of the ACM*, *50*(2), 68–73.

Krafft, P. M., & Donovan, J. (2020). Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign. *Political Communication*,

37(2), 194–214
_eprint: https://doi.org/10.1080/10584609.2019.1686094.

Kraut, R. E., & Fiore, A. T. (2014). The Role of Founders in Building Online Groups. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 722–732.

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

Kropotkin, P. (2012, May 2). *Mutual aid: A factor of evolution*. Courier Corporation. (Original work published 1902)

Kubiszewski, I., Farley, J., & Costanza, R. (2010). The production and allocation of information as a good that is enhanced with increased use. *Ecological Economics*, *69*(6), 1344–1354.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. University of Chicago Press
OCLC: 959412835.

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community Interaction and Conflict on the Web. *Proceedings of the 2018 World Wide Web Conference*, 933–943.

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

Lam, S. (K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., & Riedl, J. (2011). WP:clubhouse?: An Exploration of Wikipedia's Gender Imbalance. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 1–10.

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 543–550.

Lampe, C., Wash, R., Velasquez, A., & Ozkaya, E. (2010). Motivations to participate in online communities. *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, 1927–1936.

Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, & C. H. Page (Eds.), *Freedom and control in modern society* (pp. 18–66). Van Nostrand.

Leavitt, A., & Robinson, J. J. (2017). The role of information visibility in network gatekeeping: Information aggregation on reddit during crisis events. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1246–1261.

Leimeister, J. M., & Krcmar, H. (2005). Evaluation of a Systematic Design for a Virtual Patient Community. *Journal of Computer-Mediated Communication*, *10*.

Liang, Y. (2017). Knowledge sharing in online discussion threads: What predicts the ratings? *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 146–154.

Lin, Z., Salehi, N., Yao, B., Chen, Y., & Bernstein, M. S. (2017). Better when it was smaller? Community content and behavior after massive growth. *Eleventh International AAAI Conference on Web and Social Media*, 132–141.

Litt, E., & Hargittai, E. (2016). "Just Cast the Net, and Hopefully the Right Fish Swim into It": Audience Management on Social Network Sites. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1488–1500.

Lu, J., Sridhar, S., Pandey, R., Hasan, M. A., & Mohler, G. (2019). Investigate Transitions into Drug Addiction through Text Mining of Reddit Data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2367–2375.

Ma, X., Cheng, J., Iyer, S., & Naaman, M. (2019). When Do People Trust Their Social Groups? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Manning, C. D., Raghavan, P., Schütze, H., & Cambridge University Press. (2018). *Introduction to information retrieval*. Cambridge University Press
OCLC: 1077323048.

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015, November 24). *Political turbulence: How social media shape collective action*. Princeton University Press.

Margolin, D. B., Shen, C., Lee, S., Weber, M. S., Fulk, J., & Monge, P. (2012). Normative Influences on Network Structure in the Evolution of the Children's Rights NGO Network, 1977-2004: *Communication Research*.

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

Marwick, A. E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133.

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346.

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

Matias, J. N. (2016). Going dark: Social factors in collective action against platform operators in the Reddit blackout. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 1138–1151.

Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media + Society*, *5*(2), 1–12.

McInnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205.

McPherson, J. M. (1983). An ecology of affiliation. *American Sociological Review*, *48*(4), 519–532.

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

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

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444.

Menge, B. A. (1972). Competition for Food between Two Intertidal Starfish Species and its Effect on Body Size and Feeding. *Ecology*, *53*(4), 635–644.

Menking, A., Erickson, I., & Pratt, W. (2019). People Who Can Take It: How Women Wikipedians Negotiate and Navigate Safety. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery.

Minkoff, D. C. (1995). Interorganizational influences on the founding of african american organizations, 1955–1985. *Sociological Forum*, *10*(1), 51–79.

Morris, M. R., Teevan, J., & Panovich, K. (2010a). A Comparison of Information Seeking Using Search Engines and Social Networks. *Fourth International AAAI Conference on Weblogs and Social Media*.

Morris, M. R., Teevan, J., & Panovich, K. (2010b, April 10). What do people ask their social networks, and why? a survey study of status message q&a behavior. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1739–1748). Association for Computing Machinery.

Muhtaseb, A., & Frey, L. R. (2008). Arab Americans' Motives for Using the Internet as a Functional Media Alternative and Their Perceptions of U.S. Public Opinion. *Journal of Computer-Mediated Communication*, *13*(3), 618–657.

Nagaraj, A., & Piezunka, H. (2021). How Competition Affects Contributions to Open Source Platforms: Evidence from OpenStreetMap and Google Maps, 58.

Nardi, B. A., & O'Day, V. L. (1999). *Information Ecologies : Using technology with heart*. The MIT Press.

Novak, M., Yeakel, J. D., Noble, A. E., Doak, D. F., Emmerson, M., Estes, J. A., Jacob, U., Tinker, M. T., & Wootton, J. T. (2016). Characterizing Species Interactions to Understand Press Perturbations: What Is the Community Matrix? *Annual Review of Ecology, Evolution, and Systematics*, *47*(1), 409–432
_eprint: https://doi.org/10.1146/annurev-ecolsys-032416-010215.

Olzak, S., & Uhrig, S. C. N. (2001). The ecology of tactical overlap. *American Sociological Review*, *66*(5), 694.

Orlikowski, W. J. (1992). Learning from notes: Organizational issues in groupware implementation. *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work*, 362–369.

Ostrom, V., & Ostrom, E. (1977). Public goods and public choices. In E. S. Savas (Ed.), *Alternatives For Delivering Public Services: Toward Improved Performance* (pp. 7–49). Westview Press.

Park, R. E. (1936). Human Ecology. *American Journal of Sociology*, *42*(1), 1–15.

Peters, J. D. (1999). *Speaking into the air: A history of the idea of communication*. The University of Chicago press.

Pfaff, B. (2008). VAR, SVAR and SVEC Models: Implementation Within R Package vars. *Journal of Statistical Software*, *27*(1), 1–32.

Piantadosi, S., Byar, D. P., & Green, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology*, *127*, 893–904.

Pontikes, E., & Hannan, M. (2014). An Ecology of Social Categories. *Sociological Science*, *1*, 311–343.

Poor, N. (2005). Mechanisms of an Online Public Sphere: The Website Slashdot. *Journal of Computer-Mediated Communication*, *10*.

Poor, N. (2014). Computer game modders' motivations and sense of community: A mixed-methods approach. *New Media & Society*, *16*(8), 1249–1267.

Popielarz, P. A., & McPherson, J. M. (1995). On the Edge or In Between: Niche Position, Niche Overlap, and the Duration of Voluntary Association Memberships. *American Journal of Sociology*, *101*(3), 698–720.

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

Raban, D. R., Moldovan, M., & Jones, Q. (2010). An empirical study of critical mass and online community survival. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 71–80.

Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., & Vespignani, A. (2010). Characterizing and Modeling the Dynamics of Online Popularity. *Physical Review Letters*, *105*(15).

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

Ribeiro, M. H., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–24.

Ridings, C. M., & Gefen, D. (2004). Virtual Community Attraction: Why People Hang out Online. *Journal of Computer-Mediated Communication*, *10*(1).

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, *15*(3), 351–357.

Romer, P. M. (1990). Endogenous Technological Change. *Journal of Political Economy*, *98*, S71–S102.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Ruef, M. (2000). The Emergence of Organizational Forms: A Community Ecology Approach. *American Journal of Sociology*, *106*(3), 658–714.

Ruef, M., Aldrich, H. E., & Carter, N. M. (2003). The Structure of Founding Teams: Homophily, Strong Ties, and Isolation among U.S. Entrepreneurs. *American Sociological Review*, *68*(2), 195–222.

Schmidt, K., & Bannon, L. (1992). Taking CSCW seriously. *Computer Supported Cooperative Work (CSCW)*, *1*(1-2), 7–40.

Schoener, T. W. (1974). Resource Partitioning in Ecological Communities. *Science*, *185*(4145), 27–39.

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

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, 1461444818821316.

Shah, S. K. (2006). Motivation, governance, and the viability of hybrid forms in open source software development. *Management Science*, *52*(7), 1000–1014.

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

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, *3*(4), 329–336.

Shirky, C. (2008). *Here comes everybody : The power of organizing without organizations*. Penguin Press.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1–48.

Solomon, J., & Wash, R. (2014). Critical mass of what? exploring community growth in WikiProjects. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM '16)*.

Sørensen, J. B. (2004). Recruitment-based competition between industries: A community ecology. *Industrial and Corporate Change*, *13*(1), 149–170.

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

Starbird, K. (2012). Crowd computation: Organizing information during mass disruption events. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, 339–342.

Suchman, L. (1996). Supporting Articulation Work. In R. Kling (Ed.), *Computerization and Controversy: Value Conflicts and Social Choices* (pp. 407–423). Morgan Kaufmann.

Swaminathan, A. (2001). Resource partitioning and the evolution of specialist organizations: The role of location and identity in the U.S. wine industry. *Academy of Management Journal*, *44*(6), 1169–1185.

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

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

TeBlunthuis, N., & Hill, B. M. (2021, July 14). *Identifying Competition and Mutualism Between Online Groups*. arXiv: 2107.06970 [cs].

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

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

TeBlunthuis, N., Shaw, A., & Hill, B. M. (2020, June 19). *The population ecology of online collective action*.

Tripodi, F. (2021). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 14614448211023772.

Turner, F. (2005). Where the Counterculture Met the New Economy: The WELL and the Origins of Virtual Community. *Technology and Culture*, *46*(3), 485–512.

Ven, A. H. V. D., & Poole, M. S. (1995). Explaining Development and Change in Organizations. *Academy of Management Review*, *20*(3), 510–540.

Verhoef, H. A., & Morin, P. J. (2010). *Community ecology: Processes, models, and applications*. Oxford University Press
OCLC: 876676566.

Vincent, N., Johnson, I., & Hecht, B. (2018). Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 566:1–566:13.

von Hippel, E. (2016, November 18). *Free innovation* (1 edition). The MIT Press.

Waller, I., & Anderson, A. (2019). Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms. *The World Wide Web Conference on - WWW '19*, 1954–1964.

Wang, X., Butler, B. S., & Ren, Y. (2012). The impact of membership overlap on growth: An ecological competition view of online groups. *Organization Science*, *24*(2), 414–431.

Wang, Y., Niiya, M., Mark, G., Reich, S. M., & Warschauer, M. (2015). Coming of Age (Digitally): An Ecological View of Social Media Use among College Students. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 571–582.

Warncke-Wang, M., Ranjan, V., Terveen, L., & Hecht, B. (2015). Misalignment between supply and demand of quality content in peer production communities. *Proceedings of the Ninth International AAAI Conference on Web and Social Media (ICWSM '15)*, 493–502.

Weber, S. (2000, June). *The Political Economy of Open Source Software*.

White, R. W., Richardson, M., & Liu, Y. (2011, May 7). Effects of community size and contact rate in synchronous social q&a. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2837–2846). Association for Computing Machinery.

Williamson, O. E. (1981). The economics of organization: The transaction cost approach. *The American Journal of Sociology*, *87*(3), 548–577.

Worster, D. (1994). *Nature's economy: A history of ecological ideas*. Cambridge University Press OCLC: 855524849.

Xigen Li. (2011). Factors influencing the willingness to contribute information to online communities. *New Media & Society*, *13*(2), 279–296.

Zhang, J. S., Keegan, B., Lv, Q., & Tan, C. (2021). Understanding the Diverging User Trajectories in Highly-Related Online Communities During the Covid-19 Pandemic. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*, 12.

Zhang, J., Hamilton, W. L., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017). Community identity and user engagement in a multi-community landscape. *Proceedings of the International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, *2017*, 377–386.

Zhang, J., Pennebaker, J., Dumais, S., & Horvitz, E. (2020). Configuring Audiences: A Case Study of Email Communication. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 062:1–062:26.

Zhang, X. M., & Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review*, *101*(4), 1601–1615.

Zhao, X., Lampe, C., & Ellison, N. B. (2016, May 7). The Social Media Ecology: User Perceptions, Strategies and Challenges. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 89–100). Association for Computing Machinery.

Zhu, H., Chen, J., Matthews, T., Pal, A., Badenes, H., & Kraut, R. E. (2014). Selecting an effective niche: An ecological view of the success of online communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 301–310.

Zhu, H., Kraut, R. E., & Kittur, A. (2014). The impact of membership overlap on the survival of online communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 281–290.

Preface to Appendix A

The following appendix is published in the Proceedings of The 17th International Symposium on Open Collaboration.

Appendix A

Measuring Wikipedia Article Quality in One Dimension by Extending ORES with Ordinal Regression

Organizing complex peer production projects and advancing scientific knowledge of open collaboration each depend on the ability to measure quality. Article quality ratings on English language Wikipedia have been widely used by both Wikipedia community members and academic researchers for purposes like tracking knowledge gaps and studying how political polarization shapes collaboration. Even so, measuring quality presents many methodological challenges. The most widely used systems use labels on discrete ordinal scales when assessing quality, but such labels can be inconvenient for statistics and machine learning. Prior work handles this by assuming that different levels of quality are "evenly spaced" from one another. This assumption runs counter to intuitions about the relative degrees of effort needed to raise Wikipedia encyclopedia articles to different quality levels. Furthermore, models from prior work are fit to datasets that oversample high-quality articles. This limits their accuracy for representative samples of articles or revisions. I describe a technique extending the Wikimedia Foundations' ORES article quality model to address these limitations. My method uses weighted ordinal regression models to construct one-dimensional continuous measures of quality. While scores from my technique and from prior approaches are correlated, my approach improves accuracy for research datasets and provides evidence that the "evenly spaced" assumption is unfounded in practice on English Wikipedia. I conclude with recommendations for using quality scores in future research and include the full code, data, and models.

A.1. Introduction

Measuring content quality in peer production projects like Wikipedia is important so projects can learn about themselves and track progress. Measuring quality also helps build confidence that information is accurate and supports monitoring how well an encyclopedia includes diverse subject areas to identify gaps needing attention (Redi et al., 2021). Measuring quality enables tracking and evaluating the progress of subprojects and initiatives organized to fill the gaps (Halfaker, 2017; Warncke-Wang et al., 2015). Raising an article to a high standard of quality is a recognized achievement among contributors, so assessing quality can help motivate contributions (Ayers et al., 2008; Forte & Bruckman, 2005). In these ways, measuring quality can be of key importance to advancing the priorities of the Wikimedia movement and is also important

to other kinds of open collaboration (Champion & Hill, 2021).

Measuring quality also presents methodological and ontological challenges. How can "quality" be conceptualized so that measurement of the goals of a project and the value it produces can be precise and accurate? Language editions of Wikipedia, including English, peer produce quality labels that have been useful both for motivating and coordinating project work and for enabling research. Epistemic virtues of this approach stem from the community-constructed criteria for assessment and from formalized procedures for third-party evaluation organized by WikiProjects. These systems also have two important limitations: (1) ratings are likely to lag behind changes in article quality, and (2) quality is assessed on a discrete ordinal scale, which violates typical assumptions in statistical analysis. Both limitations are surmountable.

The machine learning framework introduced by Warncke-Wang et al. (2013), further developed by Halfaker (2017), implemented by the Objective Revision Evaluation Service[1] (ORES) article quality models and adopted by several research studies of Wikipedia article quality (e.g. Halfaker & Geiger, 2020; Kocielnik et al., 2018; Shi et al., 2019; Warncke-Wang et al., 2015) was designed to address the first limitation by using article assessments at the time they were made as "ground truth." Article quality might drift in the periods between assessments, but it seems safe to assume that new quality assessments are accurate at the time they are made. A model trained on recent assessments can predict what quality label an article would receive if assessed in its current state.

This paper introduces a method for constructing interpretable one-dimensional measures of article quality from Wikipedia quality assessments and the ORES article quality model. The method improves upon prior approaches in two important ways. First, by using inverse probability weighting to calibrate the model, it is more accurate for typical research applications, and second, it does not depend on the assumption that quality levels are "evenly spaced," which threatens the validity of prior research (Arazy et al., 2019; Halfaker, 2017). In addition, this paper helps us understand the validity of previous work by analyzing the performance of the ORES quality model and testing the "evenly spaced" assumption.

In §A.2, I provide a brief overview of quality measurement in peer production research, in which I foreground the importance of the assumptions needed to use machine learning predictions in downstream analysis—particularly the "evenly spaced" assumption used by Halfaker (2017) to justify the use of a handpicked weighted sum to combine article class probabilities. Next, in §A.3, I describe how to build accurate ordinal quality models that are appropriately calibrated for analyses of representative samples of Wikipedia articles or revisions. I also briefly explain how ordinal regression provides an interpretable one-dimensional measure of quality and how it relaxes the "evenly spaced" assumption. Finally, in §A.4 I present the results of my analysis to (1) show how the precision of the measurement depends on proper calibration and (2) demonstrate that the "evenly spaced" assumption is violated. Despite this, I find that scores from the ordinal models are highly correlated with those from prior work so the "evenly spaced" assumption may be acceptable in some applications. I conclude in §C.9 with recommendations for measuring article quality in future research.

[1] https://www.mediawiki.org/wiki/ORES (https://perma.cc/TH6L-KFT6)

A.2. Background

Measurement is important to science as available knowledge often constrains the development of improved tools for advancing knowledge. For example, in the book *Inventing Temperature*, Hasok Chang (Chang, 2004), the philosopher and historian of science, documents how extending theories of heat beyond the range of human sense perception required scientists to develop new types of thermometers. This in turn required better knowledge of heat and of thermometric materials such as the freezing point of mercury. Part of the challenge of scientific advancement is that measurement devices developed under certain conditions may give unexpected results outside of the range in which they are calibrated: a thermometer will give impossibly low temperature readings when its mercury unexpectedly freezes. Today, machine learning models are used to extend the range of quality measurements in peer production research, but state of the art machine learning can be quite sensitive to the nuances of how their training data are selected (Recht et al., 2019).

Measuring Quality in Peer Production

As described in §A.1, measuring quality has been of great importance to peer production projects like Wikipedia and in the construction of knowledge about how such projects work. The foundation of article quality measurement in Wikipedia has been the peer production of article quality assessment organized by WikiProjects who develop criteria for articles in their domain (Phoebe Ayers et al., 2008). This enables quality assessment to be consistent across different subject areas, but the procedures for assessing quality are tailored to the values of each WikiProject. Yet, like human sense perception of temperature, these quality assessments are limited in that they require human time and attention. In addition, humans' limited ability to discriminate between levels on a scale limits the sensitivity of quality assessments. Articles are assessed irregularly and infrequently at the discretion of volunteer editors. Therefore, for most article revisions, it is not known what quality class the article would be assigned if it were newly assessed.

Researchers have proposed many ideas to extend the range of quality measurement beyond the direct perception of Wikipedians, such as page length (Blumenstock, 2008), persistent word revisions (Adler & de Alfaro, 2007; Biancani, 2014), collaboration network structures (Raman et al., 2020), and template-based flaw detection (Anderka et al., 2012). Carefully constructed indexes benchmarked against English language Wikipedia quality assessments might allow quality measurement of articles that have not been assessed or in projects that have underproduced article assessments (Lewoniewski et al., 2017). However, such indexes may lack emic validity if they fail to capture important aspects of quality or if notions of quality vary between linguistic communities and might even shape the editing activity in unexpected ways that could ultimately defeat their purpose (Goodhart, 1984; Strathern, 1997). Peer-produced quality labels depend on the limited capacity of volunteer communities to coordinate quality assessment, but also provide impressive validity for evaluating projects on their own terms.

Article Quality Models Extend Measurement to Unassessed Articles

Perhaps the most successful approaches to extending the range of quality measurements use machine learning models trained on available article quality assessments to predict the quality

of revisions that have not been assessed. The ORES article quality model (henceforth ORES) implements this approach, but other similar article quality predictors have been developed (Anderka & Stein, 2012; Dang & Ignat, 2016; Druck et al., 2008; Raman et al., 2020; Sarkar et al., 2019; S. Zhang et al., 2018), and additional features including those based on language models can substantially improve classification performance compared to ORES (Schmidt & Zangerle, 2019). The ORES model is a tree-based classifier that predicts the quality class of a Wikipedia article at the time it is assessed.[2] These tree-based models are reasonable for practical purposes with the reported ability to predict within one level of the true quality class with 90% accuracy (although in §A.4 I find a decline in accuracy in a more recent dataset). Yet, since these models do not account for the ordering of quality labels, the use of these predictions in downstream analysis introduces complicated methodological challenges.

The ORES classifiers are fit using `scikit-learn`[3] through minimization of the multinomial deviance as shown (Hastie et al., 2018; Pedregosa et al., 2011):

$$L(y_i, p(x_i)) = -\sum_{k=1}^{K} I(y_i = \mathcal{G}_{i,k}) \log p_k(x_i) \tag{A.1}$$

For each article i with predictors x_i that has been labeled with a quality class y_i, the ORES model outputs an estimated probability $p_k(x_i)$ that the article belongs to each quality class $k \in \{$*stub*, *start*, *C-class*, *B-class*, *Good article (GA)*, *Featured article (FA)*$\}$. The predicted probabilities $p(x_i)$ sum to one so the ORES model outputs a unit vector for each article. If $\mathcal{G}_{i,k}$, the most probable quality class (MPQC) according to the model, is the true label, then $I(y_i = \mathcal{G}_{i,k})$ equals 1 (I is the indicator function) and the log predicted probability $p_k(x_i)$ of the correct class is subtracted from the loss $L(y_i, p(x_i))$. Note that this model does not use the fact that article quality classes are ordered. If it did, then it would have to penalize an incorrect classification of a *Good article* as *C-class* more than a classification of a *Good article* as *B-class*. In this model, different quality classes have no intrinsic rank or ordering and thus are akin to different categories of article subjects like animals, vegetables, or minerals.

The MPQC is perhaps the most natural way to use the ORES output to measure quality. It has been used in several studies including to provide evidence that politically polarized collaboration on Wikipedia leads to high quality articles (Shi et al., 2019) and to understand the relationship between article quality and donation (Kocielnik et al., 2018). However, the MPQC is limited in that it does not measure quality differences between articles that have the same MPQC. Consider two hypothetical articles; the first has the multinomial prediction $(0.1, 0.3, 0.4, 0.075, 0.075, 0)$ and the second has the prediction $(0.075, 0.075, 0.4, 0.3, 0.1, 0)$. The MPQC will assign both the *C-class* label even though the first article has an even chance at being a *Stub* or *Start-class* while the second article has an even chance at being a *B-class* or even a *Good article*. At best, the MPQC has limited sensitivity to subtle variations or gradual changes in quality (Halfaker, 2017).

[2] The system uses cross-validation to select among candidates that include random-forest and boosted decision tree models.

[3] https://scikit-learn.org/stable/(https://perma.cc/5Y8B-W8T5)

Combining Scores for Granular Measurement

To further extend the range of article quality measurement within article quality classes, Halfaker (2017) constructed a numerical quality score using a linear combination (a weighted sum) of the elements of the multinomial prediction $p(x_i)$. This is advantageous from a statistical perspective as it naturally provides a continuous measure of quality which can typically justify a normal or log-normal statistical model. It can also support higher-order aggregations for measuring the quality of a set of articles (Halfaker, 2017). Halfaker handpicks the coefficients $[0, 1, 2, 3, 4, 5]$ to make a linear combination of the predictions under the assumption "that the ordinal quality scale developed by Wikipedia editors is roughly cardinal and evenly spaced," which I refer to the "evenly spaced" assumption. It essentially says that a *Start-class* article has one more unit quality of a *Stub-class* article, and that a *C-class* article has one more unit of quality than a *Start-class* article and so on. This approach is being adopted by other researchers including Arazy et al. (2019).

The considerable degree of effort and expertise required to raise articles to higher levels of quality raises doubt in the assumption (Jemielniak, 2014). Higher quality levels correspond to increasing completeness, encyclopedic character, usefulness to wider audiences, incorporation of multimedia, polished citations, and adherence to Wikipedia's policies. The English language Wikipedia editing guideline on content assessment[4] defines a *Good article* as "useful to nearly all readers, with no obvious problems" and a *Featured article* article as "professional, outstanding and thorough." According to Wikipedians, it can take "three to six months of full time work" to write a *Featured article*.[5] Are we to assume that the difference in quality between a *Good article* and a *Featured article* is measurably the same as that between a *Stub* defined as as "little more than a dictionary definition" and a *Start-class* that is "a very basic description of the topic?" How could we even answer this question?

If the "evenly spaced" assumption is reasonable, then Halfaker's weighted sum approach is too. But if increasing Wikipedia article classes do not represent roughly equal improvements in quality, this may threaten the accuracy of analysis dependent on the assumption. Suppose that a *B-class* has not 1, but 2 units of quality greater than a *C-class* article, then Halfaker could have underestimated the improvement in the knowledge gap of women scientists, which was considerably driven by improvement in *B-class* articles. In the next section, I provide a straightforward extension of the ORES article quality model based on ordinal regression can both relax the "evenly spaced" assumption and provide a better calibrated and more accurate one-dimensional measure of quality.

A.3. Data, Methods and Measures

I use Bayesian ordinal regression models that use the ORES predicted probabilities to predict the quality class labels and quantify the distance between quality classes. I now provide a brief overview of ordinal regression as needed to explain my approach to measuring quality. Under-

[4] https://en.wikipedia.org/w/index.php?title=Wikipedia:Content_assessment&oldid=1023695750 (https://perma.cc/2JUV-6SD)

[5] Public statement by Stuart Yeates, an expert Wikipedian; quoted with permission. https://lists.wikimedia.org/hyperkitty/list/wiki-research-l@lists.wikimedia.org/message/7U35LHAXRWEPABN75DOTPOIEA2VYCTQQ/ (https://perma.cc/9V4P-WRXR)

standing ordinal regression depends on background knowledge of odds and generalized linear models. I recommend McElreath and Safari (2018) for reference.

Bayesian Ordinal Regression

Ordinal regression predicts quality class membership using a single linear model for all classes and identifies boundaries between classes using the log cumulative odds link function shown below in Eq. A.2. The log cumulative odds is not the only possible choice of link function, but it is the most common, is the easiest to interpret, and is appropriate here.

$$\log \frac{\Pr(y_i \leq k)}{1 - \Pr(y_i \leq k)} = \alpha_k - \phi_i \tag{A.2}$$
$$\phi_i = Bx_i$$

As in Eq. A.1, y_i is the quality label for article i. The left hand side of Eq. A.2 gives the log odds that y_i is less than or equal to quality level k. The ordinal quality measure is given by a linear model $\phi_i = Bx_i$ (x_i is a vector of transformed ORES scores for article i). Key to interpreting ϕ_i as a quality measure are the intercept parameters a_k for each quality level k. The log cumulative odds (the log odds that the article y_i has quality less than or equal to k) are given by the difference between the intercept and the linear model a_k - ϕ_i. Therefore, if $\phi_i = \alpha_k$ then the chances that $i <= k$ equal the chances that $i > k$. When ϕ_i is less than α_k, the quality of article i is probably less than or equal to quality level k. As $\phi_i - \alpha_k$ increases so do the chances that article i is of quality better than k. In this way, the threshold parameters a_k define quantitative article quality levels on the scale of the ordinal quality measure ϕ_i.

Informally, an ordinal regression model maps a linear regression model to the ordinal scale using the log cumulative odds link function. It does this by inferring thresholds that partition the range of linear predictions. When the linear predictor for an article crosses a threshold, the probability that the article has quality greater than that corresponding to the threshold begins to increase.

Bayesian inference allows interpreting model parameters like ϕ_i and α_k as random variables and provides accurate quantification of uncertainty in thresholds and predictions. I fit models using the R package Bayesian regression modeling using Stan (`brms`) (Bürkner, 2017) version 2.15.0. I use the default priors for ordinal regression, which are weakly informative. Due to the large sample size, the data overwhelm the priors and the priors have little influence over results. I confirmed this by fitting equivalent frequentist models using the `polr` function in the `MASS` R package (Venables et al., 2002) and found that the estimates of intercepts and coefficients were very close.

The six quality scores output by the ORES article quality classifier are perfectly collinear by construction because they sum to one. This means they cannot all be included in the same regression model. Since interpreting the coefficients is not important, I take the linear transformation of the ORES scores using appropriately weighted principle component analysis and use the first five principle components as the independent variables. This is simpler and more statistically efficient than a model selection procedure.

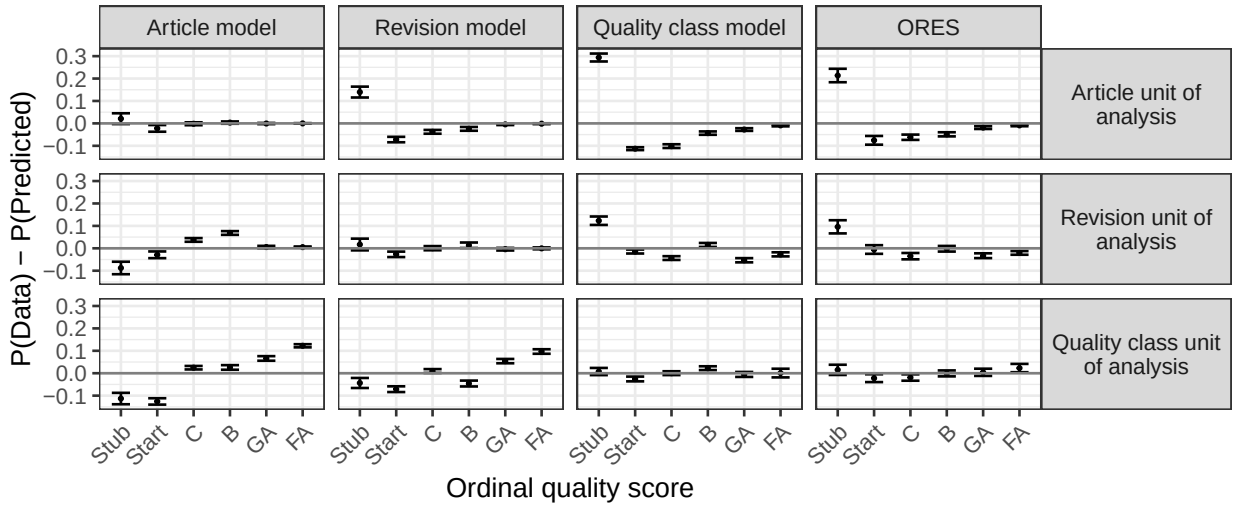


Figure A.1: Calibration of each predictive quality model on datasets representative of each unit of analysis (article, revision, quality class). Each chart shows, for each quality class, the miscalibration of a model (columns) with respect to a dataset weighted to represent a unit of analysis (rows). The y-axis shows difference between the true probability of the quality class and the average predicted probability of that class, given a chosen unit of analysis. Points close to zero indicate good calibration. For example, the top-left chart shows that the article model is well-calibrated to the dataset on which it was fit and the middle-left chart shows that the article model predicts that articles are *Stubs* with probability greater than the frequency of *Stubs* in a random sample of revisions. Error bars show 95% confidence intervals.

Dataset and Model Calibration

I draw a new random sample of 5,000 articles from each quality class to develop my models. I first reuse code from the `articlequality`[6] Python package to process the March 2020 XML dumps for English Wikipedia and extract up-to-date article quality labels. I then select pages that have been assessed by a member of at least one WikiProject. Following prior work, if an article is assessed at different levels according to more than one WikiProject, I assign it to the highest such level and I drop articles having the rarely used *A-class* quality level (Halfaker, 2017; Warncke-Wang et al., 2015; Warncke-Wang et al., 2013). Next, I use the `revscoring`[7] Python package to obtain the ORES scores of the labeled article versions. Some of these versions have been deleted leading to missing observations at each quality level. Table A.1 shows the number of articles sampled in each quality class. I reserve a random sample of 2000 articles which I use in reporting my results and fit my ordinal regression models on the remainder.

The ORES article quality classifiers are fit on a "balanced" dataset having an equal number of articles in each quality class. Thus, an ORES score is the probability that an article is a member of a quality class under the assumption that the article was drawn from a population where each quality class contains an equal number of articles. Simply put, the model has learned from its training data that each quality class is about the same size.

[6] https://pypi.org/project/articlequality (https://perma.cc/8R4H-MAZ9)
[7] https://pypi.org/project/revscoring (https://perma.cc/3HFN-V23Z)

Table A.1: Number of articles sampled at each quality level

Label	No. of articles	No. of revisions	Sample size	Article weights	Revision weights
Stub	3,359,351	12,005,611	4,969	4.23	2.52
Start	1,019,038	7,828,335	4,979	1.28	1.64
C	235,655	3,889,639	4,988	0.30	0.81
B	128,875	3,640,591	4,990	0.16	0.76
GA	31,808	924,468	4,999	0.04	0.19
FA	7,438	365,255	4,995	0.01	0.08

This is not representative of the overall article quality on Wikipedia, which is highly skewed with over 3 million *Stubs* but only around *7,000 Featured articles* as shown in Table A.1. Although using a balanced dataset likely improves the accuracy of the ORES models, for the ordinal regression models, the choice of unit of analysis presents a trade-off between accuracy in a representative sample of articles or revisions and accuracy within each quality class. Constructing a balanced dataset by oversampling is a common practice in machine learning because it can improve predictive performance. However, oversampling can also lead to badly calibrated predictive probabilities as shown in Fig. A.1. Calibration means that, on average, the predicted probability of a quality class equals the average true probability of that class for the unit of analysis.

The "balanced" dataset on which ORES is trained has the *quality class* unit of analysis because each quality class has equal representation. However, researchers are more interested in analyzing representative samples of *articles* or *revisions*. For example, the article unit of analysis would be used to estimate the average quality of a random sample of articles and the revision unit of analysis might be used to model the change in the quality of an encyclopedia over time. Weighting allows the use of the balanced dataset to estimate a model as if the dataset were a uniform random sample of a different unit of analysis. My method uses a balanced dataset to fit ordinal regression models with inverse probability weighting to calibrate each model to the unit of analysis of a research project. For example, each article in the model calibrated to the article unit of analysis is weighted by the probability of its quality class in the population of articles divided by the probability of its quality class in the sample. The size of the sample and the weights for the article and revision levels of analysis are also shown in Table A.1.

A.4. Results

I first report my findings about the spacing of the quality classes in each of the models in §A.4. Quality classes are not evenly spaced, especially when articles or revisions are the unit of analysis. Next, in §A.4, I report the accuracy of each of the models and the uncertainty of the ordinal quality scale. All models perform similarly to or better than the MPQC within the pertinent unit of analysis. The unweighted model provides the best accuracy and lowest uncertainty across the entire range of quality levels, but is poorly calibrated for other units of analysis. Finally, in §A.4, I show that all quality measures are highly correlated, but the ordinal quality measures agree with one another more than with the "evenly spaced" measure.

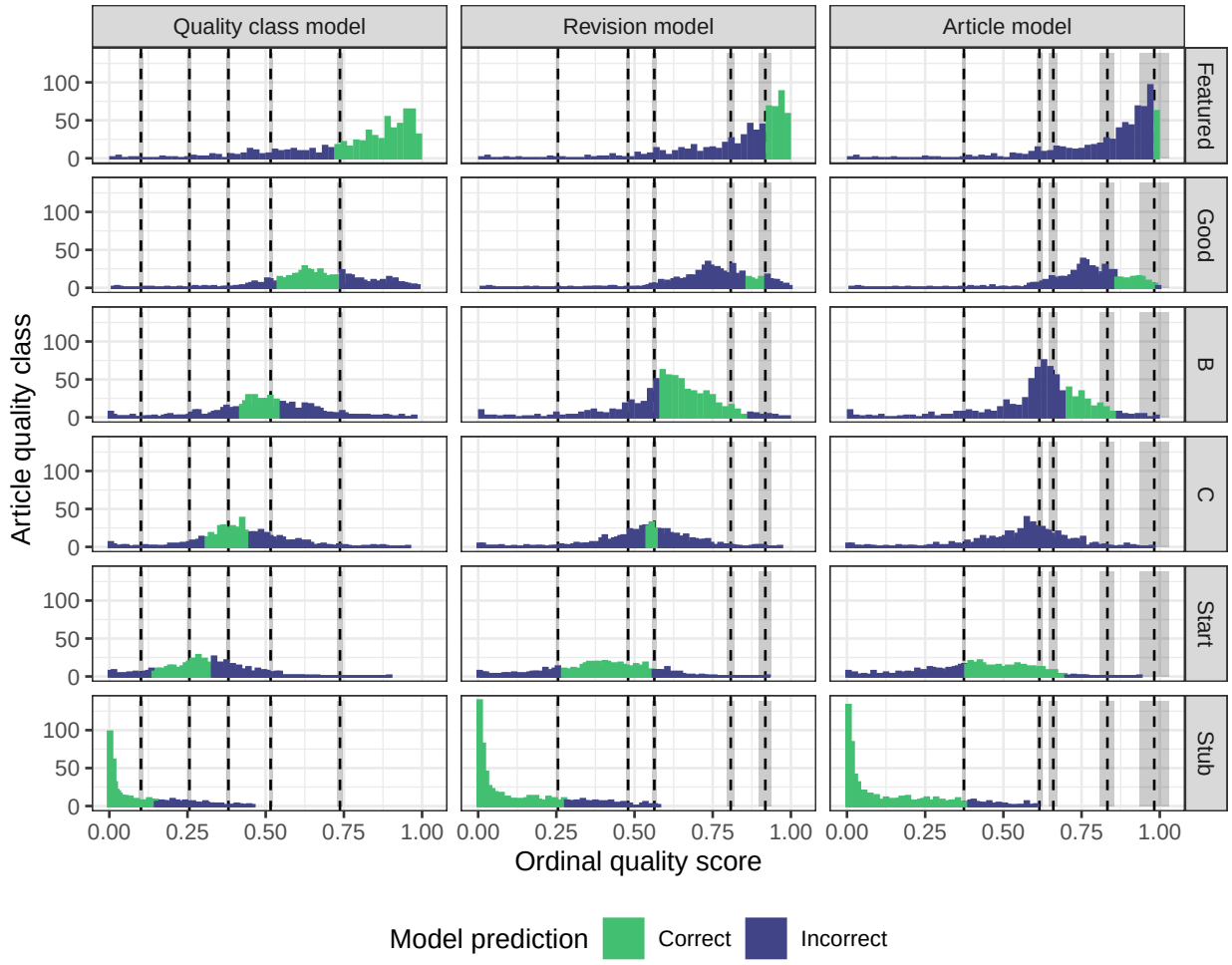


Figure A.2: Quality scores and predictions of the ordinal regression models. Columns in the grid of charts correspond to the ordinal quality model calibrated to the indicated unit of analysis and rows correspond to sampled articles having the indicated level of quality as assessed by Wikipedians. Each chart shows the histogram of scores, thresholds inferred by the ordinal model with 95% credible intervals colored in gray, and colors indicating when the model makes correct or incorrect predictions. The thresholds are not evenly spaced, especially in *revision model* and *article model* that has more weight on lower quality classes. These two models infer that the gaps between *Stub* and *Start* and between *Start* and *C-class* articles are considerably wider than the gap between *C-class* and *B-class* articles.

Spacing of Quality Classes

The grid of charts in Fig. A.2 shows quality scores and thresholds for each model (columns) and article quality level (rows). Each chart shows the histogram of quality scores ϕ_i given to articles having the true quality label corresponding to the row of the grid. The histograms are colored to indicate regions where the model correctly predicts that articles belong to their true class. Vertical dashed lines show the thresholds inferred by the model with 95% credible intervals colored in gray. Different models have different ranges of scores, so Fig. A.2 shows results normalized between 0 and 1.

No matter the unit of analysis, article quality classes are not evenly spaced. The quality class model provides a quality scale in which *Featured* articles take up 27% of the scale and are expected to score in the range of $[0.73, 1]$, but probable *C-class* articles only span 14% of the scale in the range $[0.31, 0.45]$. Researchers are likely to be interested in models calibrated to the article or revision units of analysis, and in these cases, the quality classes are far from evenly spaced. The *revision model* assigns 28% of the scale to *Stubs*, from 0 to 0.28. It assigns *C-class* articles the smallest part of the scale, only 4% of it, from 0.54 to 0.58. The *article model* is even more extreme. It assigns *Stubs* to the interval $[0, 0.39]$, 39% of the scale, and the space between thresholds defining the range of *C-class* articles is so narrow that it virtually never predicts that an article will be C-class. In general terms, the *quality class model* gives relatively equal amounts of space to each quality class compared to the other models, while reserving nearly the top half of the scale for the top 2 quality classes. The *revision model* and *article model* do the opposite and use the bottom half of the scale to account for differences within the bottom two quality classes, leave some room for *B-class* articles, but squeeze the top end of the scale and *C-class* articles into relatively small intervals.

Accuracy and Uncertainty

I evaluate predictive performance in terms of *accuracy*, the proportion of predictions of article quality that are correct. To allow comparison with the reported accuracy of the ORES quality models, I also report *off-by-one accuracy*, which includes predictions within one level of the true quality class among correct predictions.

As shown in Table A.2, the ordinal regression models have better predictive ability than the MPQC except when the unit of analysis is the quality class. In this case, the best ordinal quality model has worse accuracy than the MPQC but slightly better off-by-one accuracy. Table A.2 shows accuracy and off-by-one accuracy weighted for each unit of analysis. Accuracy for a given unit of analysis depends on having a model fit to data representative of that unit of analysis. Accuracy scores are higher when greater weight is placed on lower article quality classes, suggesting that it is easier to discriminate between these classes.

The ORES article quality model has been quickly adopted by researchers, but its accuracy is limited. While off-by-one accuracy is above 90% when the article is the unit of analysis, the MPQC only predicts the correct quality class 55% of the time when the quality class is the unit of analysis.

The trade-offs in selecting a unit of analysis on which to calibrate the models are further illustrated by Fig. A.3, which plots the size of the 95% credible intervals as a function of the quality scores for each model. As in Fig. A.2, quality scores in this plot are rescaled between 0

Table A.2: Accuracy of quality prediction models depends on the unit of analysis. The greatest accuracy and off-by-one accuracy scores are highlighted. Models are more accurate when calibrated on the same unit of analysis on which they are evaluated. Compared to the MPQC, the ordinal quality models have better accuracy when revisions or articles are the unit of analysis. When the quality class is the unit of analysis, the ordinal quality model has worse accuracy, but predicts within one quality class with slightly better accuracy.

Unit of analysis	Model	Ordinal model?	Accuracy	Off-by-one accuracy
Quality class	Article	Yes	0.33	0.75
Quality class	Revision	Yes	0.44	0.84
Quality class	Quality class	Yes	0.52	0.87
Quality class	ORES MPQC	No	0.55	0.86
Revision	Article	Yes	0.57	0.87
Revision	Revision	Yes	0.61	0.92
Revision	Quality class	Yes	0.54	0.88
Revision	ORES MPQC	No	0.58	0.9
Article	Article	Yes	0.76	0.97
Article	Revision	Yes	0.73	0.96
Article	Quality class	Yes	0.63	0.92
Article	ORES MPQC	No	0.65	0.94

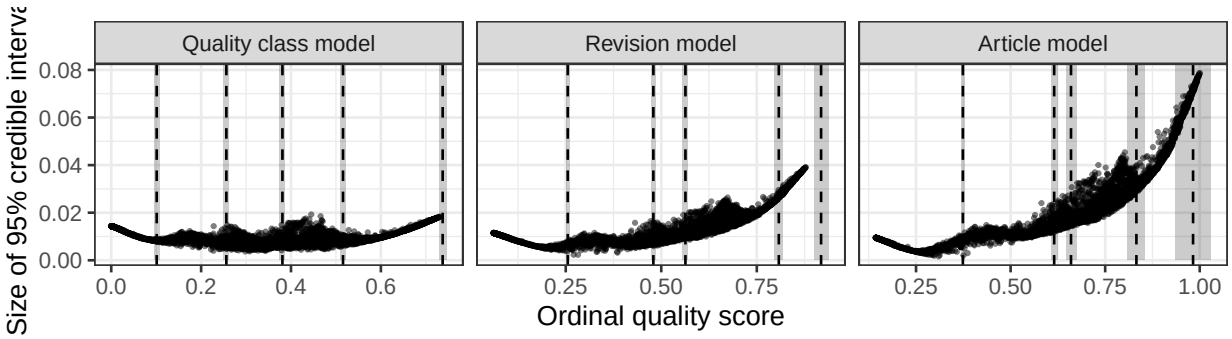


Figure A.3: Uncertainty in ordinal quality scores for models calibrated at each unit of analysis. Points show the size of the 95% credible interval for the ordinal quality score for each article in the dataset. The quality class model has low uncertainty across the range of quality. Models calibrated to the revision and article levels of analysis have less uncertainty at the low end of the quality scale, but greater uncertainty at the higher end of the scale.

and 1. The models calibrated to articles or revisions have more certainty in the lower range of the quality scale compared to the model that places equal weight in all quality classes. This comes with a trade-off for the higher range of quality. While the *quality class model* has relatively low uncertainty across the entire range of quality, the *revision model* and *article model* have greater uncertainty at higher levels of quality.

Correlation Between Scores

Although the models have different predictive performances and uncertainties, as measures of quality, they are nearly perfectly correlated with one another as shown in Fig. A.4. For each quality score, including the "evenly spaced" weighted sum, Fig. A.4 shows a scatter plot and two correlation statistics: Kendall's τ and Pearson's r. Pearson's r is the standard linear correlation coefficient and Kendall's τ is a nonparametric rank-based correlation defined as the probability that the quality scores will agree about which of any two articles has higher quality minus the probability that they will disagree.

According to Pearson's r all the quality scores are highly correlated with correlation coefficients of about 0.98 or higher. Kendall's τ measures nonlinear correlation and reveals discrepancies between the ordinal models and the "evenly spaced" measures. The Pearson correlation between scores from the *revision model* and the scores from the *quality class model* are about the same as the correlation between the *revision model* scores and the "evenly spaced" scores ($r = 0.98$). However, according to Kendall's τ, scores from the *revision model* are more similar to those from the *quality class model* ($r = 0.98$) than to the scores from the "evenly spaced" approach ($r = 0.9$).

The evenly spaced model is more likely to disagree with the model-based scores than any of the model-based scores are to disagree with one another as visualized in the scatter plots in Fig. A.4. Disagreement between the "evenly spaced" method and the ordinal models is greatest among articles in the middle of the quality range.

A.5. Discussion

Past efforts to extend the measurement of Wikipedia article quality from peer-produced article quality assessments to unassessed versions of articles and from the discrete to the continuous domain have relied upon machine learning and expedient but untested assumptions like that quality levels are "evenly spaced." While I suggest technical improvements for statistical models for measuring quality, I also find that scores from my models are highly correlated to those obtained under the "evenly spaced" assumption.

I set out to provide a better way to convert the probability vector output by the ORES article quality model into a continuous scale and to test the assumption that the quality levels are evenly spaced. I used ordinal regression models to infer spacing between quality levels and used the linear predictor of these models as a continuous measure of quality. While I found in §A.4 that the quality levels are not evenly spaced and that the spacing depends on the unit of analysis to which the models are calibrated, I also showed in §A.4 that the model-based quality measures are highly, although not perfectly, correlated with the "evenly spaced" measure. This provides some assurance that past results built on this measure are unlikely to mislead. That said, I recommend that future work adopt appropriately calibrated model-based quality measures instead of the "evenly spaced" approach, and I argue that it is important to improve the accuracy of article quality predictors to enable more precise article quality measurement.

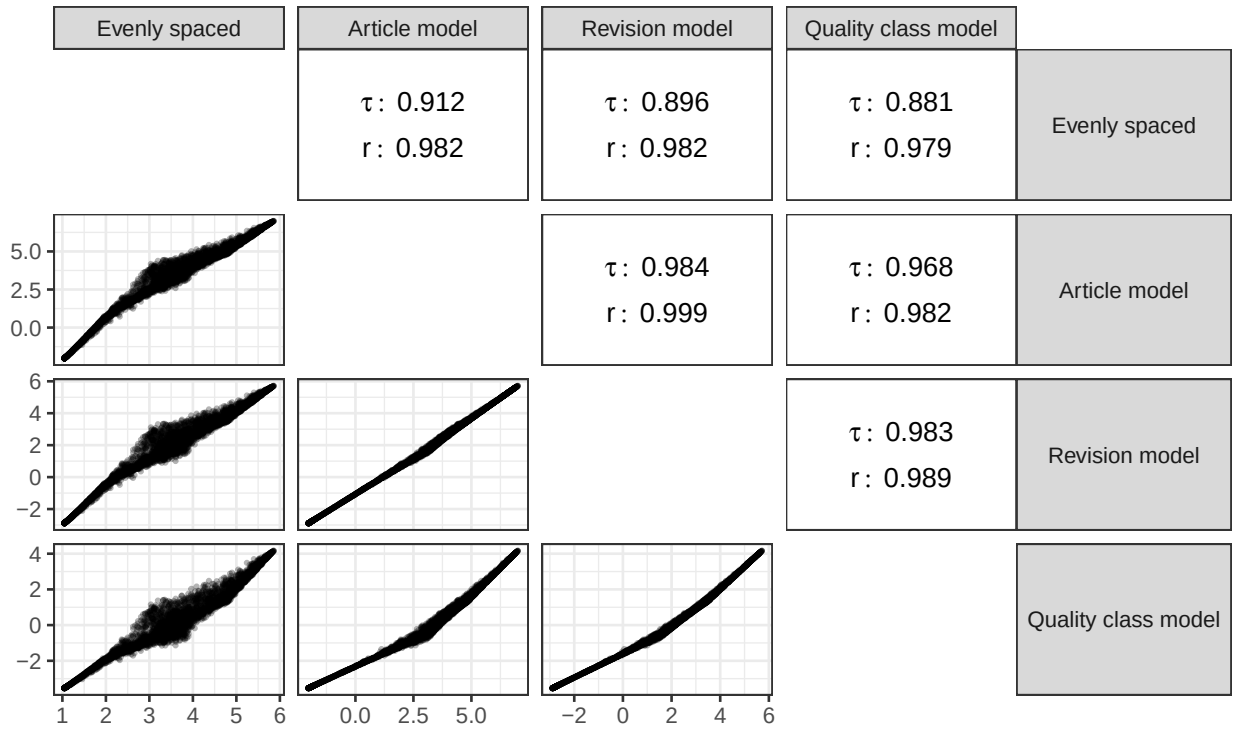


Figure A.4: Correlations between quality measures show that the different approaches to measuring quality are quite similar. "Evenly spaced" uses a weighted sum of the ORES scores with handpicked coefficients (Halfaker, 2017). Lower values of Kendall's τ, a nonparametric rank correlation statistic, compared to Pearson's r suggest nonlinear differences between the weighted sum and the other measures.

Recommendations for Measuring Article Quality

How should future researchers approach the question of how to measure Wikipedia article quality? While I cannot provide a final or complete answer to the question, I believe the exercise reported in this paper provides some insights on which to base recommendations. It is important to note that I consider here only approaches to measuring quality that assume the use of a good predictor of article quality assessment, such as the ORES quality model. I do not consider other based approaches such as those based on indexes (Lewoniewski et al., 2017) described in §A.2.

Use the principle components of ORES scores for statistical control of article quality In many statistical analyses, the only purpose of measuring quality will be as a statistical control or adjustment. For example, A. F. Zhang et al. (2017) used the MPQC as a control variable in a propensity score matching analysis of promotion to *Featured article* status, but as argued in §A.3, the MPQC provides less information than the vector of ORES scores. Using the principle components is simpler than using an ordinal quality model. I recommend obtaining ORES scores for your dataset, taking the principle components, and dropping the least significant one to remove collinearity.

Use ordinal quality scores when article quality is an independent variable In other cases, research questions will ask how article quality is related to an outcome of interest, like how Kocielnik et al. (2018) set out to explore factors associated with donations to the Wikimedia Foundation. They use the MPQC as an independent variable, which complicates their analysis. Although they conclude that "pages with higher quality attract more donations," this is not strictly true. They actually found a nonlinear relationship where readers of *B-class* articles were more likely to donate than readers of *Featured articles*. Using a continuous measure of quality is more convenient when the average linear relationship is the target of inference.

I recommend using an ordinal regression model appropriate to the downstream unit of analysis because this will justify the interpretation of the measure. If the downstream unit of analysis differs substantively from those used here, such as if different selection criteria are applied, I recommend reusing my code to calibrate a new ordinal regression model to a new dataset. Otherwise, reusing one of my models should be adequate. Finally, in the Bayesian framework, the scores are interpretable as random variables. This provides a justification for incorporating the variance of these scores as measurement errors to improve estimation in downstream analysis (McElreath & Safari, 2018).

Use the MPQC or ordinal quality scores when article quality is the dependent variable Using the MPQC as the outcome in an ordinal regression model, as is done by Shi et al. (2019) in their analysis of Wikipedia articles with politically polarized editors, is a reasonable choice as long as it provides sufficient variation and a more granular quality measure is not needed. Although it is theoretically possible that using the MPQC might introduce statistical bias because it less accurate than ordinal quality scores for units of analysis other than the quality class and omits variation within quality classes, such threats to validity do not seem more significant than the threat introduced by inaccurate predictions. If the MPQC does not provide sufficient granularity and a continuous measure is desired as in Halfaker (2017) or Arazy et al. (2019), I recommend using a measure based on ordinal regression as described in §A.5.

Limitations

Although intuitions about the varying degrees of effort required to develop articles with different levels of quality led me to question the "evenly spaced" assumption, my findings that quality classes are not evenly spaced do not necessarily reflect relative degrees of effort. Rather, spaces between levels are chosen to link a linear model to ordinal data. The spacing of intervals depends on the ability of the ORES scores to predict quality classes. The ORES article quality model has relative difficulty classifying *C-class* and *B-class* articles (Halfaker, 2017). Perhaps, the differences between these quality classes are minor compared to the other classes. Maybe ORES lacks the features or ability to model these differences and the space between these classes will grow if its predictive performance improves.

The usefulness of article quality scores depends on the accuracy of the model. The ORES quality models are accurate enough to be useful for researchers, but they still only predict the correct quality class 55% of the time on a balanced dataset. Of course, this limits the accuracy of the ordinal regression models reported here. Furthermore, while the ORES quality models were designed with carefully chosen features intended to limit biases (Halfaker & Geiger, 2020), it is still quite plausible that the accuracy of predictive quality models may vary depending

on characteristics of the article (Kleinberg et al., 2016). Such inaccuracies may introduce bias, threaten downstream analysis or lead to unanticipated consequences of collaboration tools built upon the models (TeBlunthuis et al., 2021). Therefore, improving the accuracy of article quality prediction models is important to the validity of future article quality research. Adopting machine learning models that can incorporate ordinal loss functions is a promising direction and can reduce the need for auxiliary ordinal regression models (Cardoso et al., 2007).

This paper only considers measuring article quality for English language Wikipedia, but expanding knowledge of collaborative encyclopedia production depends on studying other languages as audiences and collaborative dynamics can greatly vary between projects (Hecht & Gergle, 2010; Lemmerich et al., 2019; TeBlunthuis et al., 2019). Other languages carry out quality assessments (Lewoniewski et al., 2017), and some of these have been used to build ORES article quality models. Future work should extend this project to provide multilingual article quality measures in one continuous dimension.

An additional limitation stems from the likelihood that peer-produced quality labels are biased. For instance, the English Wikipedia community has a well-documented pattern of discrimination against content associated with marginalized groups such as biographies of women (Menking et al., 2019; Tripodi, 2021) and indigenous knowledge (van der Velden, 2013). Although demonstrating biases in article quality assessment is a task for future research, if Wikipedians' assessments of article quality are biased then model predictions of quality will almost certainly be as well.

A.6. Conclusion

Measuring article quality in one continuous dimension is a valuable tool for studying the peer production of information goods because it provides granularity and is amenable to statistical analysis. Prior approaches extended ORES article quality prediction into a continuous measure under the "evenly spaced" assumption. I showed how to use ordinal regression models to transform the ORES predictions into a continuous measure of quality that is interpretable as a probability distribution over article quality levels, provides an account of its own uncertainty and does not assume that quality levels are "evenly spaced." Calibrating the models to the chosen unit of analysis improves accuracy for research applications. I recommend that future work adopt this approach when article quality is an independent variable in a statistical analysis.

A.7. Code and Data Availability

Code, data and instructions for replicating or reusing this analysis are available in the Harvard Dataverse at https://doi.org/10.7910/DVN/U5V0G1.

Acknowledgements

I am grateful to the members of the Community Data Science Collective for their feedback on early drafts of this work including Kaylea Champion, Sneha Narayan, Jeremy Foote, and Benjamin Mako Hill. I would also like to thank Aaron Halfaker for encouraging me to write this after seeing a preliminary version. Thanks to Stuart Yeates and other participants in the

wiki-research-l mailing list wiki-research-l@lists.wikimedia.org for answering my questions about measuring article quality and effort. Finally, thank you to the anonymous OpenSym reviewers whose careful and constructive feedback improved the paper.

References

Adler, B. T., & de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. *Proceedings of the 16th International Conference on World Wide Web*, 261–270.

Anderka, M., & Stein, B. (2012). A breakdown of quality flaws in Wikipedia. *Proceedings of the 2Nd Joint WICOW/AIRWeb Workshop on Web Quality*, 11–18.

Anderka, M., Stein, B., & Lipka, N. (2012). Predicting quality flaws in user-generated content: The case of Wikipedia. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 981–990.

Arazy, O., Lindberg, A., Rezaei, M., & Samorani, M. (2019). The evolutionary trajectories of peer-produced artifacts: Group composition, the trajectories' exploration, and the quality of artifacts. *MIS Quarterly*.

Ayers, P., Matthews, C., & Yates, B. (2008). *How Wikipedia works and how you can be a part of it*. No Starch Press.

Biancani, S. (2014). Measuring the Quality of Edits to Wikipedia. *Proceedings of The International Symposium on Open Collaboration*, 33:1–33:3.

Blumenstock, J. E. (2008). Size matters: Word count as a measure of quality on wikipedia. *Proceeding of the 17th International Conference on World Wide Web - WWW '08*, 1095.

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Cardoso, J. S., Cardoso, J., & Pt, I. (2007). Learning to Classify Ordinal Data: The Data Replication Method. *Journal of Machine Learning Research*, *8*, 37.

Champion, K., & Hill, B. M. (2021). Underproduction: An approach for measuring risk in open source software. *IEEE International Conference on Software Analysis, Evolution and Reengineering*.

Chang, H. (2004). *Inventing temperature.* OUP
OCLC: 538097673.

Dang, Q. V., & Ignat, C.-L. (2016). Quality Assessment of Wikipedia Articles Without Feature Engineering. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 27–30.

Druck, G., Miklau, G., & McCallum, A. (2008). Learning to Predict the Quality of Contributions to Wikipedia. *WikiAI*, 6.

Forte, A., & Bruckman, A. (2005). Why Do People Write for Wikipedia? incentives to Contribute to Open-Content Publishing. *Proceedings of GROUP*, 6.

Goodhart, C. A. E. (1984). Problems of Monetary Management: The UK Experience. In C. A. E. Goodhart (Ed.), *Monetary Theory and Practice: The UK Experience* (pp. 91–121). Macmillan Education UK.

Halfaker, A. (2017). Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect. *Proceedings of the 13th International Symposium on Open Collaboration*, 1–9.

Halfaker, A., & Geiger, R. S. (2020). ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *4*(148), 37.

Hastie, T., Friedman, J., & Tisbshirani, R. (2018). *The Elements of statistical learning: Data mining, inference, and prediction*. Springer
OCLC: 1085863671.

Hecht, B., & Gergle, D. (2010). The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 291–300.

Jemielniak, D. (2014). *Common Knowledge?: An Ethnography of Wikipedia*. Stanford University Press.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*.

Kocielnik, R., Keyes, O., Morgan, J. T., Taraborelli, D., McDonald, D. W., & Hsieh, G. (2018). Reciprocity and Donation: How Article Topic, Quality and Dwell Time Predict Banner Donation on Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–20.

Lemmerich, F., Sáez-Trumper, D., West, R., & Zia, L. (2019). Why the World Reads Wikipedia: Beyond English Speakers. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 618–626.

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017). Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics*, *4*(4), 43.

McElreath, R., & Safari, a. O. M. C. (2018). *Statistical Rethinking*
OCLC: 1107423386.

Menking, A., Erickson, I., & Pratt, W. (2019). People Who Can Take It: How Women Wikipedians Negotiate and Navigate Safety. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Phoebe Ayers, Charles Matthews, & Ben Yates. (2008). *How Wikipedia Works*. No Starch Press.

Raman, N., Sauerberg, N., Fisher, J., & Narayan, S. (2020). Classifying Wikipedia Article Quality With Revision History Networks. *Proceedings of the 16th International Symposium on Open Collaboration*, 1–7.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet? *International Conference on Machine Learning*, 5389–5400.

Redi, M., Gerlach, M., Johnson, I., Morgan, J., & Zia, L. (2021). A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft). *arXiv:2008.12314 [cs]*.

Sarkar, S., Reddy, B. P., Sikdar, S., & Mukherjee, A. (2019). StRE: Self Attentive Edit Quality Prediction in Wikipedia. *arXiv:1906.04678 [cs]*.

Schmidt, M., & Zangerle, E. (2019). Article quality classification on Wikipedia: Introducing document embeddings and content features. *Proceedings of the 15th International Symposium on Open Collaboration*, 1–8.

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, *3*(4), 329–336.

Strathern, M. (1997). 'Improving ratings': Audit in the British University system. *European Review*, *5*(3), 305–321.

TeBlunthuis, N., Bayer, T., & Vasileva, O. (2019). Dwelling on Wikipedia: Investigating time spent by global encyclopedia readers. *OpenSym '19, The 15th International Symposium on Open Collaboration*, 14.

TeBlunthuis, N., Hill, B. M., & Halfaker, A. (2021). Effects of Algorithmic Flagging on Fairness: Quasi-experimental Evidence from Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 56:1–56:27.

Tripodi, F. (2021). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 14614448211023772.

van der Velden, M. (2013). Decentering Design: Wikipedia and Indigenous Knowledge. *International Journal of Human–Computer Interaction*, *29*(4), 308–316
_eprint: https://doi.org/10.1080/10447318.2013.765768.

Venables, W. N., Ripley, B. D., & Venables, W. N. (2002). *Modern applied statistics with S*. Springer
OCLC: 1058013209.

Warncke-Wang, M., Ayukaev, V. R., Hecht, B., & Terveen, L. G. (2015). The Success and Failure of Quality Improvement Projects in Peer Production Communities. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 743–756.

Warncke-Wang, M., Cosley, D., & Riedl, J. (2013). Tell me more: An actionable quality model for Wikipedia. *Proceedings of the 9th International Symposium on Open Collaboration*, 1–10.

Zhang, A. F., Livneh, D., Budak, C., Robert, L. P., & Romero, D. M. (2017). Crowd Development: The Interplay between Crowd Evaluation and Collaborative Dynamics in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–21.

Zhang, S., Hu, Z., Zhang, C., & Yu, K. (2018). History-Based Article Quality Assessment on Wikipedia. *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 1–8.

Preface to Appendix B

The following appendix is a collaborative work with Tilman Bayer and Olga Vasileva and is published in the Proceedings of The 15th International Symposium on Open Collaboration.

Appendix B

Dwelling on Wikipedia: Investigating time spent by global encyclopedia readers

Much existing knowledge about global consumption of peer-produced information goods is supported by data on Wikipedia page view counts and surveys. In 2017, the Wikimedia Foundation began measuring the time readers spend on a given page view (dwell time), enabling a more detailed understanding of such reading patterns. In this paper, we validate and model this new data source and, building on existing findings, use regression analysis to test hypotheses about how patterns in reading time vary between global contexts. Consistent with prior findings from self-report data, our complementary analysis of behavioral data provides evidence that Global South readers are more likely to use Wikipedia to gain in-depth understanding of a topic. We find that Global South readers spend more time per page view and that this difference is amplified on desktop devices, which are thought to be better suited for in-depth information seeking tasks.

B.1. Introduction

How do Wikipedia readers vary across different geographic and developmental contexts? A recent study of readers of different Wikipedia language editions found that readers in countries with a lower human development index (HDI) were more likely to read for in-depth understanding compared to readers in high-HDI countries (Lemmerich et al., 2019). However, this study is limited by the use of self-reported data, which can be biased by effects of social desirability and self-selection due to the volunteer nature of web-based surveys (Antin & Shaw, 2012; Hill & Shaw, 2013; Kiesler & Sproull, 1986; Phillips & Clancy, 1972). This study provides additional support for this finding from large-scale observation of reading behavior across contexts with varying levels of development.

Wikipedia contributors generally start as Wikipedia readers Therefore understanding and better supporting readership is important for the continued growth of the Wikimedia movement (Preece & Shneiderman, 2009). In 2017, the Wikimedia Foundation's web team introduced new instrumentation to measure the amount of time Wikipedia readers spend on the pages they view. We utilize this newly available data source, which provides additional information over the widely used page view data. With reading times in our field of view, it becomes clear that not all views are created equal. Some page views seem to involve in-depth reading, yet most are quite short.

We begin our analysis by evaluating the quality of the adopted approach for measuring reading times. We find limitations including a high rate of missing data on mobile devices and a low

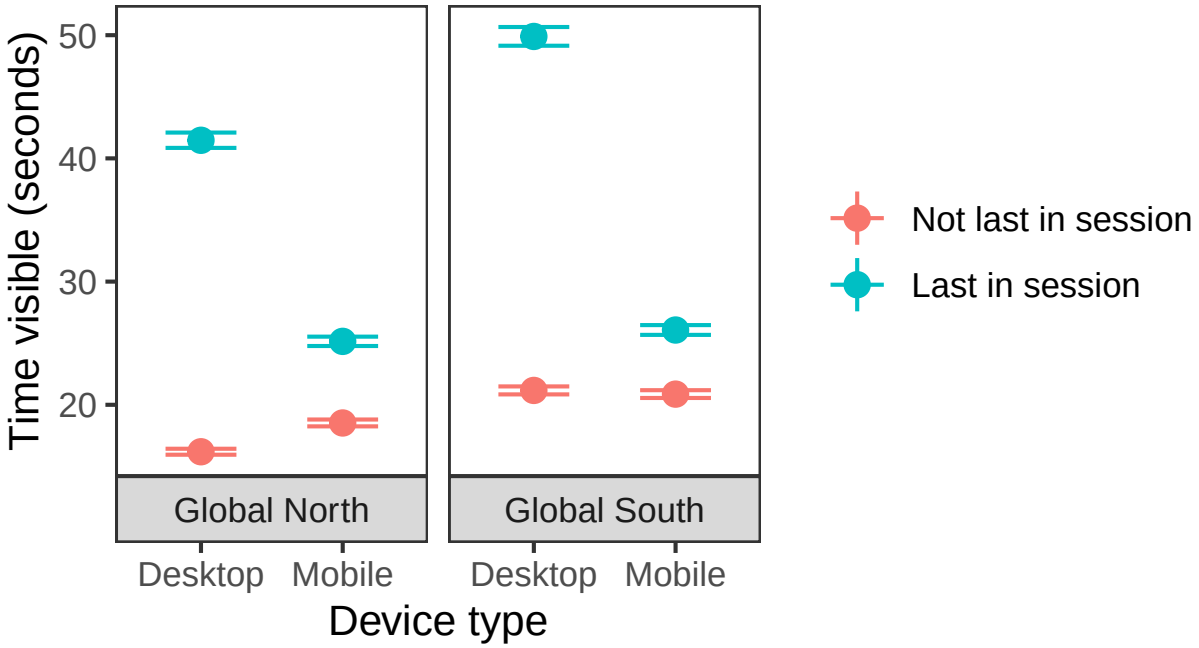


Figure B.1: Marginal effects plot showing dwell times on Wikipedia pages predicted by our regression model. Compared to readers in the Global North, readers in the Global South spend substantially more time reading when on desktop devices.

rate of invalid (missing or negative) measurements. However, we believe that the data can be generally informative as long as these limitations are considered. We then present a summary of the data and estimate the total time spent reading Wikipedia.

Next we evaluate probability models for reading time data. In addition to validating assumptions that underlie the use of parametric statistics and regression models used in answering our research questions, model selection can also help evaluate theorized data generating processes that predict when a given model will be a good fit for the data (Mitzenmacher, 2004; Stumpf & Porter, 2012). For instance, Liu et al. (2010) analyze dwell times using Weibull models, finding evidence for "screen-and-glean" patterns in which people first spend a short amount of time to assess a web page, and then decide whether to read it in-depth (Liu et al., 2010). We evaluate several probability distributions on the data from Wikipedia readers, and find that the Weibull model is not a good fit, but that the log-normal distribution fits the data well enough to justify using the geometric mean as a metric.

Finally, we return to our study of global reading behavior. Consistent with the results of Lemmerich et al., we find that readers in countries with lower HDI or in the so-called Global South spend more time reading per page view compared to readers in the Global North or in countries with higher HDI (Lemmerich et al., 2019). Moreover, this difference is amplified where we would expect users to consume information in depth: on the desktop (non-mobile) site. While we also hypothesized that the difference would likewise be greater in the last page-

view in a session, this idea was not supported by our data analysis. We demonstrate these patterns using both multivariate regressions and a simple non-parametric analysis.

B.2. Background

Wikipedia readership

Reading behavior on Wikipedia has been studied extensively, with a 2014 literature review listing 99 publications by 2011 (Okoli et al., 2014). Page view count data is central to this body of work when it comes to quantifying the attention readers give to particular topics or entire Wikipedia language editions. According to Priedhorsky et al., "the most common application [of page view data] is detection and measurement of popular news topics or events," with other uses including forecasting attempts (of e.g. box office revenues) and the study of Wikipedia's own processes (Priedhorsky et al., 2017). As an example of research using it to examine information imbalances, building on earlier work by Gorbatâi and others, Warncke-Wang et al. compared page view data with article quality ratings, and found "misalignment between supply and demand", as the Wikipedia articles with the most views were often not the highest quality (Gorbatai, 2011; Warncke-Wang et al., 2015). Other, less frequently used research strategies include using click streams and session lengths (Halfaker et al., 2015; Paranjape et al., 2016).

Surveys are another important source of information about Wikipedia readership (Okoli et al., 2014). As mentioned in the introduction, such voluntarily self-reported data are subject to participation and social desirability biases. Participation biases from self-selection may have had significant effects in the case of a previous Wikipedia reader and editor survey (Hill & Shaw, 2013).

Some previous research on Wikipedia readership has already used an approximation of reading time that assumes that the end of a page view is always marked by a new web-request originating from the same IP and user agent (Singer et al., 2017). Apart from the limitations arising from using the IP/user agent combination as a substitute for a user ID, this approach also does not allow measuring the dwell time for the last page view in a session.

Dwell times and information seeking

It has long been observed that page view numbers can paint a misleading picture of the amount of attention spent by web readers, or the information value a web site provides to them. An early study of search engine users found in 2003 that typical reading times were "substantially less than has been previously reported using survey data" (Jansen & Spink, 2003). In more recent years, metrics based on page dwell time (or total time spent on a site) have been adopted more widely. A prominent example is the online publishing platform Medium.com, which in 2013 declared "Total Time Reading" (TTR) as their "Only Metric That Matters." Distancing themselves from widely adopted web analytics metrics such as page views or active users, they argue that the act of reading should be seen as the most relevant form of user engagement for content websites (Davies, 2013).

Much prior work on web page dwell times focuses on applications in information retrieval and content recommendation (e.g., Kim et al., 2014; Yi et al., 2014; Yin et al., 2013). Long dwell times can signal successful information retrieval in search applications because they suggest that

the user has found sought information (Kim et al., 2014). Liu et al. analyzed dwell time data collected through a web browser plugin to characterize types of web content (Liu et al., 2010). However, factors beyond content may influence dwell times including psychological processes of decision making and individualized styles of content consumption (Yin et al., 2013). As we compare Wikipedia readers using mobile and desktop devices it is worth noting that dwell times are likely to be longer on desktop computers compared to mobile devices (Yi et al., 2014).

Global device and knowledge gaps

We seek to understand differences in Wikipedia's audience between the areas roughly known as the Global North and the Global South. Lemmerich et al. show empirical differences between self-reported information seeking behavior between such contexts (Lemmerich et al., 2019). These differences are likely related to digital divides or gaps between the knowledge, information and technology resources commonly available in different contexts, which can lead to systematic differences in reading behavior.

For people to use the Internet (or Wikipedia), they have to be able to connect to it, but not all forms of access are equally suited for a given task (van Deursen & van Dijk, 2015). Deursen et al. suggest that personal computers will be better for in-depth information seeking, while mobile devices, which are often close at hand, have advantages for social interaction (van Deursen & van Dijk, 2015). As Internet access becomes more ubiquitous, gaps in skills and knowledge about how to use the Internet are increasingly salient digital dividers and can be reinforced by device gaps (Deursen et al., 2017; Hargittai, 2002). For instance, in many parts of the non-western world, mobile phones diffused before PCs, and skills for PC usage may be less widespread (Napoli & Obar, 2014; Pearce & Rice, 2013). We contribute new information about the interaction between device use around the world and how people read Wikipedia.

Gaps in skills and knowledge may also help explain gaps in who contributes to Wikipedia (Shaw & Hargittai, 2018). Wikipedia promises to advance over traditional modes of knowledge production in which dominant western attitudes shape what people and places will be included and how they will be represented in authoritative sources like encyclopedias (Graham et al., 2014). In theory, peer production can empower people around to the world to add their local knowledge of their places to Wikipedia. Yet even as global access to Wikipedia grows, it is slow to fulfill these promises. Gaps in coverage of cultural knowledge reflect and reinforce structural digital divides at many levels that "disadvantage many of the world's informational peripheries" (Graham et al., 2014). These gaps in Wikipedia's coverage help motivate a better understanding of global readership.

In this paper, we use the Human Development Index (HDI) and the Global South/Global North regional classification as means of comparing countries separated by varying levels of development. We recognize that both are insufficient for defining economic development. Furthermore, these concepts and our measures of them only provide an incomplete understanding of the unique identities and motivations of cultures within an information-seeking context. What's more, they do not take into consideration inequality within a geographic region due to minority populations, which may affect the utility of averages such as GDP, income, and life expectancy. We hope that this work provides a basis of study that may be continued with work that takes into account individual cultural context, internet accessibility, and internal inequality.

B.3. Methods

Collecting reading time data

Our data collection instrument, the reading depth plugin uses the page visibility API to measure *time visible*, the total amount of time that the page was in a visible browser tab.[1] The instrument also records a second candidate measure of reading time: *total time*. This is simply the entire time the page was loaded in the browser. We used this variable for data validation and in robustness checks. We chose to focus on *time visible* because it excludes time when the user could not possibly have been reading the page. This is similar to the client-side approach described in Yi et al. (2014) (Yi et al., 2014).

Beginning November 20th 2017, we logged events from a 0.1% sample of visitor sessions.[2] The sampling rate was increased to 10% on September 25, 2018 to support future studies at a higher level of granularity.

Since we care about the reading behavior of humans, we identify bots using user agent strings and exclude them from all of our analyses.[3]

Missing data

We are only able to collect data from web browsers that support the APIs on which the instrument depends. Also, we excluded certain user agents that were found to send data unreliably in our testing, namely the default Android browser, versions of Chrome earlier than 39, Safari, and all browsers running on versions of iOS older than 11.3. We also do not collect data from browsers that have not enabled JavaScript or that have enabled Do Not Track.[4]

Even when the above conditions are met, in some cases we are still not able to collect data. Sometimes we observe a page loaded event, indicating that a user in our sample opened a page, but we do not observe a corresponding event indicating that the user has left the page (a page unloaded event). This issue affects 57% percent of records on the mobile site and about 5% of records on the desktop site. The likely explanation for why many mobile views are affected is that many mobile browsers will fail to send a page-unloaded event in certain situations, such as when the user closes the browser app using the app switcher.[5] We only include page views for which we observe exactly 1 page loaded event and 1 page unloaded event and remove 0.016% of page unloaded events where, for unknown reasons, the instrument recorded a page visible time that was less than 0 or undefined.

[1] See https://meta.wikimedia.org/wiki/Schema:ReadingDepth *archived at* https://perma.cc/JK75-Y6DH and https://developer.mozilla.org/en-US/docs/Web/API/Page_Visibility_API *archived at* https://perma.cc/79PB-389J

[2] Sessions are based on a random identifier recorded in the browser's *sessionStorage*, which expires at the end of each browser session. This is more privacy-friendly than the common approach (as used in e.g. Google Analytics) of tracking users via a cookie, in that the session identifier is not sent with every request to Wikimedia servers. It also differs from session cookies in that a new identifier will be used for links opened in a new browser tab or window.

[3] See https://meta.wikimedia.org/wiki/Research:Page_view/Tags#Spider *archived at* https://perma.cc/3NSL-X6L2

[4] See https://en.wikipedia.org/wiki/Do_Not_Track *archived at* https://perma.cc/J368-ZYBD

[5] We are planning to remedy this issue in future versions of the instrumentation, by making use of alternatives to the page unloaded event available in modern browsers, e.g. the Page Lifecycle API introduced in Google Chrome in 2018.

Taking a sample

Because Wikipedia is so widely read, even a 0.1% sample results in an amount of data exceeding the statistical requirements of this analysis. We therefore conduct our analysis on random subsamples of the collected data.

To ensure that all Wikipedia language projects are fairly and adequately represented in our sample, we use stratified sampling by assigning a *weight* to each group that adjusts the probability that members of the group are chosen in the sample. This introduces a *known bias* in the resulting sample, which is corrected using the *weights* in ways analogous to weighted averaging. For estimating total reading time, and for distribution selection, we stratify by wiki, taking up to 20,000 data points for each wiki and excluding wikis that have fewer than 300 data points. This leaves us with 242 wikis in our sample. In the multivariate analysis below, we stratify by wiki, by the country of the reader's approximate location, and by whether or not we think that the user is on a mobile device. We sample up to 200 data points for each stratum and analyze a sample of 285 wikis.

Ethical considerations

Our approach in this paper relies on large-scale observational data collected by monitoring the behavior of Wikipedia visitors. We neither see nor speak to the humans on the other side of the screen. In addition to the empirical limitations discussed below, this approach is subject to epistemic limitations. It makes those behaviors that we can observe through browser APIs visible, while obscuring those we cannot. It cannot speak to how people in different countries understand their experience of Wikipedia (Graham & Shelton, 2013). Furthermore, "big data" approaches carry critical and novel ethical risks that are not easily understood in conventional informed-consent and human subjects research frameworks (boyd & Crawford, 2012).

Wikimedia's privacy policy endeavors to clearly communicate that the information we use here will be collected, but we do not consider this an ethical license to use this data however we see fit.[6] We chose an analysis that we believe poses minimal risk to Wikipedia visitors' expectations, trust, and autonomy (Fiesler & Proferes, 2018).[7] Each observation of individuals in our study was aggregated with many others at a high level of granularity. We chose to study the country level partly because our geolocation measure is most accurate at that level, but also because it is very coarse. We do not track people from one session to another, and do not look at the content of the pages they visit other than the page length. We exclude people from our analysis who indicate a wish for privacy by enabling Do Not Track in their browsers, and will discard any session identifiers remaining in the data collected for this analysis after it is complete.

B.4. Distribution of reading times

Here we present summary statistics and a high level description of reading behavior on Wikipedia in terms of dwell times. When someone opens a given page on Wikipedia, how long do they typically stay on the page? Are reading times highly skewed? How much does reading behavior

[6] See https://foundation.wikimedia.org/wiki/Privacy_policy *archived at* https://perma.cc/C4VQ-HWRT

[7] We followed the Wikimedia Foundation's (WMF) guidelines and processes for conducting research. As it is not a federally funded institution, research at the WMF is not supervised by an institutional review board (IRB).

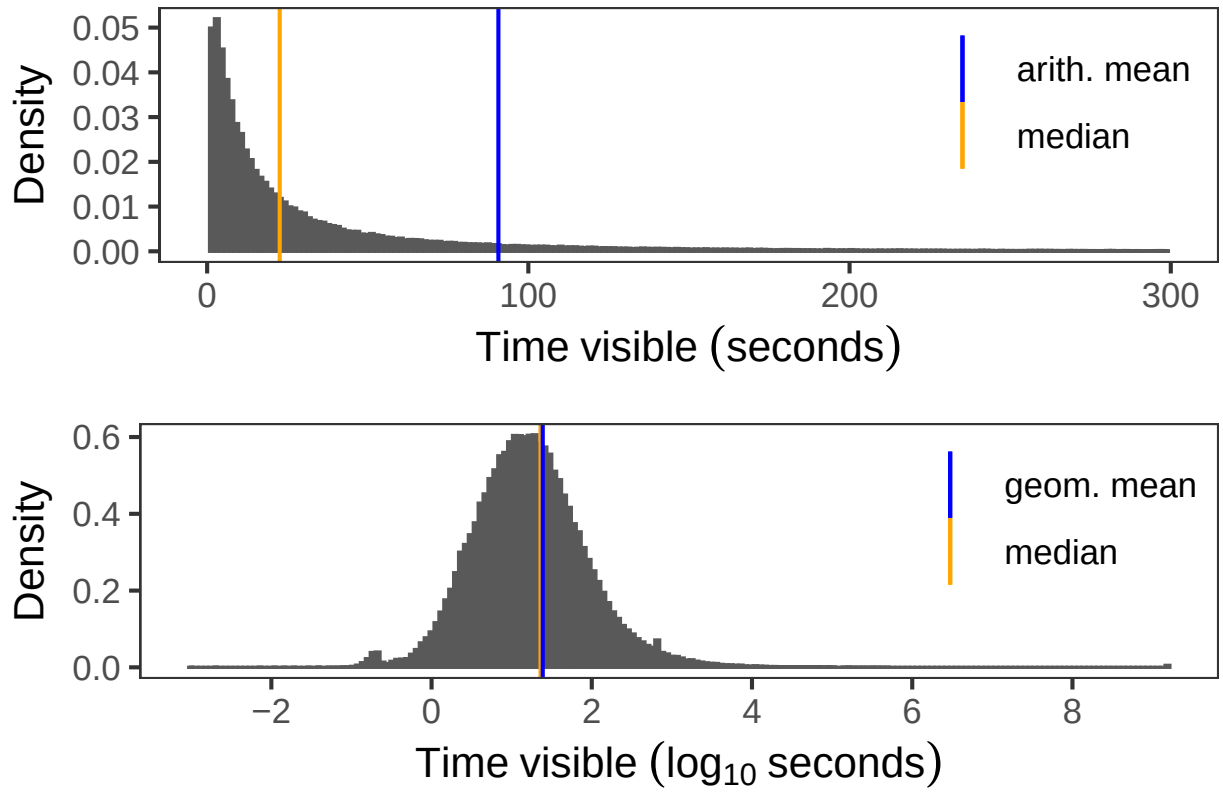


Figure B.2: The distribution of dwell times across 242 language editions of Wikipedia. The top chart shows a histogram of dwell times less than one hour long (the x-axis is truncated to 300 seconds for clarity). In this chart we can see that the median dwell time is about 25 seconds long and that the distribution of dwell times is very skewed, with the arithmetic mean far from the median. The y-axis represents the probability that a given page view is in a given box. In the lower figure, the dwell times are log-transformed and the data appear bell-shaped, with some skew to the right.

vary across different language editions of Wikipedia? How much time does all of humanity spend reading Wikipedia?

Wikipedia as a whole

In general, the distribution of reading times is very skewed (see Figure B.2). The median reading time is 25 seconds and the 75th percentile is 75.1 seconds. This skewness pushes the arithmetic mean far from most of the mass of the distribution. Therefore, the geometric means, medians, and other percentiles have more utility within our discussion of reading times.

Total time spent

Based on our data, we estimate that humanity spent about *672,349 years* reading Wikipedia from November 2017 through October 2018. We calculated this estimate as the product of

wiki	5%	25%	50%	75%	95%
all wikis	1.8	8.0	25.0	75.1	439.1
ar	5.2	5.2	21.5	69.9	371.7
de	14.1	14.1	14.1	56.6	482.7
en	37.2	37.2	37.2	37.2	262.4
es	23.3	23.3	23.3	65.5	616.4
hi	2.5	11.4	31.4	82.6	360.5
nl	6.1	6.1	15.9	60.1	441.8
pa	2.0	7.2	19.5	55.4	303.1

Table B.1: Percentiles for reading times (in seconds) on selected Wikipedia editions

the mean reading time on each Wikipedia wiki by the number of page views on that wiki, excluding readers using the mobile apps and identified bots. It is possible that some people leave Wikipedia pages visible in their browsers for extended periods of time without reading. To make our estimates of total reading time in this section somewhat conservative, we rounded all page views down to 1 hour.

Variation between different language editions

Figure B.3 shows kernel density estimates of the distribution of page visible times on several Wikipedia language editions selected to highlight projects of different sizes and of different cultures. These are Arabic (ar), German (de), English (en), Spanish (es), Hindi (hi), Dutch (nl) and Punjabi (pa). As above, we place unscaled data side-by-side with log-transformed data. Only the log-transformed plots show the full range of the data. Similar kernel density plots for other languages as well as box-and-whisker plots are available in our online supplement.[8]

B.5. Univariate model selection

Motivation

Analysts of reading times on Wikipedia will wish to make parametric assumptions to justify the use of statistical models for evaluating experiments, drawing comparisons between different samples of reading times, and performing multivariate analyses as we do below. This requires assuming a probability distribution with interpretable parameters such as mean, variance, and shape parameters. Fitting parametric distributions to data allows us to estimate these parameters and to statistically test changes in the parameters. However, parametric models can mislead if they don't fit the data well. Below, we evaluate several models.

Candidate models

We consider the following distributions in our model-selection process.

[8] Available at https://w.wiki/5Jo.

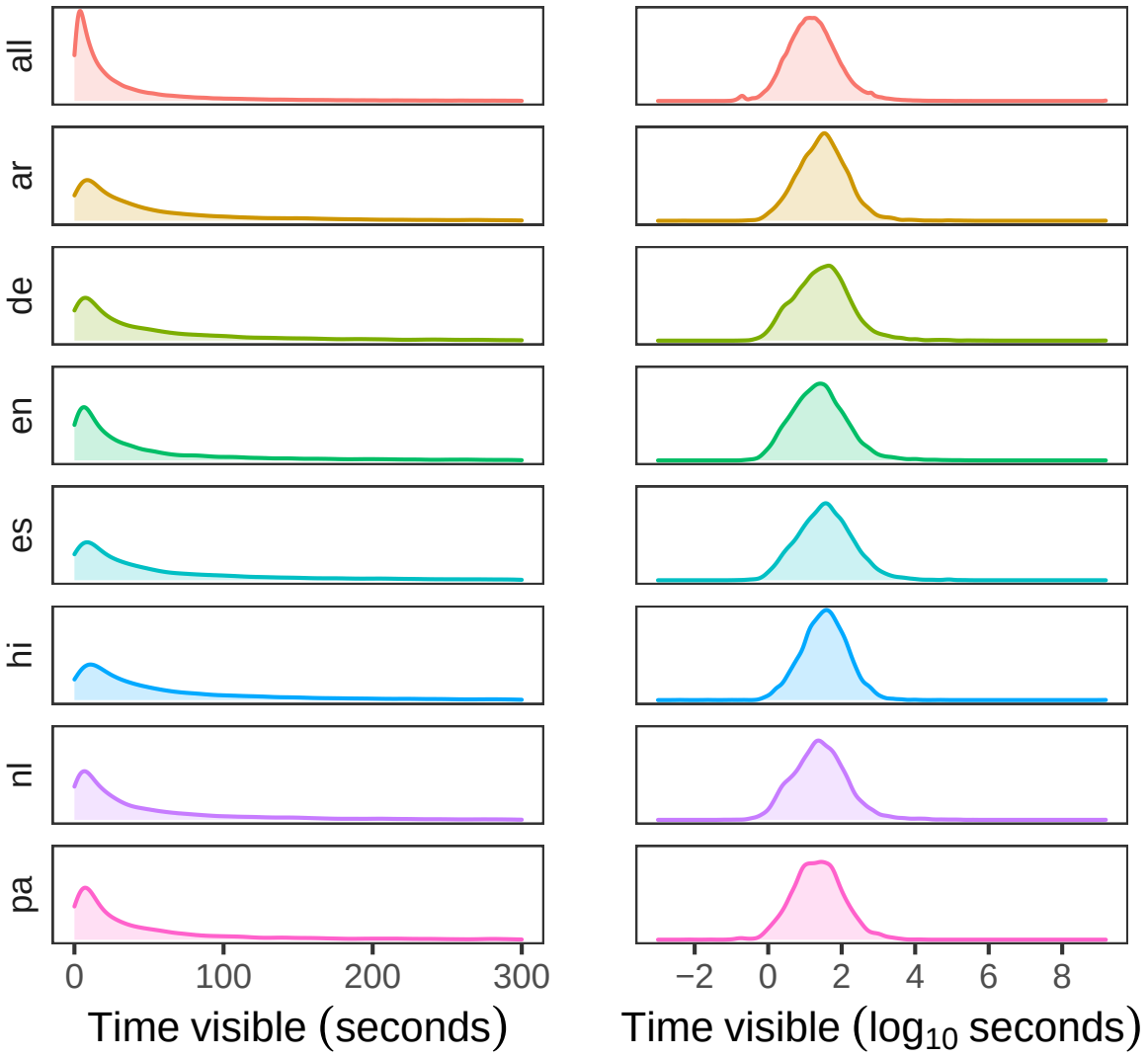


Figure B.3: Kernel density plots of the distribution of dwell times on a selection of wikis. Spanish, Hindi, and Arabic appear to have longer reading times while English and Punjabi appear to have somewhat shorter reading times. In general, the distribution is very skewed, as these example wikis demonstrate.

Log-normal distribution: This is a normal distribution, but on a logarithmic scale. Differences in means between log-normal samples can be tested using t-tests. Such advantages make the log-normal distribution a common choice in analyzing skewed data, even when it is not a perfect fit.

Lomax (Pareto Type II) Distribution: Datasets on human behavior often exhibit power-law distributions, meaning that the probability of extreme events, while still low, is much greater than would be predicted by a normal (or log-normal) distribution (Clauset et al., 2009). We fit the Lomax Distribution, a commonly used long-tailed distribution with two parameters that assumes that power law dynamics occur over the whole range of the data.

Weibull Distribution: Liu et al. model reading times on web pages using a Weibull Distribution (Liu et al., 2010). This model has two parameters: λ, a scale parameter, and k, a shape parameter. The Weibull distribution can be a useful model because of the intuitive interpretation of k. If $k > 1$, then reading behavior exhibits positive aging, which means that the longer someone stays on a page, the more likely they are to leave the page at any moment. Conversely $k < 1$ is interpreted as negative aging, which means that as someone remains on a page, they become less likely to leave the page at any given moment. The Weibull distribution is often used in the context of reliability engineering for modeling the chances that a given part will fail at a given moment.

Exponentiated Weibull Distribution: The Weibull model assumes that the rate of readers leaving a page changes monotonically over time. This implies there must be either negative aging, positive aging, or no aging. It excludes more complicated dynamic processes where positive aging gives way to negative aging after a point in time. The exponentiated Weibull distribution is a three-parameter generalization of the Weibull distribution that relaxes this constraint (Pal et al., 2006). The extra degree of freedom will allow this model to fit a greater range of empirical distributions compared to the two-parameter Weibull model.

Methods

Our method for model selection is inspired in part by Liu et al., who compared the log-normal distribution to the Weibull distribution of dwell times on a large sample of web pages (Liu et al., 2010). They fit both models to data for each web page, and then compare two measures of model fit: the log-likelihood, which measures the probability of the data given the model (higher is better), and the Kolmogorov-Smirnov distance (KS-distance), which is the maximum difference between the model CDF and the empirical CDF (lower is better). For the sample of web pages they consider, the Weibull model outperformed the log-normal model in a large majority of cases according to both goodness-of-fit measures.

Similar to the approach of Liu et al., we fit each of the models we consider on reading time data, separately for each Wikipedia project (Liu et al., 2010). In addition to the KS-distance, we also use KS-tests of the null hypothesis that the model is a good fit for the data to evaluate goodness-of-fit (Clauset et al., 2009). For the samples sizes we use, passing the KS-test is a high bar.

Adding parameters can increase model fit without improving out-of-sample predictive performance or explanatory power. To make fair comparison between models with different numbers of parameters, we use the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) instead of the log-likelihood. Both criteria attempt to quantify the amount of

model	AIC rank		BIC rank		ks rank		KS p-value		KS 95%		KS 97.5%	
	mean	med.	mean	med.	mean	med.	mean	med.	mean	passing	mean	passing
Lomax	1.78	2	1.70	1	2.09	2	0.26	0.17	0.79	192	0.87	211
Log-normal	2.20	2	2.10	2	2.33	2	0.27	0.17	0.71	173	0.79	191
Expon. Weibull	2.15	2	2.34	3	2.11	2	0.29	0.23	0.77	187	0.84	203
Weibull	3.98	4	3.94	4	3.84	4	0.07	0.00	0.24	59	0.30	72

Table B.2: Goodness of fit statistics resulting from the model selection process on 242 wikis. The Lomax, log-normal, and exponentiated Weibull distributions fit the data reasonably well, but the Lomax most often fits the best. The "mean" columns under KS 95%, and KS 97.5% refer to the proportion of wikis passing KS-tests at the 95% and 97.5% significance levels, and the "passing" columns states the absolute number.

information lost by the model (lower is better), by evaluating the log likelihood, and adding a penalty for the model parameters. The difference between AIC and BIC is that BIC maintains the penalty for larger sample sizes.[9]

Following Liu et al., we build these goodness-of-fit measures for each wiki and rank them from best to worst (Liu et al., 2010). For each distribution, we report the mean and median of these ranks. In addition, we report the mean and median p-values of the KS-tests as well as the number and proportion of wikis that pass the KS-test for each model.

We fit the models using SciPy. The exponentiated Weibull, Weibull, and Lomax models were fit using maximum likelihood estimation and the log-normal distributions were fit using the method of moments.

B.6. Results

Table B.2 below shows the results of this procedure. The Lomax, exponentiated Weibull, and Log-normal all fit the data reasonably well. All pass the KS-test for many wikis, and are in a three-way tie for best median rank according to AIC. Despite this, none of our candidate models pass the KS test for all wikis: There are 28 wikis where all 4 models fail to pass at the 95% level, and 13 wikis where they all fail at the 97.5% level.

The Lomax distribution is the best fit across all wikis according to all metrics. With only 2 parameters, it has a lower AIC and BIC than the three-parameter exponentiated Weibull distribution and passes the KS-test 79% of the time at the 95% confidence level. The exponentiated Weibull model fits the data better than the log-normal model in terms of passing KS-tests and with respect to AIC. However, the log-normal is better in terms of BIC, which imposes a greater penalty on the additional parameter of the exponentiated Weibull model.

The Weibull model fits substantially worse than the Lomax, log-normal, and exponentiated Weibull in terms of all of our goodness-of-fit metrics. In this respect, our results differ from those of Liu et al., who observed the Weibull model fitting dwell time data better than the Log-normal model (Liu et al., 2010). We observe that for dwell times on Wikipedia, the Log-normal model is the better fit. While substantially worse than the Lomax model, the Log-normal model still passes the KS-test at the 95% level for about 71% of wikis in the sample.

[9] We provide a more detailed example of this procedure in our online supplement at https://w.wiki/5Jo.

Discussion

We found that the Lomax, exponentiated Weibull, and log-normal models all fit the data within reason. We now discuss how each of these models can be applied to understanding Wikipedia reading behavior.

Lomax (Pareto Type II) Distribution: That the Lomax model fits well suggests that Wikipedia reading times may follow a power law. Mitzenmacher (2004) describes several possible data generating processes for power law (Pareto) and log-normal distributions (Mitzenmacher, 2004). Rich-get-richer dynamics such as preferential attachment are commonly associated with power law distributions, and a mixture of Log-normal distributions can also generate a power law (Mitzenmacher, 2004). Deeper exploration of potential power-law dynamics in reading behavior is a potential avenue for future research.

Log-Normal Distribution: The log-normal model does not fit the data perfectly, but it fits well enough to be useful. It frequently passes KS-tests, and is preferred to the exponentiated Weibull by the BIC. Even though the Lomax model typically fits the data better, assuming a log-normal model justifies using t-tests to compare differences in geometric means when evaluating experiments. Furthermore, assuming log-normality can help justify using ordinary least squares to estimate regression models in multivariate analysis (as we do below) instead of models that require maximum likelihood estimation.

Weibull Distribution: The Weibull model did not fit the data well. While Liu et al. observed that the Weibull model out-performed the log-normal model on their datasets, we (along with Yin et al. (2013)) observe the opposite. However, the exponentiated Weibull model generalizes the Weibull, is a good fit for the data, and can help us explain why the Weibull does not fit the data well.

Exponentiated Weibull Distribution: The exponentiated Weibull has 3 parameters (Pal et al., 2006). Two are shape parameters ($\alpha > 0$ and $\gamma > 0$) and one is a scale parameter ($\lambda > 0$). The major qualitative distinctions in interpreting the model depend on the shape parameters. In many cases the parameters can be interpreted in terms of a transition from negative to positive aging (or visa-versa) after some threshold. However, if either $\gamma > 1$, $\alpha < 1$ or $\gamma < 1$, $\alpha > 1$ then qualitative interpretation may require closer inspection of estimated hazard functions.

Inconveniently, we estimated $\alpha > 1$ and $\gamma < 1$ for all but one of the 285 Wikipedia projects we analyzed. This limits the usefulness of exponentiated Weibull models for large-scale analysis on many wikis because the parameters are outside the area where the model leads directly to intuitive qualitative interpretations. However, by plotting the estimated hazard function we can see over what range of the data the hazard function is decreasing or increasing, accelerating or decelerating.

In figure B.4 we observe that, on English Wikipedia, the log-normal and exponentiated Weibull models both indicate a brief period of positive aging, during which the instantaneous rate of page-leaving increases, followed by negative aging. This helps explain why the Weibull model is not a good fit for the data compared to the log-normal and exponentiated Weibull models: the Weibull distribution cannot model a non-monotonic hazard function. While Liu et al. found it to be a good model for the distribution of dwell times in data collected through a web browser plugin, our analysis suggests that the behavior of Wikipedia readers may be somewhat more complex. Perhaps whereas Liu et al. operationalized “screen-and-glean” as a monotonically decreasing hazard function, Wikipedia readers require more than 1 or 2 seconds to "screen"

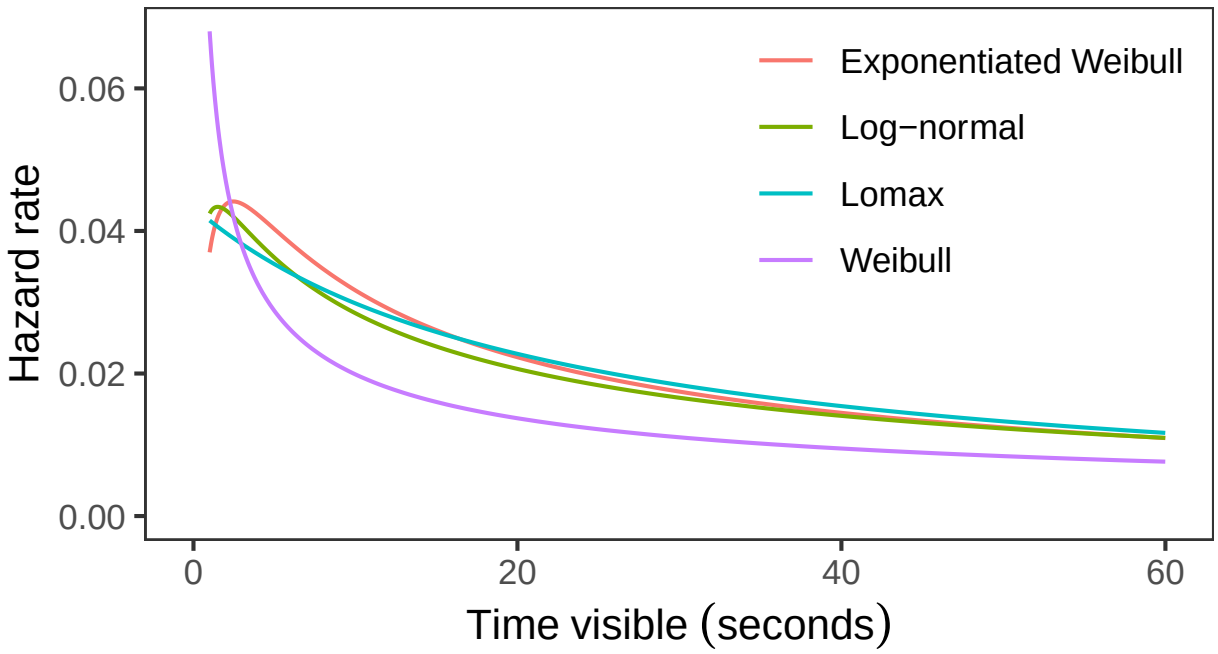


Figure B.4: Hazard functions for the parametric models estimated on English Wikipedia. The exponentiated Weibull model (the best fit to the data) indicates that the hazard rate increases in the first seconds of a page view, after which we observe negative aging.

the page and during these first few moments, their hazard of leaving it increases.

B.7. Reading time and global contexts

Now we return to our analysis of Wikipedia readers in a global context. Our analysis is most closely inspired by Lemmerich et al.'s large-scale global survey of Wikipedia readers. They found that readers in lower-HDI countries are more likely to use Wikipedia in educational contexts and for intrinsic learning, but not for fact-checking (Lemmerich et al., 2019). Such motivations and contexts are likely to involve longer sessions and dwell times compared to fact-checking (Lemmerich et al., 2019; Singer et al., 2017). Therefore, we predict that readers in lower-HDI countries and in the Global South are more likely to have longer dwell times on Wikipedia articles.

H1: Readers in countries with lower HDI (or the Global South) are more likely to spend more time reading each page they visit compared to readers in countries with higher HDI (or the Global North).

We also test a second prediction of the theory that Global South readers are more likely to use Wikipedia for in-depth understanding. If desktop devices have advantages for reading to gain in-depth understanding then users may be more likely to choose these devices for such tasks (when they have the choice). Furthermore, Global South readers may also experience gaps limiting their access to desktop devices, and when they do have access may be likely to take advantage of such opportunities by reading longer. Therefore, we expect users in countries

within the Global South designation (or with lower HDIs) to read even longer on desktop devices.

H2: The difference between the reading times of readers in countries with lower HDI compared to readers in higher-HDI countries will be greater on desktop than on mobile devices.

Based on the "screen-and-glean" model of information seeking behavior that Liu et al. observed on the web (Liu et al., 2010), we propose that reading of articles for in-depth understanding is most likely to take place in the last page view in a session. Differences in reading time in other page views might be attributable to less efficient "screening"—gaps in the skills required to efficiently sift through Wikipedia pages to find the page with the information sought. However, the final page view in a session may reflect "gleaning"—information consumption. If so, then the last page view in a session provides an opportunity to isolate information consumption from information seeking.

Therefore if the gap between low and high development context readers is attributable to types of information seeking tasks, and in-depth reading tasks require more time spent "gleaning," then we predict that the gap between reading time in low versus high HDI countries will also be amplified on the last page view in a session.

H3: The difference between reading times in countries with lower HDI and countries with higher HDI will be greater on the last page view in a session than on other page views.

On the other hand, a "skills gap" with respect to information screening may drive an opposite result. The gap between reading times in the Global South and the Global North may shrink on the last page view in a session if Global South readers are less efficient at filtering information.

Methods and measures

The EventLogging system records the date and time the page was viewed. We include *Day-Of-Week* and *Month* as statistical controls for seasonal and weekly reading patterns. Including *NthInSession* statistically adjusts for the number of pages a reader has viewed so far in the session. *Revision Length*, the size of the wiki-page, measured in bytes, roughly accounts for the amount of content on the page. We use two other measures from the instrument to statistically adjust for page load time: *time till first paint*, the time from the request until the browser starts to render any part of the page; and *dom interactive time*, the time from the request until the user can interact with the page.[10]

We obtain the *page length*, measured in bytes at the time the page was viewed, by merging the EventLogging data with the edit history. To understand how reading behavior on *mobile* devices differs from behavior on non-mobile (i.e. desktop) devices, we assume that visitors to mobile web-hosts (e.g. en.m.wikipedia.org) are using mobile devices and that visitors to non-mobile web-hosts (e.g. en.wikipedia.org) are on non-mobile (desktop) devices.

We determine the approximate country in which a reader is located from the MaxMind GeoIP database which is integrated with the Wikimedia analytics pipeline.[11] We use the United

[10]See https://developer.mozilla.org/en-US/docs/Web/API/PerformanceNavigationTiming/domInteractive *archived at* https://perma.cc/RRA8-8SQG, DOM refers the page's "document object model" structure

[11]See https://wikitech.wikimedia.org/wiki/Analytics/Systems/Cluster/Geolocation *archived at* https://perma.cc/C36T-2E4E

Nations' human development index (*HDI*) to measure the development level of the country.[12] We lack geolocation data before March 3rd 2018, which limits our analysis of reading times in the global context to the period from then until September 28th 2018. We standardize the HDI by centering to 0 and scaling it by the standard deviation (taken at the country level) because the partial residual plots of interaction term between (unscaled) HDI and mobile were very skewed. This also allows us to interpret results in terms of standard deviations.

We also use the established regional classifications of Global North and Global South[13] as a second, dichotomous, measure of development. Finally, the EventLogging instrumentation retains a session token with which we measure whether or not a given page view is the *last-in-session*. We also statistically adjust for the number of pages viewed in the session so far (*Nth in session*).

Models We test the three hypotheses using two regression models that differ only in how they represent economic development. *Model 1a* uses the human development index (HDI) and *model 1b* uses the Global North / Global South regional classification. Here is the specification of *model 1a*:

$$Y = B_0 + B_1 HDI + B_2 Mobile + B_3 Mobile\ x\ HDI$$
$$+B_4 RevisionLength + B_5 DayOfWeek + B_6 Month$$
$$+B_7 NthInSession + B_8 LastInSession$$
$$+B_9 HDI\ x\ LastInSession + B_{10} Mobile\ x\ LastInSession$$
$$+B_{11} FirstPaint + B_{12} DomInteractiveTime$$

The formula for *model 1b* is the same except for using *GlobalNorth* terms instead of *HDI*.

We consider **H1** supported if $B_1 < 0$ in both models; **H2** if $B_3 > 0$; and **H3** if $B_9 < 0$. Because interaction terms can be difficult to interpret qualitatively, we will present marginal effect (ME) plots to assist in qualitative interpretation of the observed relationships (Pepinsky, 2018).

We explored alternative model specifications that include higher order terms and additional interaction terms. We choose to present *model 1a* and *model 1b* because more complex models neither substantively improve the explained variance and the predictive performance nor lead to qualitatively different conclusions. We fit both models using weighted ordinary least squares estimation in R on a stratified sample of size 9,873,641.

Non-parametric Analysis

Our multivariate regression analysis assumes a parametric model and as we saw in the univariate analysis above, the assumption of log-normality may not be valid for every Wiki. Therefore, we also provide a simple non-parametric analysis based on median reading times. Unlike the regression analysis, the non-parametric analysis does not include statistical controls or afford

[12] From http://hdr.undp.org/en/data *archived at* https://perma.cc/SLQ3-HS8S. The HDI is a number between 0 and 1.

[13] See https://meta.wikimedia.org/wiki/List_of_countries_by_regional_classification *archived at* https://perma.cc/WHN7-GB9D

statistical hypothesis tests, but it avoids having to depend on assumptions about the distribution. We construct a 3x3 table of users depending on whether they are in the Global North or Global South, on a mobile or desktop device, or on the last page view in their session. The medians of each cell of the table validate that our findings are not driven by the normality assumption alone.

B.8. Results

We use marginal effects (ME) plots to interpret our regression models.[14] A marginal effects plot shows how the model's predicted outcome varies with respect to one or more of the predictors when other terms of the model are held constant at some typical value (Pepinsky, 2018). Since we are interested in comparing reading times between last-in-session page views and other page views, we create two marginal effects plots for each model: one for last-in-session page views and one for non-last-in-session page views. Similarly, we also break down predicted reading times by device type.

For each marginal effects plot, the y-axis shows the model predicted values and the x-axis shows the values of the predictor variables. In the marginal effects plots shown here, uncertainty intervals represent confidence intervals of the parameter estimates, not uncertainty about the model predictions. Uncertainty about model predictions in this case is generally very high, as our models explain only a small fraction (about 7%) of the variance in reading times.

Hypothesis 1: Global context and reading times

We find support for **H1**: that readers in higher-HDI countries ($\mathrm{B} = -0.20,\ SE = 0.002$) or in the Global North ($\mathrm{B} = -0.27,\ SE = 0.002$) are likely to spend less time on each page than readers in lower HDI countries or in the Global South. For illustration, our ME plot for *model 1a* (figure B.5) shows that, for non-last-in-session page views, a prototypical reader on a desktop device in a country with an HDI one standard deviation below the mean is predicted to spend about 25 seconds on a given non-last-in-session page view compared to the predicted 18 seconds spent by an average reader in a country with an HDI one standard deviation above the mean. Similarly, per our ME plot for *model 1b* (figure B.1), for last-in-session page views on desktop devices, a prototypical Global North reader is predicted to spend around 42 seconds per page view compared to the 50 seconds spent by a prototypical Global South reader.

Hypothesis 2: Global context and mobile devices

We also find support for **H2**: that readers in the Global North ($\mathrm{B} = 15,\ SE = 0.002$) or higher-HDI ($\mathrm{B} = 0.11,\ SE = 0.002$) countries are likely to spend even less time reading compared to Global South or lower-HDI readers when they are on a desktop device compared to a mobile device. This is clearly visible as a differences in slopes in figure B.5. Indeed, for pages views other than the last-in-session, the predicted reading times for prototypical readers in countries 1 standard deviation below the mean decreases from 25 seconds on desktop devices to 22 seconds on mobile devices, but the reverse is true for readers in higher-HDI countries. In a country 1 standard deviation above the mean, an otherwise comparable reader is predicted to read for

[14] Full regression tables are available in the appendix.

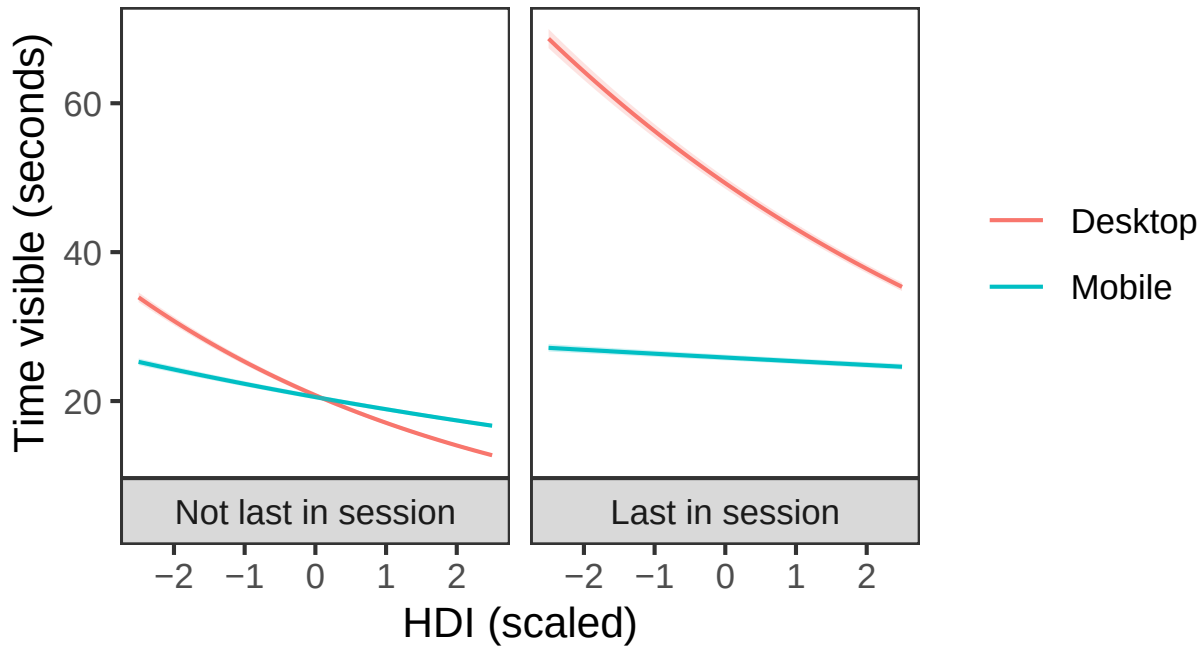


Figure B.5: Marginal effects plot showing the relationship between HDI and reading time predicted by *model 1a*. The negative slope of the lines shows that lower-HDI readers have longer reading times, and the difference in slopes between devices shows that the relationship between HDI and reading time is more pronounced on desktop devices. The ribbons reflect 95% confidence intervals of the model coefficients. The x-axis units represent standard deviations from the mean HDI.

about 19 seconds on mobile and about 17 seconds on desktop. The ME plot for *model 1b* (figure B.1) shows that for the prototypical reader, the gap between Global South and Global North is greater on desktop devices (about 5 seconds) than on mobile devices (about 3 seconds).

Hypothesis 3: Global context and last-in-session

Based on the "screen-and-glean" results by Liu et al, we expected in-depth reading to be most likely in the last page view in a session, and thus predicted **H3:** the difference in reading times between lower-HDI countries and higher-HDI countries will be amplified in the last page view in a session. However, we do not find support for this hypothesis, which would have been indicated by a negative regression coefficient for the interaction term between development and last-in-session. Instead we find a positive coefficients for *HDI:Last in session* ($\mathrm{B} = 0.63,\ SE = 0.002$) in *model 1a* and for *Global North:Last in session* ($\mathrm{B} = 0.08,\ SE = 0.002$) in *model 1b*.

Non-parametric Analysis

Table B.3 shows the median time pages are visible by the user's economic region, device and whether a page is the last viewed in the user's session. Consistent with **H1**, median users in

the Global South spend more time on pages compared to median users in the Global North regardless of device or session stage. Consistent with **H2**, the difference between Global South and Global North users is clearly more pronounced on desktop compared to mobile. In contrast to the prediction of **H3**, but in line with the findings from our parametric analysis, we do not observe an accentuation of the difference between Global South and Global North users in the last page view in a session.

Page length

In addition to the above results on reading times and global contexts, we also examined how reading times relate to page length. The association between page length and reading times is small and positive ($\mathrm{B} = 0.17, SE = 0.0004$). Pages on Wikipedia vary greatly in length: from just a few bytes up to 2,000,000 bytes. If a page were to double its length, our model would predict a marginal increase in reading times of a factor of 1.2. For example, a page with 10000 bytes has a predicted reading time of 25 seconds, which for a page with twice that length (20000 bytes) increases to 30 seconds.[15]

B.9. Limitations

Two important technical limitations of our dwell time data affect our ability to compare reader behavior between mobile phone and PC devices. The first is missing data on mobile devices, discussed above. This missing data likely introduces a negative bias to our measures of reading time on mobile devices because we believe observations are more likely to be lost when users

[15]See our online supplement at https://w.wiki/5Jo for a marginal effects plot. Page length refers to the size of the wikitext source of the page measured in bytes. Not every byte corresponds to a character of readable text. Wikitext source also includes code for formatting, using templates, or embedding images. Additionally, some characters, especially in non-Latin alphabets, may take up multiple bytes. Still our results confirm that for longer Wikipedia articles, only a fraction of the text is read in a typical page view. Assuming a reading speed of around 250 words per minute and an average word length of 5 characters in English (not including spaces and punctuation), these 30 seconds would only suffice to read through less than 1000 of these 20000 bytes (Bell, 2001/00/00; Bochkarev et al., 2012).

Economic-region	Desktop	Last-in-session	Time-visible
North	False	False	20.1
South	False	False	21.5
North	True	False	16.1
South	True	False	21.8
North	False	True	28.1
South	False	True	28.7
North	True	True	39.8
South	True	True	43.6

Table B.3: Table of median reading times by last-in-session, economic region, and device type. Reading times in the Global South are greater than in the Global North in all categories, and are markedly greater on desktop compared to mobile devices.

switch tasks from the browser, and subsequently return to reading. This bias may be quite significant as the issue affects a large proportion of our sample.

The second limitation occurs when readers leave a page visible in the browser at times when they are not directly reading it. For example, a user may have multiple windows visible while only looking at one of them, or may leave a browser window visible and move away from the computer for a long period of time. In general, the best we can hope to observe is that a page is visible in a browser. We cannot, through this instrument alone, know with confidence that an individual is reading. This limitation leads to positive bias in our measures of reading time. To partially address this limitation, we fit regression models on data with dwell times greater than 1 hour removed (assuming that it contains a higher ratio of those "visible but not reading" cases), and found that our results were not substantively affected by the change.

It is possible that this positive bias may correlate with our analytic variables. Perhaps last-in-session views may be particularly subject to this source of bias and may contribute to the gap we observe between reading times in last-in-session page views compared to others. We designed our analysis of H1 and H2 to account for differences between last-in-session and other page views, and found that the sign of the observed differences remained the same whether the view was the last in a session or not. We did not find support for H3, which considered differences within last-in-session page views.

Additional steps could be taken to construct new measures of reading that would not suffer this limitation through browser instrumentation to track mouse movements or scroll positions. However, such steps should be taken with care as additional data collection may negatively affect users in terms of privacy, browser responsiveness, page load times, and power consumption.

Finally, readers should keep in mind that we analyzed observational, not experimental, data with the intention to describe correlations between our variables, not to demonstrate causal relationships. We used ordinary least squares analysis, but future analysis might better account for the hierarchical structure of our data using multilevel modeling.

Alternative explanations

Furthermore, there are several plausible alternative explanations that we cannot rule out in the presented analysis. The observed reading time gap between more and less developed countries may be due to factors other than the types of information seeking tasks in which readers are engaged. For instance, if readers experience knowledge gaps in less developed countries, they may be likely to read in languages that are not their primary language, and thus spend more time reading regardless of task (Graham et al., 2014). A future iteration of this project may partially address this limitation by accounting for whether a Wikipedia edition is a common primary language in the reader's country.

Another alternative explanation may be that the gap between readers in more and less developed countries is partly due to time spent on exploration ("screening") rather than on content consumption ("gleaning"). Our finding rejecting **H3**, suggests this, as Global South readers have longer dwell times on non-last-in-session page views compared to Global North readers. We also observe shorter non-last-in-session page views on desktop devices compared to mobile for Global North readers, but for Global South readers such page views are about the same length no matter what device is used. This unexpected result would be consistent with a skills gap experienced by Global South readers who may have greater difficulty finding sought information,

especially when using desktop devices (van Deursen & van Dijk, 2015). The present analysis offers only tentative support for this claim, but we suggest it as an avenue for future research.

Global South readers may also be more sensitive to the price of downloading data and thus they may avoid opening pages that they are unlikely to read in-depth. Future work might use data from the Wikipedia Zero project to study the relationship between price sensitivity and Wikipedia audiences. More generally, drawing conclusions about information seeking from our analysis rests on strong assumptions about relationships between task type and reading times. Future work on information seeking behavior on Wikipedia testing these assumptions would help validate such conclusions.

B.10. Discussion and Conclusion

In an analysis of novel data from Wikipedia, measuring the time that web pages are visible in the browser window as an approximation of reading time, we investigated patterns of reader behavior across global contexts and found systematic differences consistent with greater use for in-depth understanding in lower-HDI countries compared to higher-HDI countries. We believe this analysis should strengthen confidence in similar findings from surveys of reader behavior because our data have complementary strengths and limitations compared to self-report data.

We conclude that Global South readers are more likely to engage in in-depth information seeking when reading Wikipedia compared to Global North readers. Consistent with Lemmerich et al.'s survey results (Lemmerich et al., 2019), we find that readers in lower-HDI countries have longer reading times than readers in higher-HDI countries, and that this difference is greater for users of non-mobile (desktop) devices.

The observed relationships are quite similar whether measured using the human development index (HDI) or dichotomized economic region (Global South / Global North). These relationships are supported not only by the regression models, but also by non-parametric analysis. While Wikipedia readers increasingly use mobile devices to visit Wikipedia, they are likely to spend the most time reading when they are in the last page view of a desktop session. This is exactly when we expect them to gain in-depth understandings of topics.

We lack evidence to fully explain our findings in terms of structural and socioeconomic differences between the Global North and Global South. One possibility is that the gap in reading times reflects differences in information seeking and content understanding skills (Shaw & Hargittai, 2018; van Deursen & van Dijk, 2015). That we did not observe the gap between global contexts widen in last-in-session page views tentatively suggests that Global South readers are more likely to struggle to find and filter information on Wikipedia compared to Global North readers.

However, given the evidence that Wikipedia readers in the Global South are more likely to engage in deeper information seeking tasks (Lemmerich et al., 2019), we conjecture that the gap in reading times may be explained by the quality and accessibility of the information on Wikipedia relative to alternatives available in the reader's contexts. Wikipedia may not be perfect, but given historical inequalities in education, and knowledge production between the Global South and Global North (Graham et al., 2014), it still might be competitive compared to other sources, especially when it comes to encyclopedic content about the Global South, content in local languages, and information not otherwise available for free to Internet users. This would explain why Global South readers would be more likely to choose Wikipedia when

seeking in-depth information. Future research might test this hypotheses in audience surveys or by adapting approaches previously applied to gender comparisons on English Wikipedia (Reagle & Rhue, 2011).

Another contribution of this study is to vet the reading time data to understand its limitations and to conduct model selection to justify parametric assumptions for future analysts. We found a high rate of missing data on mobile, among other less significant irregularities. Future analysts should keep this in mind and work to improve the coverage. We found that the log-normal distribution often fits the data well, and therefore adopted the use of geometric means as a metric for comparing samples reading times. This also helped support our decision to adopt ordinary least squares regression analysis for multivariate comparison. However, we also found that exponentiated Weibull and Lomax probability models were often an even better fit. Future researchers might explore how reader behavior may generate data in processes consistent with these models.

The reading time data we used in this study is a promising tool for future researchers to improve upon studies of page views for understanding Wikipedia's audiences. For example, recent research has shown widespread misalignment between how often articles are visited and the quality of those articles (Warncke-Wang et al., 2015). However, we have observed that not all views are created equal. Future studies on the relationship between content production and content consumption on Wikipedia might use reading time data to learn about how content consumption might change depending on article quality.

Acknowledgements

We are grateful to the anonymous reviewers, whose observations helped improve the paper. Specials thanks to the web team at the Wikimedia Foundation that built the instrumentation, to Zareen Farooqui who conducted initial data quality vetting as part of her data analyst Outreachy internship at WMF, and to the Foundation's Analytics Engineering team for supporting the data analysis infrastructure used in this work. Thanks to those who provided comments on various stages of this research, including Kaylea Champion, other members of the Community Data Science Collective, Johnathan Morgan, Aaron Halfaker, Isaac Johnson, Miriam Redi, Abbey Ripstra, and other members of the Wikimedia research team. Special thanks to Benjamin Mako Hill for his comments and advice. This work was completed while Nathan TeBlunthuis was a PhD student at the University of Washington and in his capacity as a Wikimedia Foundation contractor and affiliate. It was also supported by the National Science Foundation (GRFP-2016220885).

References

Antin, J., & Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2925–2934. https://doi.org/10.1145/2207676.2208699

Bell, T. I. (2001/00/00). Extensive Reading: Speed and Comprehension. *Reading Matrix: An International Online Journal*, *1*(1).

Bochkarev, V. V., Shevlyakova, A. V., & Solovyev, V. D. (2012). Average word length dynamics as indicator of cultural changes in society. *arXiv:1208.6109 [cs]*. Retrieved April 1, 2019, from http://arxiv.org/abs/1208.6109

boyd, d., & Crawford, K. (2012). Critical Questions For Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Clauset, A., Shalizi, C., & Newman, M. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Davies, P. (2013). Medium's metric that matters: Total Time Reading. Retrieved March 30, 2019, from https://medium.com/data-lab/mediums-metric-that-matters-total-time-reading-86c4970837d5

Deursen, A. J. A. M. V., Helsper, E., Eynon, R., & van Dijk, J. A. G. M. (2017). The compoundness and sequentiality of digital inequality. *International Journal of Communication*, *11*, 452–473.

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, *4*(1), 2056305118763366. https://doi.org/10.1177/2056305118763366

Gorbatai, A. D. (2011). Exploring Underproduction in Wikipedia. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 205–206.

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty. *Annals of the Association of American Geographers*, *104*(4), 746–764. https://doi.org/10.1080/00045608.2014.910087

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, *3*(3), 255–261. https://doi.org/10.1177/2043820613513121

Halfaker, A., Keyes, O., Kluver, D., Thebault-Spieker, J., Nguyen, T., Shores, K., Uduwage, A., & Warncke-Wang, M. (2015). User Session Identification Based on Strong Regularities in Inter-activity Time. *Proceedings of the 24th International Conference on World Wide Web*, 410–418. https://doi.org/10.1145/2736277.2741117

Hargittai, E. (2002). Second-Level Digital Divide: Differences in People's Online Skills. *First Monday*, *7*(4). https://doi.org/10.5210/fm.v7i4.942

Hill, B. M., & Shaw, A. (2013). The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS ONE*, *8*(6), e65782.

Jansen, B. J., & Spink, A. (2003). An Analysis of Web Documents Retrieved and Viewed. *International Conference on Internet Computing*, 65–69.

Kiesler, S., & Sproull, L. S. (1986). Response Effects in the Electronic Survey. *Public Opinion Quarterly*, *50*(3), 402–413. https://doi.org/10.1086/268992

Kim, Y., Hassan, A., White, R. W., & Zitouni, I. (2014). Modeling Dwell Time to Predict Click-level Satisfaction. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 193–202. https://doi.org/10.1145/2556195.2556220

Lemmerich, F., Sáez-Trumper, D., West, R., & Zia, L. (2019). Why the World Reads Wikipedia: Beyond English Speakers. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 618–626.

Liu, C., White, R. W., & Dumais, S. (2010). Understanding Web Browsing Behaviors Through Weibull Analysis of Dwell Time. *Proceedings of the 33rd International ACM SIGIR Con-*

ference on Research and Development in Information Retrieval, 379–386. https://doi.org/10.1145/1835449.1835513

Mitzenmacher, M. (2004). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, *1*(2), 226–251. https://doi.org/10.1080/15427951.2004.10129088

Napoli, P. M., & Obar, J. A. (2014). The Emerging Mobile Internet Underclass: A Critique of Mobile Internet Access. *The Information Society*, *30*(5), 323–334. https://doi.org/10.1080/01972243.2014.944726

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, *65*(12), 2381–2403. https://doi.org/10.1002/asi.23162

Pal, M., Ali, M. M., & Woo, J. (2006). Exponentiated Weibull distribution. *Statistica*, *66*(2), 139–147. https://doi.org/10.6092/issn.1973-2201/493

Paranjape, A., West, R., Zia, L., & Leskovec, J. (2016). Improving Website Hyperlink Structure Using Server Logs. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 615–624. https://doi.org/10.1145/2835776.2835832

Pearce, K. E., & Rice, R. E. (2013). Digital Divides From Access to Activities: Comparing Mobile and Personal Computer Internet Users. *Journal of Communication*, *63*(4), 721–744. https://doi.org/10.1111/jcom.12045

Pepinsky, T. B. (2018). Visual heuristics for marginal effects plots. *Research & Politics*, *5*(1), 2053168018756668. https://doi.org/10.1177/2053168018756668

Phillips, D. L., & Clancy, K. J. (1972). Some Effects of "Social Desirability" in Survey Studies. *American Journal of Sociology*, *77*(5), 921–940. https://doi.org/10.1086/225231

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

Priedhorsky, R., Osthus, D., Daughton, A. R., Moran, K. R., Generous, N., Fairchild, G., Deshpande, A., & Del Valle, S. Y. (2017). Measuring Global Disease with Wikipedia: Success, Failure, and a Research Agenda. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1812–1834. https://doi.org/10.1145/2998181.2998183

Reagle, J., & Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communication*, *5*(0), 21. Retrieved June 24, 2019, from https://ijoc.org/index.php/ijoc/article/view/777

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

Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., & Leskovec, J. (2017). Why We Read Wikipedia. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, 1591–1600. https://doi.org/10.1145/3038912.3052716

Stumpf, M. P. H., & Porter, M. A. (2012). Critical Truths About Power Laws. *Science*, *335*(6069), 665–666. https://doi.org/10.1126/science.1216142

van Deursen, A. J. A. M., & van Dijk, J. A. G. M. (2015). Toward a Multifaceted Model of Internet Access for Understanding Digital Divides: An Empirical Investigation. *The Information Society*, *31*(5), 379–391. https://doi.org/10.1080/01972243.2015.1069770

Warncke-Wang, M., Ranjan, V., Terveen, L., & Hecht, B. (2015). Misalignment between supply and demand of quality content in peer production communities. *Proceedings of the Ninth International AAAI Conference on Web and Social Media (ICWSM '15)*, 493–502.

Yi, X., Hong, L., Zhong, E., Liu, N. N., & Rajan, S. (2014). Beyond Clicks: Dwell Time for Personalization. *Proceedings of the 8th ACM Conference on Recommender Systems*, 113–120. https://doi.org/10.1145/2645710.2645724

Yin, P., Luo, P., Lee, W.-C., & Wang, M. (2013). Silence is Also Evidence: Interpreting Dwell Time for Recommendation from Psychological Perspective. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 989–997. https://doi.org/10.1145/2487575.2487663

Table B.4: Regression tables for models 1a and 1b.

	Model 1a	Model 1b
Intercept	1.3660 (0.0085)***	1.3791 (0.0085)***
Global North		−0.2680 (0.0022)***
mobile : Global North		0.1490 (0.0024)***
mobile : Last in Session	−0.6332 (0.0021)***	−0.6349 (0.0021)***
Global North : Last in Session		0.0830 (0.0024)***
Human development index	−0.1961 (0.0018)***	
mobile : HDI	0.1133 (0.0019)***	
HDI : Last in Session	0.0632 (0.0019)***	
Revision length (bytes)	0.1752 (0.0004)***	0.1758 (0.0004)***
time to first paint	−0.0164 (0.0006)***	−0.0171 (0.0006)***
time to dom interactive	0.0025 (0.0009)**	0.0024 (0.0009)**
mobilemobile	−0.0118 (0.0023)***	−0.0142 (0.0023)***
sessionlength	−0.0001 (0.0000)***	−0.0001 (0.0000)***
Last in session	0.8632 (0.0023)***	0.8575 (0.0023)***
nthinsession	0.0002 (0.0000)***	0.0002 (0.0000)***
dayofweekMon	0.0939 (0.0020)***	0.0926 (0.0020)***
dayofweekSat	0.0169 (0.0020)***	0.0175 (0.0020)***
dayofweekSun	0.0322 (0.0020)***	0.0332 (0.0020)***
dayofweekThu	0.0561 (0.0019)***	0.0548 (0.0019)***
dayofweekTue	0.0349 (0.0020)***	0.0326 (0.0020)***
dayofweekWed	0.0757 (0.0019)***	0.0743 (0.0019)***
usermonth4	0.0095 (0.0096)	0.0083 (0.0096)
usermonth5	0.0108 (0.0095)	0.0104 (0.0095)
usermonth6	−0.0102 (0.0097)	−0.0103 (0.0097)
usermonth7	−0.0494 (0.0097)***	−0.0491 (0.0097)***
usermonth8	−0.0119 (0.0097)	−0.0121 (0.0097)
usermonth9	0.0382 (0.0076)***	0.0370 (0.0076)***
usermonth10	−0.0004 (0.0075)	0.0010 (0.0075)
R^2	0.0721	0.0725
Adj. R^2	0.0720	0.0725
Num. obs.	9873641	9873641
RMSE	14.2330	14.2297

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

Preface to Appendix C

The following appendix is a collaborative work with Benjamin Mako Hill and Aaron Halfaker and is published in the Proceedings of ACM on Human-Computer Interaction: Computer Supported Cooperative Work.

Appendix C

Effects of Algorithmic Flagging on Fairness: Quasi-experimental Evidence from Wikipedia

Online community moderators often rely on social signals such as whether or not a user has an account or a profile page as clues that users may cause problems. Reliance on these clues can lead to "overprofiling" bias when moderators focus on these signals but overlook the misbehavior of others. We propose that algorithmic flagging systems deployed to improve the efficiency of moderation work can also make moderation actions more fair to these users by reducing reliance on social signals and making norm violations by everyone else more visible. We analyze moderator behavior in Wikipedia as mediated by RCFilters, a system which displays social signals and algorithmic flags, and estimate the causal effect of being flagged on moderator actions. We show that algorithmically flagged edits are reverted more often, especially those by established editors with positive social signals, and that flagging decreases the likelihood that moderation actions will be undone. Our results suggest that algorithmic flagging systems can lead to increased fairness in some contexts but that the relationship is complex and contingent.

C.1. Introduction

Online community moderators are responsible for reviewing the torrents of user-generated content for spam, vandalism, attacks, and other violations of community norms and rules. In many large online communities, a small number of moderators—often volunteers—will be responsible for reviewing thousands or millions of actions and taking steps to stop and mitigate problematic behaviors (Gillespie, 2018). To help focus their attention within this deluge, moderators typically rely on social signals (Donath, 2014) that indicate that a user's contributions are made in good faith and of high quality (Kraut et al., 2012). Common signals include visible reputation scores, user profiles, experience, and registration status (Broughton, 2008; Kraut et al., 2012). For example, since new users are often more likely to engage in bad behaviors, moderators might scrutinize contributions from newcomers more closely (Kraut et al., 2012; Potthast et al., 2008). However, directing limited moderation attention based on social signals can introduce unfairness through "overprofiling" that occurs when moderators focus their attention on users with signals associated with bad behaviors while ignoring others engaged in similar or worse behaviors (de Laat, 2016). For this reason, and because relying on social signals can still place enormous demands on limited moderator resources, online communities are increas-

ingly adopting algorithmic flagging systems to direct moderators toward problematic actions (Chandrasekharan et al., 2019; Halfaker & Geiger, 2020).

Although the consequences are very different, these systems share salient commonalities with algorithmic flagging systems used in employment, college admissions, and criminal justice. All of these systems use predictions of whether an outcome will occur to flag certain individuals as more or less likely sources of problems. All leave final decisions to a human judge. The use of these systems when people's lives are at stake has rightfully attracted criticism based of how algorithms engage in misrepresentation and discrimination (Barocas et al., 2019; Campolo et al., 2017; O'Neil, 2018). On the other hand, advocates of algorithmic prediction in criminal justice argue that algorithms—even those that are measurably biased in their predictions—might still be less discriminatory than decisions made by biased human judges alone (Kleinberg et al., 2018; Stevenson, 2017).

Can algorithmic flagging systems in online community moderation similarly reduce reliance on social signals and lead to more fair outcomes? We aim to answer this question through a field evaluation of an algorithmic flagging system called RCFilters, which was deployed on 23 different Wikipedia language editions from January 2019 to March 2020. RCFilters flags contributions identified by the Objective Revision Evaluation Service (ORES) machine learning system as likely to be damaging (Halfaker & Geiger, 2020). These flags are shown along with existing social signals of quality. We take advantage of a set of arbitrary thresholds built into RCFilters to conduct a quasi-experimental analysis that estimates the causal effect of algorithmic flagging on moderation decisions and that seeks to measure whether algorithmic flags lead to better or worse outcomes for users who are likely to be overscrutinized *ex ante*. Our results suggest that algorithmic flagging can lead to more fair outcomes but that this effect may depend on the specifics of the social signals in question.

Our paper makes several contributions. First, our work answers calls to analyze the impacts of algorithms *in situ* (Selbst et al., 2019; Stevenson, 2017; Zhu et al., 2018) by offering an empirical evaluation of an algorithmic flagging system in an important social computing context. Second, our analysis contributes to an ongoing debate over when and how algorithms might lead to more or less fair outcomes for individuals subject to profiling by human decision makers. Third, our work offers a methodological contribution by presenting a novel quasi-experimental approach that can act as a template for future non-interventional studies of causal effects of algorithmic decision support systems. Finally, our work contributes to social computing system design by suggesting improvements to algorithmic flagging and filtering systems.

C.2. Background

Moderation in Online Communities

Contemporary online communities are flooded with harassment, spam, misinformation, disinformation, and hate. Users of social media systems frequently and flagrantly violate community and platform rules, various laws, and norms of decency and decorum. Even users acting in good faith can do damage by taking conversations off-topic, undermining the stated purpose of communities, and lowering the quality of discourse or the knowledge goods being produced. Protecting online communities from unwanted activity are content moderators—many of them volunteers—that Gillespie (2018) has described as "custodians of the Internet." Moderation work

typically involves three tasks: namely, reviewing content or activity, mitigating damage caused by a problematic behavior, and sanctioning users in different ways (Gillespie, 2018; Jiang et al., 2019; Kiene et al., 2019; Seering et al., 2019).

Grimmelmann (2015) defined moderation as "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." Discussions of content moderation often focus on individuals occupying formal roles as moderators with special rights and responsibilities. For example, several of the moderators in Gillespie's (Gillespie, 2018) account are professional moderators working for major platforms such as Facebook and Twitter. Several moderators, and nearly all of them on platforms such as Reddit and Discord (Jiang et al., 2019; Kiene et al., 2019; Matias, 2016), work as volunteers but occupy similar positions of formal authority and responsibility. That said, the work of moderation is also distributed across regular community members (Kiene et al., 2016; C. Lampe & Resnick, 2004). In Wikipedia, for example, the bulk of moderation activity as defined by Grimmelmann occurs as normal users review, vet and undo the work of others to mitigate damage and sanction users they believe have behaved badly (Piskorski & Gorbatâi, 2017).

Sanctions Sanctioning involves enforcing norms in ways that attempt to discourage future misbehavior. It is a core part of moderation work because it encourages compliance with norms by communicating that rules will be enforced (Jhaver, Appling, et al., 2019; Srinivasan et al., 2019). Although it also serves to mitigate damage, removing content is a common form of sanctioning because it communicates that an action was inappropriate (Piskorski & Gorbatâi, 2017). Halfaker et al. (2011) showed that removing content is an effective sanction and results in higher quality contributions by the reverted contributor in Wikipedia. Similarly, Srinivasan et al. (2019) found that people whose comments were removed from Reddit were less likely to violate norms in the future.

Although the goal of most sanctioning is to steer participants toward productive behaviors, the effect is often to deter participation. This can be particularly problematic with well-meaning newcomers who often violate norms because they have not yet learned the ropes (Adler & de Alfaro, 2007; Halfaker et al., 2013; Halfaker et al., 2011). Sanctioned newcomers are less likely to continue participating, especially in the absence of clear explanations from moderators (Halfaker et al., 2013; Jhaver, Appling, et al., 2019; Kiesler et al., 2012; Potthast et al., 2008; TeBlunthuis et al., 2018). On Wikipedia and similar communities, high rates of sanctioning can help explain declines in participation and may be an obstacle to building a community that includes diverse participants (Halfaker et al., 2013; Lam et al., 2011; TeBlunthuis et al., 2018).

Meta-norms No moderation system is perfect. Moderators inevitably make mistakes and apply sanctions in ways that are arbitrary and unfair. This is particularly challenging to avoid in distributed moderation models used on sites such as Slashdot or Wikipedia where moderation is conducted by large and diverse groups of untrained and loosely coordinated users. Sanctions can be particularly demotivating to newcomers who feel that sanctions are unfair and incorrect (Gillespie, 2018; Jhaver, Appling, et al., 2019; Srinivasan et al., 2019). Consequently, steps that make moderation more fair might decrease the negative effects of sanctions on community growth.

One way to improve fairness in moderation is through governance structures that enforce accountability (Frey et al., 2019). Toward this end, Slashdot famously created tools for "meta-moderation" that allowed all users to evaluate the decisions of moderators (C. Lampe & Resnick, 2004). Users whose moderation decisions were controversial or at odds with the opinions of other Slashdot members would not be given moderation privileges again. Although formal systems for meta-moderation remain rare, behaviors that take action against controversial sanctions are common and serve a similar social function (Crawford & Gillespie, 2016). "Meta-norms," which prescribe when and how one should issue sanctions against violations of first-order norms (Horne, 2001) are particularly relevant. Reagle (2010) documented the formalization of meta-norms on Wikipedia and Piskorski and Gorbatâi (2017) showed how Wikipedia users maintain meta-norms by undoing sanctions in ways that effectively sanction the originally sanctioning user.

Flagging and Algorithmic Triage Moderators in large online communities can face incredible challenges in scaling their work to handle an enormous mass of content and user activity (Gillespie, 2018; Kiene et al., 2019; Seering et al., 2017; Seering et al., 2019). In interviews conducted by Kiene et al. (2019), small teams of volunteers tasked with maintaining order in large communities described their work as akin to "running a small city." Some platforms deal with scale by employing more paid moderators. However, the work involved can be exploitative, challenging, traumatizing, and expensive (Roberts, 2016). Volunteer moderator teams frequently find it difficult to identify, train, and integrate new members as they grow (Kiene et al., 2016). On average, teams become less likely to add new members as their communities grow (Shaw & Hill, 2014).

For these reasons and others, it is often impossible for communities to scale moderation resources such that human moderators can review all activity. As a result, many moderation systems implement flagging so that a wider group of users can report content for review by moderators (Grimmelmann, 2015). If users reliably flag problematic behaviors, flagging can mitigate issues of scale because moderators focus their attention on behavior that is flagged. Obviously, flagging is far from a perfect solution. From the perspective of a flagged user, flagging can seem arbitrary and opaque (Crawford & Gillespie, 2016). From a moderator perspective, flagging is flawed because disgruntled users can coordinate to overwhelm moderators and target opposing viewpoints (Crawford & Gillespie, 2016). Finally, given that traditional flagging systems continue to rely on volunteer labor, they often fail to fully address issues of scale, leaving many bad actions unflagged, unreviewed, and unsanctioned.

To address this final limitation, communities have turned to algorithmic flagging systems that use computer programs to automatically mark content for review by human moderators (Kiene & Hill, 2020; Kiene et al., 2019; Seering et al., 2017). Although some of these systems rely on keywords, regular expressions, or heuristics, more advanced and flexible versions of these systems use predictions from machine learning models. These systems are seen as promising answers to the problem of moderation at scale because they can be easily be used to review an enormous volume of behaviors, they may be less vulnerable to strategic flagging, and they may be more reliable than human reviewers.

Algorithmic flagging systems can be thought of as human-in-the-loop versions of similar computational systems that engage in fullly automated moderation. For example, numerous digital platforms utilize the PhotoDNA system to automatically identify and remove child pornog-

raphy (Gillespie, 2018). Similarly, Wikipedia's ClueBot NG uses a machine learning predictor to automatically remove vandalism (Geiger & Halfaker, 2013). Although they play a critical role in reducing moderation workloads, fully automated systems are uncertain enough in most of their assessments that they are typically only considered useful in defending against the most clear-cut examples of misbehavior (Gillespie, 2018).

Some machine learning systems that are designed to classify bad behavior are used as a form of algorithmic triage. While the most egregious examples of bad behavior are dealt with by automatic systems, other possible norm violations are flagged for review by human moderators. For example, Reddit allows moderators to define a system of rules based on regular expressions to automatically flag content for further review (Jhaver, Birman, et al., 2019). Algorithmic flagging systems based on machine learning occupy the vanguard of online activity regulation and numerous examples have been described in recent scholarship. Chandrasekharan et al. (2019) described a system for Reddit communities to share information and collaborate on automatic flagging that accounts for differences between rules of different communities. Wulczyn et al. (2017) presented a system for classifying harassing behavior on Wikipedia. Finally, Halfaker and Geiger (2020) developed the ORES system to predict the quality of contributions and content on Wikipedia.

Will Algorithmic Flagging Decrease Discrimination Of Overprofiled Users?

One of the most important debates in contemporary technology policy is the degree to which the introduction of algorithms into socially consequential decision making leads to more or less fair outcomes (Chouldechova, 2017; Kleinberg et al., 2018; O'Neil, 2018; Selbst et al., 2019). Much of this debate focuses on arguments about whether algorithms will amplify or entrench discrimination and on biases introduced by training data (Barocas et al., 2019; Campolo et al., 2017; Sap et al., 2019). Discrimination is the deferential treatment of individuals based on membership in a group. Economists of discrimination distinguish between taste-based and statistical discrimination (Becker, 1957; Bertrand & Duflo, 2016; Phelps, 1972). Taste-based discrimination is driven by preferences for members of one group and includes both ideologically-driven racism and implicit bias. Statistical discrimination occurs when social signals—visible and socially salient characteristics, such as group memberships—are instrumental in driving decisions. Statistical discrimination can also lead to unequal outcomes for certain groups.

Social Signals Although most discussions of discrimination focus on high-stakes contexts such as banking, labor markets, and criminal justice, moderation in online communities is also ripe for statistical discrimination based on visible social signals. When interacting in face-to-face groups, people can observe—and discriminate on the basis of—visible signals of status, group membership, psychological states, or cultural identity (Donath, 2014; Pentland, 2008; Ridgeway, 2019). Because the invisibility of these signals in online communities creates a barrier to regulation, sociability, and cooperation, communities use devices such as profile images and biographies, avatars, or visualizations of activity as tools for self-presentation and signals of membership (Donath, 2014; C. A. Lampe et al., 2007). Disclosing information on profiles can provide signals helpful for people using prototypes (Grabner-Kräuter & Bitter, 2015), building social capital (N. B. Ellison et al., 2011), and developing trust (Ma et al., 2017). Formal reputation systems such as karma on Reddit and Slashdot or badges on StackExchange can be im-

portant signals of commitment, quality, and trustworthiness (Grimmelmann, 2015; C. Lampe, 2012; Merchant et al., 2019)

Even without user profiles or formal reputation systems, participants in online communities use subtle signals to draw conclusions about each other (Donath, 2007; N. Ellison et al., 2006; Jacobson, 1999). Sparse cues such as usernames or communication styles can be signals of personality, gender, and identity Donath, 2014; Hancock and Dunham, 2001; Herring, 2000. Tests of community-specific technical or cultural knowledge can identify newcomers and, similar to formal reputation systems, they may be more challenging to fake than biographical information (Bernstein et al., 2011; Donath, 2014; Grimmelmann, 2015). In peer production projects, prior contributions can be inspected for information about expertise, work styles, and the future value of a newcomer (Marlow et al., 2013).

In several online communities such as Wikipedia, users can elect to participate anonymously, under more-or-less stable pseudonyms, or using their real names. Masking signals of gender, race, age, (dis)ability, or status can appear to equalize and free individuals from oppressive prejudices and stereotypes (Dubrovsky et al., 1991; Friedman & Resnick, 2001). On the other hand, the presence or absence of a stable user identity is itself an essential signal because persistent identities make it possible to build up reputation, social capital, and trust and the inability to do so is associated with misbehavior (Grabner-Kräuter & Bitter, 2015; Hill & Shaw, 2020).

Will algorithmic flagging reduce overprofiling? Online community moderators can use social signals to discover and respond to misbehavior, but this can lead to statistical discrimination. Wikipedia's *Missing Manual* advises would-be vandal fighters on Wikipedia to "consider the source" when "estimating the likelihood that an edit is vandalism" (Broughton, 2008). Because newcomers and anonymous users are more likely to violate rules, moderators may rely on social signals of newness to find bad behaviors or to decide if an ambiguous contribution was made in bad faith. Increased scrutiny and skepticism can translate into an increased likelihood of sanction, simply for being new or anonymous. Statistical discrimination emerges because moderators are more likely to scrutinize and sanction new or anonymous contributors who have legitimate reasons for contributing.

Ethical philosophers have objected to the way social signals are used in online moderation activity. Dutch philosopher Paul de Laat adopted the concept of "profiling" from legal scholar Frederick Schauer to argue against the use—and even the public display of—social signals such as registration status and experience levels in the user interfaces used for moderation because they are prone to "overuse" (de Laat, 2015, 2016). It should be noted that discriminating by attributes such as newness does not raise the same legal or constitutional concerns as discrimination against protected classes such as race or religion. Online communities establish their own norms and may choose to protect or target certain attributes on the basis of a specific community's values. For example, while discussing Wikipedia, de Laat argues that "overuse" is unethical, immoral, and inconsistent with the community's founding principles of transparency and equality. Drawing on de Laat, we refer to individuals with social signals that elicit undue scrutiny as "overprofiled."

Although an important debate continues over the use of algorithmic predictions in domains like criminal sentencing, proponents of algorithms argue that they could reduce discrimination and inequality (Kleinberg et al., 2018; Stevenson, 2017). Algorithms can reproduce statistical

discrimination, but they might be less biased than the alternative: human decisions that would presumably rely heavily, if perhaps subconsciously, on salient social signals such as race. Critics suggest that algorithms simply obscure this discrimination behind complex mathematical models that are difficult to understand, interrogate, or challenge.

Although this debate is challenging to resolve in the case of criminal justice, algorithmic flagging in online community moderation provides a setting with lower stakes and more detailed data. If we apply arguments proposing that algorithms can reduce discrimination to community moderation, we would conclude that algorithmic triage systems would reduce the impact of discrimination among overprofiled individuals by making misbehavior by all kinds of users visible to moderators. If algorithmic flagging reduces overprofiling bias, then it will have a smaller effect on overprofiled users than on others. If algorithms simply reproduce discrimination, we would find no such difference. This leads us to our first research question: ***[RQ1]*** *How will flagging an action change the likelihood an action is sanctioned for overprofiled editors compared with others?*

Algorithmic fairness researchers use specific criteria to quantify biases encoded in algorithmic predictors and the fairness of resulting decisions (Barocas et al., 2019; Chouldechova, 2017; Mitchell et al., 2020). These criteria are often developed for settings where model predictions are equivalent to decisions. For example, Kusner et al. (2017) define demographic parity in terms of model predictions, whereas Mitchell et al. (2020) define it in terms of human decisions. In algorithmic flagging, decisions are informed by algorithms but left to humans. Therefore, we distinguish between the fairness of predictions and the fairness of decisions and refer to our criteria as "decision system fairness metrics" following Mitchell et al.'s (Mitchell et al., 2020) use of the term "decision system."

We first consider demographic parity, as shown in Equation C.1, which means that the probability of a decision (D) is statistically independent of a protected attribute (A) (Barocas et al., 2019; Kusner et al., 2017):

$$P(\widehat{D}|A = 0) = P(\widehat{D}|A = 1) \tag{C.1}$$

An algorithmic flagging system will have demographic parity concerning registration status if the probability that an action is flagged is the same for actions by overprofiled and underprofiled editors. Our analysis of RQ1 thus evaluates how flagging shapes demographic parity for sanctioning decisions.

Will Algorithmic Flagging Increase Fairness?

A system might lack demographic parity by sanctioning one group more than others but still be justifiable if all sanctions are fair. What does it mean for a sanction to be fair? The subject of fairness in algorithmic systems is a major subject of debate in computing and AI. There are several different approaches to conceptualizing fairness, and no algorithmic predictor can satisfy them all (Barocas et al., 2019; Caraban et al., 2019; Kleinberg et al., 2016; Mitchell et al., 2020; Wallach, 2019; Yin et al., 2019). While such approaches focus on discrimination built into machine learning programs, we seek a concept of fairness that reflects the standards of relevant communities of practice. We find one in the concept of "meta-norms" from social psychology and James Coleman's sociological conception of norm maintenance. Drawing from these sources, we define unfair sanctions as those that a community is unwilling to let stand—i.e.,

sanctions that are themselves the subject of sanction (Coleman, 1988; Horne, 2001; Piskorski & Gorbatâi, 2017). For example, norms in Wikipedia govern right and wrong ways of editing wiki pages. Sanctions of first-order norm violations are governed by meta-norms about what sorts of contributions merit sanction. Following Piskorski and Gorbatâi (2017), we describe a sanction as *controversial*—i.e., in likely violation of a meta-norm—if it, in turn, is sanctioned by a third community member.

A controversial sanction suggests that the initial edit was not truly damaging (i.e., $D = 1$ but $Y = 0$ where $Y = 1$ means an edit was truly damaging). Thus, a controversial sanction is analogous to false positive classification by a machine predictor ($\widehat{Y} = 1$ but $Y = 0$, where $\widehat{Y} = 1$ means the machine predicts that an edit is damaging). The false positive rate quantifies the amount of unfair treatment a group experiences, but it does not compare unfair treatment between groups. Therefore, is not strictly speaking an algorithmic fairness criterion. However, changes in the false positive rate of the decision system (shown in Equation C.2) quantify how flagging is increasing or decreasing the rate of unfair sanctions.

$$P(D = 1|Y = 0, \widehat{Y} = 1) - P(D = 1|Y = 0, \widehat{Y} = 0) \tag{C.2}$$

Relying on this definition of fairness, our second research question asks how algorithmic flagging shapes the fairness of sanctioning in terms of the rate of sanctions for meta-norm violations: ***[RQ2]*** *How will flagging an action change the chances it receives a controversial sanction?*

Influential theoretical frameworks in social computing seem to predict competing answers to this second question. First, dual-process models of behavioral economics suggest that people will tend to rely on "salient signals" for rapid decision making in conditions of uncertainty and imperfect information (Bordalo et al., 2012; Kleinberg et al., 2018; Tversky & Kahneman, 1974). When human moderators use social signals to choose behavior to review or sanction, these attributes serve as salient signals but remain far from perfect signals of quality. Algorithmic flags provide an additional salient signal but are also far from perfect (Halfaker & Geiger, 2020). Indeed, algorithmic flagging systems are typically designed to minimize the risk of missing bad behaviors by surfacing large numbers of false positives (i.e., non-problematic behaviors) and relying on human moderators to make final decisions. Of course, if human moderators use algorithmic flags as salient signals, they may reproduce algorithms' false predictions. In this case, controversial sanctions will increase.

A second perspective suggests that algorithmic flags can increase fairness. Several online communities have institutionalized rules, norms and meta-norms and act as highly bureaucratic organizations (Butler et al., 2008; Piskorski & Gorbatâi, 2017). Max Weber described how bureaucratic organizations construct and use two concepts of what he called "rationality:" substantive rationality and formal rationality (Weber, 1978). Substantive rationality refers to how bureaucratic organizations use policies, routines and hierarchy to define their collective values and goals. Formal rationality refers to the use of calculated decision making, such as that involving productivity or financial metrics, in the pursuit of goals (Lindebaum et al., 2019). Following Weber, Kreiss et al. (2011) argued that increasing substantive rationality through bureaucratic policies in online communities can lead to more fair outcomes.

Although less explored by scholars of online communities, there are also reasons to believe that increasing formal rationality in moderation decisions might also enhance fairness, at least in online communities with mature normative systems. In such contexts, algorithmic flagging sys-

tems can enact formal rationality by estimating the probability and displaying an authoritative signal that an action runs afoul of shared behavioral standards. Adopting algorithmic flagging can thus mark a shift away from idiosyncratic individual decision-making and toward increasing the use of formalized rationality. Through this lens, an algorithmic flagging system—even one that encodes biases—can be a "carrier of formal rationality" (Lindebaum et al., 2019), leading to governance that is more in line with community meta-norms and to a decrease in controversial sanctions.

Next, we consider how changes in the false positive rate of the decision system depends on overprofiling. This corresponds to evaluating decision system fairness in terms of equality of opportunity (shown in Equation C.3) (Hardt et al., 2016; Mitchell et al., 2020):

$$P(D = 1|Y = 0, A = 0) = P(D = 1|Y = 0, A = 1) \tag{C.3}$$

Equality of opportunity is satisfied when the false positive rate of a decision system does not depend on the protected attribute. Equality of opportunity for registration status would mean that registered and unregistered editors that make good edits have equal chances of having their contributions accepted.

Our third research question asks whether algorithmic flagging systems will increase or decrease equality of opportunity: ***[RQ3]*** *Within the set of sanctioned actions, how will the effect of flagging an action on controversial sanctions depend on whether contributors are overprofiled?*

Once again, influential theoretical frameworks in social computing research seem to point in opposite directions. Under dual-process psychological models, both social signals and algorithmic flags might cue moderators to issue sanctions and might substitute for one another. In this case, we would hypothesize that flagging would have a more positive effect on controversial sanctions among underprofiled contributors, who had previously been relatively ignored, than it does among the overprofiled individuals, who were always scrutinized. Conversely, if the larger effect of algorithmic flagging is helping moderators comply with meta-norms, it simply will not matter whether contributors are overprofiled.

C.3. Empirical Setting

We aim to answer our three research questions through a field evaluation of an algorithmic flagging system called RCFilters, which was deployed on 23 different Wikipedia language editions between January 2019 and March 2020. RCFilters stands for "Recent Changes filters." The term "Recent Changes" refers to a page on Wikipedia that allows viewers to see the most recent changes made to the site.[1] As Figure C.1 shows, RCFilters adds a set of flags represented as colored dots on the left side of the list of recent contributions. Social signals are also visible, including registration status and whether a user has created a profile page. Although dense with information regarding recent edits and hyperlinks, the page is immediately understandable to Wikipedia moderators. When deployed, the RCFilters interface appears both on "Recent Changes" as well as on "watchlists"—a special version of "Recent Changes" that shows only edits to the subset of pages that a user has elected to follow. RCFilters must be enabled by each user on their Wikipedia user preferences page.

[1]For example, the Recent Changes page for English Wikipedia is available here: https://en.wikipedia.org/wiki/Special:RecentChanges (Archived: https://perma.cc/BNZ3-E9D5)

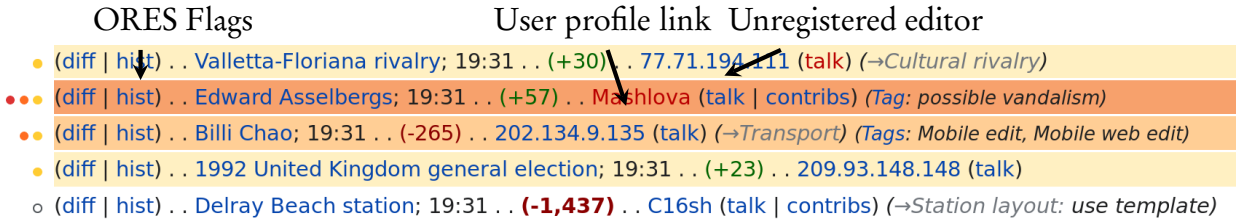


Figure C.1: Screenshot of Wikipedia edit metadata on Special:RecentChanges with RCFilters enabled. Highlighted edits with a colored circle to the left side of other metadata are flagged by ORES. Different circles and highlight colors (white, yellow, orange and red in the figure) correspond to different levels of confidence that the edit is damaging. Users can configure which colors are shown. Visible social signals include registration status (i.e., whether a username or an IP address is shown) and whether an editor's user page and user talk page exist. RCFilters does not specifically flag edits by new accounts, but does support filtering changes by newcomers.

Algorithmic flagging in the RCFilters system is powered by the ORES edit quality models trained to predict whether edits are labeled "damaging" or "not damaging." The models are gradient boosted decision trees trained on a mixture of human-labeled Wikipedia edits and edits made by established editors that are assumed to be "not damaging."

It should be noted that ORES models do not merely reproduce profiling patterns typical of moderation on Wikipedia. The interface for labeling training data obscures social signals from the volunteer Wikipedians doing labeling work and its models are predictive of damage from users that are not anonymous or newcomers. Nevertheless, as discussed in §C.8, ORES encodes biases against unregistered editors and—to a lesser extent—against editors without user pages. ORES was designed neither to merely support quality control in Wikipedia, nor to optimize precision, recall, or fairness but to enact Wikipedian principles of openness, transparency, and community accountability—to "deploy efficient machine learning at scale for content moderation … in ways that enable volunteers to develop and deploy advanced technologies on their own terms" (Halfaker & Geiger, 2020). More information on the philosophy, design and implementation of ORES can be found in Halfaker and Geiger (2020).

C.4. Methods

Our analysis is based on a regression discontinuity design (RDD) that aims to estimate causal the effects of flagging by RCFilters on moderator behavior in Wikipedia (Imbens & Lemieux, 2008; Jacob et al., 2012; Lee & Lemieux, 2010). Common in empirical economics, RDDs are quasi-experimental in that they resemble a randomized control trial for data points in the neighborhood of an arbitrary cutoff (Jacob et al., 2012; Lee & Lemieux, 2010). RDDs model how an outcome depends on this cutoff and a continuous "forcing variable." The idea behind an RDD is that observations immediately below and above the cutoff will be equal in expectation after adjusting for any underlying (i.e., "secular") trend. For example, RDDs used in econometrics might estimate the effect of passing a test by comparing the outcomes of people who barely passed and failed. One benefit of an RDD over a field experiment based on A/B tests is that it can provide ecological validity and support causal claims without subjecting users to intervention without consent (Barocas & Nissenbaum, 2015; Jouhki et al., 2016). Although they remain

rare in computing, RDDs have been used in recent publications in social computing (Hill & Shaw, 2020; Narayan et al., 2019).

Our forcing variable is the score from the ORES machine learning system. Our cut-off variables are a set of arbitrarily chosen operating points used by RCFilters. Our outcomes are constructed by creating two variables that indicate whether a revision's author is overprofiled as well as variables that indicate whether each revision was reverted or subject to a controversial revert. We discuss each in turn before introducing our analytic approach.

Data and Measures

We build our dataset from two publicly available tables of Wikimedia history published by the Wikimedia Foundation (WMF).[2] Although Wikipedia is published and collaborated on in several languages, the vast majority of knowledge regarding collaboration on Wikipedia is derived from studies of English Wikipedia (Hara et al., 2010; Hecht & Gergle, 2010). To support generalizability, we analyze data from 23 language editions of Wikipedia where edit quality flags are displayed in the RCFilters interface. To ensure that we have variation in our outcomes, we exclude wikis with less than three edits above and below each threshold (see §C.4) from each sub-analysis. For all of our analyses, our unit of analysis is the *revision*. Revisions correspond to a single edit to a page by a participant on Wikipedia. We exclude revisions by bots since we care about how algorithmic flagging and social signals are used by human moderators. Following guidance for RDDs (Lee & Lemieux, 2010), we include only revisions very near to RCFilters thresholds, with ORES scores within 0.03 of the thresholds.

To manage the total size of our dataset, we analyze a sample that we construct by stratifying along several dimensions: Wikipedia language edition; user registration status (§C.4); whether the editor has a user page or not (§C.4); whether an edit was reverted in 2 hours, 48 hours, or 30 days; and whether the revert was controversial (§C.4). Then, we sample 5000 edits from within unique combinations of the variables. If there are less than 5000 edits in the given strata, we include all of them. We adjust for this stratification using sample weights throughout our analysis. Since RCFilters was introduced to different wikis at different times, we sample edits during the period immediately following the introduction of ORES but weight our sample according to the number of edits to each wiki over the entire study period. The numbers of observations sampled at each threshold, from each Wiki, and for each model are available in the supplementary material.

ORES scores and RCfilter thresholds The continuous forcing variable used in our RDD analysis is a score from the ORES algorithm described in §C.3. Scores range from 0 to 1 and reflect the predicted probability that a revision is damaging. Because the ORES system has been under continual development over time, we obtain ORES scores created at the times revisions were made from a log maintained by the WMF. The treatments in our analysis are whether edits to Wikipedia are flagged by RCFilters. These flags are applied if, and only if, a score from ORES exceeds a threshold. This use of thresholds at arbitrary operating points is a feature of most algorithmic flagging systems. The intuition behind our RDD is that—after adjusting for small

[2] https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Edits/Mediawiki_history (Archived: https://perma.cc/CPM6-PY6F; https://dumps.wikimedia.org/other/mediawiki_history/readme.html (Archived: https://perma.cc/3DDJ-9FXS)

differences in quality associated with marginally higher or lower scores—edits with ORES scores immediately above and below an arbitrary threshold will be similarly likely to receive both first-order and controversial sanctions. Consequently, any discontinuous change in reverts at one of the thresholds used by RCFilters can be attributed to the flag.

RCFilters uses multiple thresholds corresponding to green, yellow, orange, and red flags. By default, only orange, and red flags are shown, but users can configure which colors to display. Green flags and filters are to help Wikipedia editors find good edits. Our analysis considers only red, orange, and yellow flags, which correspond to thresholds making different trade-offs between precision (the proportion of flagged edits that are truly damaging) and recall (the proportion of truly damaging edits that are flagged). The red flag is labeled "very likely damaging" and corresponds to a high precision threshold. Orange flags corresponds to a "likely damaging" label with greater recall but less precision. Edits with a yellow flag are "maybe damaging" with a high recall but lower precision. RCFilters' thresholds are truly arbitrary and have changed over time and across language editions in response to shifts in the precision and recall of ORES models and in response to community feedback. We were able to collect data on threshold configuration, fully trained ORES models, code, and the precise time that changes were deployed in the WMF server admin log. We combined these data to identify the precise thresholds that were active for each revision in our dataset.

Sanctions Our outcome variable for answering RQ1 must capture sanctioning in Wikipedia. Following a large body of other social computing research, we measure sanctions as identity reverts (e.g., Halfaker et al., 2013; Halfaker et al., 2011; Piskorski & Gorbatâi, 2017; TeBlunthuis et al., 2018). Identity reverts occur when a user undoes another user's edit by restoring a page to an earlier state and are measured by comparing hashes of page revisions (Halfaker et al., 2011).

That said, identity reverts are an imperfect measure of sanctioning. It is also possible for an individual to "self-revert" by undoing their own edit. We therefore only treat a revision as reverted if it was undone, but not by a self-revert. We also limit our measure of sanctioning to revisions that are undone within 48 hour to avoid problems related to mass revert actions such as "blanking" of pages that result in false positives. We are confident that 48 hours is a reasonable window because most damage to Wikipedia will be undone within that amount of time (Geiger & Halfaker, 2013) and a 48 hours window will include reverts caused by RCFilters since any effect of RCFilters is likely to occur quickly.

Controversial sanctions Our outcome variable for answering RQ2 and RQ3 measures controversial sanctions. We follow Piskorski and Gorbatâi (2017) by measuring controversial sanctions as identity reverts that are subsequently reverted by a third party. Specifically, we label a sanction as controversial if the sanction is undone by a third editor who was not the original editor or the reverting editor. Such interactions likely correspond to cases in which a third party observes the initial revert, disagrees with the initial sanction and then acts to reverse the sanction.

Social signals Answering our RQ1 and RQ3 requires that we identify underprofiled and overprofiled individuals in our empirical setting. Drawing from research and documentation for Wikipedia moderators, we identify two such measures shown in the RCFilters interface shown in Figure C.1. Our first measure is whether an editor was logged into an account. Unregistered

Threshold	Edit type	N.	Prop.
Maybe dam.	Not reverted	12,403,717	0.87
Maybe dam.	Rev. controversial	69,395	0.00
Maybe dam.	Rev. not cont.	1,757,866	0.12
Maybe dam.	**Total**	14,230,978	1.00
Likely dam.	Not reverted	1,254,219	0.55
Likely dam.	Rev. controversial	31,652	0.01
Likely dam.	Rev. not cont.	1,009,108	0.44
Likely dam.	**Total**	2,294,979	1.00
V. likely dam.	Not reverted	58,474	0.15
V. likely dam.	Rev. controversial	12,545	0.03
V. likely dam.	Rev. not cont.	323,762	0.82
V. likely dam.	**Total**	394,781	1.00

(a) Counts and proportions of edits by whether an edit was reverted or controversially reverted in the neighborhood of each threshold.

Threshold	Editor type	N.	Prop.
Maybe dam.	Reg. No User Page	4,006,466	0.28
Maybe dam.	Reg. User Page	3,797,451	0.27
Maybe dam.	Unregistered	6,415,271	0.45
Maybe dam.	**Total**	14,219,188	1.00
Likely dam.	Reg. No User Page	281,964	0.12
Likely dam.	Reg. User Page	26,459	0.01
Likely dam.	Unregistered	1,982,985	0.87
Likely dam.	**Total**	2,291,408	1.00
V. likely dam.	Reg. No User Page	21,630	0.05
V. likely dam.	Reg. User Page	687	0.00
V. likely dam.	Unregistered	371,499	0.94
V. likely dam.	**Total**	393,816	1.00

(b) Counts and proportions of edits by whether an editor was registered or had a user page in the neighborhood of each threshold.

Table C.1: Summary statistics from our full dataset.

editors act on Wikipedia without logging in and registered contributors are those that edit with accounts. Because they are identified by their IP address rather than by a chosen username, unregistered editors are also referred to as "IP editors" or "anons." Unregistered editors are associated with misbehavior and have long had a controversial status on Wikipedia (McDonald et al., 2019). Geiger and Ribes described how tools for moderators highlight unregistered editors (Geiger & Ribes, 2010). De Laat argued that unregistered editors on Wikipedia are overprofiled in that they are at higher risk to have their contributions rejected unfairly (de Laat, 2015, 2016).

Second, the RCFilters interface indicates whether the editor has created a user page. User pages are Wikipedia's version of profile pages. Not having a user page is a social signal of newness because most committed users will create a user page early into their experience in Wikipedia (Ayers et al., 2008). The presence or absence of pages in Wikipedia is indicated with a subtle user interface clue: links to pages that do not exist are rendered in red, whereas links to pages that exist are blue. For example, Figure C.1 shows the user "Mashlova" whose name is shown in red and would be identified as a newcomer. De Laat cited the absence of a user page as a second example of an indicator of vandalism that will result in overprofiling (de Laat, 2016). We measure whether an editor's user page exists at the time of a given contribution by matching the titles of user pages against the editor's username and checking if the creation of the user page was prior to the edit in question. We only include registered editors in our analysis of overprofiling based on user pages.

C.5. Analytic plan

Our analysis comprises Bayesian logistic regression models in two parallel analyses. The first analysis treats our dichotomous measure of whether edits are reverted as an outcome. This begins with an "adoption check" (§C.6) that describes the causal effects of flagging on reverts in general. The adoption check is a prerequisite to answering our research questions. The rest of the first analysis (§C.7) answers RQ1 by comparing the effect of RCFilters on edits by

overprofiled users to its effect on other editors. Our second analysis is very similar but uses controversial reverts as the outcome, and analyzes only reverted edits to model the probability that a revert is controversial. It begins by answering RQ2 (§C.7) in an analysis similar to the adoption check but with controversial sanctions as an outcome and with a dataset limited to overprofiled users. The rest of the second analysis (§C.7) answers RQ3 and is similar to RQ1 but with controversial reverts as the outcome in place of reverts.

Although our models use different sets of edits and outcomes, they all have the same logistic regression structure shown in Equation C.4.

$$\log\left(\frac{P\left(Y_r\right)}{1-P\left(Y_r\right)}\right) = \alpha_1\left(score_r - c_{jw}\right) + \tau_j \mathbf{1}\left[score_r > c_{jw}\right] + \alpha_2\left(score_r - c_{jw}\right)\mathbf{1}\left[score_r > c_{jw}\right] + \alpha_w \tag{C.4}$$

Our goal is to estimate τ_j which is the causal effect of being flagged at level j, where $j \in \{1, 2, 3\}$ corresponds to labels of "maybe damaging," "likely damaging" and "very likely damaging." For each cutoff on each wiki, we select revisions whose ORES scores are within a ± 0.03 window of the cutoff (c_{jw}). Following established approaches to RDD, we fit "kink" models that allow for a change in slope at the discontinuity (Lee & Lemieux, 2010; Litschig & Morrison, 2013). The slope before the discontinuity is α_1 and the change in slope is α_2. The indicator function is represented by $\mathbf{1}$. Our models include fixed effects for wiki (α_w) to account for differences in the rates of sanctioning between wikis.

We use Bayesian inference to estimate our models for two reasons. First, virtually all edits above the "very damaging" level are reverted in some of the wikis we analyze. The presence of near-perfect "separation" creates estimation problems for classical numerical approaches (Allison, 2004). Preferred solutions to this problem in non-Bayesian frameworks include penalized likelihood methods that introduce bias. Our Bayesian approach uses weakly-informative priors that are conservative but avoid the problem of separation. The second reason we use Bayesian inference is that it makes it easy to compare estimates across models. Prior work at CSCW by Gan et al. (2018) used a similar rationale for adopting Bayesian logistic regression. In Bayesian analysis, fitted models take the form of posterior distributions constituting a probability distribution of model coefficients conditional on our model, data and priors. We consider a hypothesis supported if it is consistent with at least 95% of posterior draws. In other words, we accept a given hypothesis if our parameter estimate has the predicted sign and the 95% credible interval does not contain 0. This is the Bayesian analog to testing a hypothesis with $\alpha = 0.05$. We fit our models using the rstanarm package (version 2.19.3) and the default priors that are provided for reference in the supplementary material.

C.6. Adoption Check

Before presenting results from hypothesis tests associated with our research questions, we first establish that RCFilters was adopted by Wikipedia moderators and that it had an effect on sanctioning behavior. This establishes a baseline necessary to answer RQ1 regarding the differential effects of RCFilters between overprofiled users and others. Null effects in RQ1 might simply reflect that the system was not used. A successful adoption check rules out this possibility and sets up a credible null hypothesis test for RQ1.

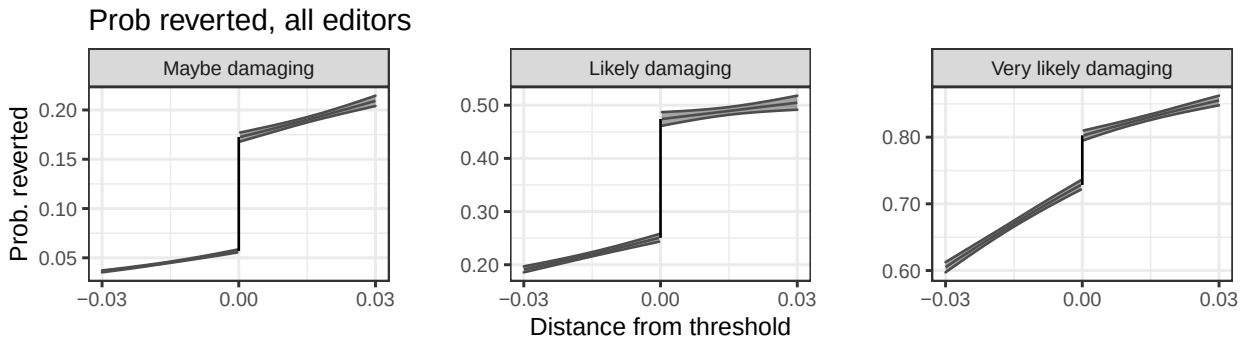


Figure C.2: Marginal effects plot showing model predicted relationship between ORES score and the probability that an edit will be reverted around the cutoffs for all contributors with 95% credible intervals.

We test the hypothesis that flagging increases the probability that an edit is reverted to demonstrate that RCFilters flags are being used by Wikipedia moderators. Our estimates for τ_j—as described in §C.5—should be positive if Wikipedia moderators are using flags in RCFilters to review potentially damaging edits.

We find strong evidence that RCFilters was adopted and impacted sanctioning. Figure C.2 visualizes this evidence: a marginal effects plot that illustrates our models' predicted likelihood of reverts across different ORES scores in the neighborhood of the thresholds. In each such plot, the x-axis shows the distance from the threshold such that discontinuities at `0` represent the effect of being flagged. The plots show modeled values for the English language edition of Wikipedia but are representative of relationships across all wikis.[3] Figure C.2 shows discontinuous increases in the likelihood of reversion at the "maybe damaging" and "likely damaging" thresholds in the left and center panels. We find the greatest effect at the "maybe damaging" threshold ($\tau_1 = 1.23$ $[1.19; 1.28]$).[4] The effect at the "very likely damaging" threshold shown in the right-most panel is smaller ($\tau_3 = 0.41$, $[0.35; 0.46]$).

The impacts of the "maybe damaging" and "likely damaging" flags on the likelihood of sanctioning are enormous. Figure C.2 shows that likelihood of a revert for an edit just below the "maybe damaging" threshold is between 5.5% and 5.8%, indicating that reverts of unflagged edits are relatively rare. Being flagged with the "maybe damaging" flag causes a dramatic increase in the reversion probability to between 16.8% and 17.7% for edits just above the threshold. The effect of algorithmic flags at the "likely damaging" level is even more stark. We estimate that edits just below the "likely damaging" threshold are likely to be reverted between 24.3% and 25.8% of the time, whereas similar edits just above the threshold are reverted between 46.1% and 48.7% of the time. Being flagged at the "very likely damaging" threshold causes an increase in reversion probability from between 72.1% and 73.5% to between 79.5% and 81%.

[3]Because intercepts are the only part of our model that depend on Wikis, slopes and the discontinuities caused by algorithmic flagging represent our inference over all our data.

[4]All τ parameter estimates are reported as log-odds ratios. The bracket notation indicates the 95% credible interval. In other words, the most likely value of the parameter is 1.23, but there is a 95% probability that the parameter lies in the interval $[1.19; 1.28]$.

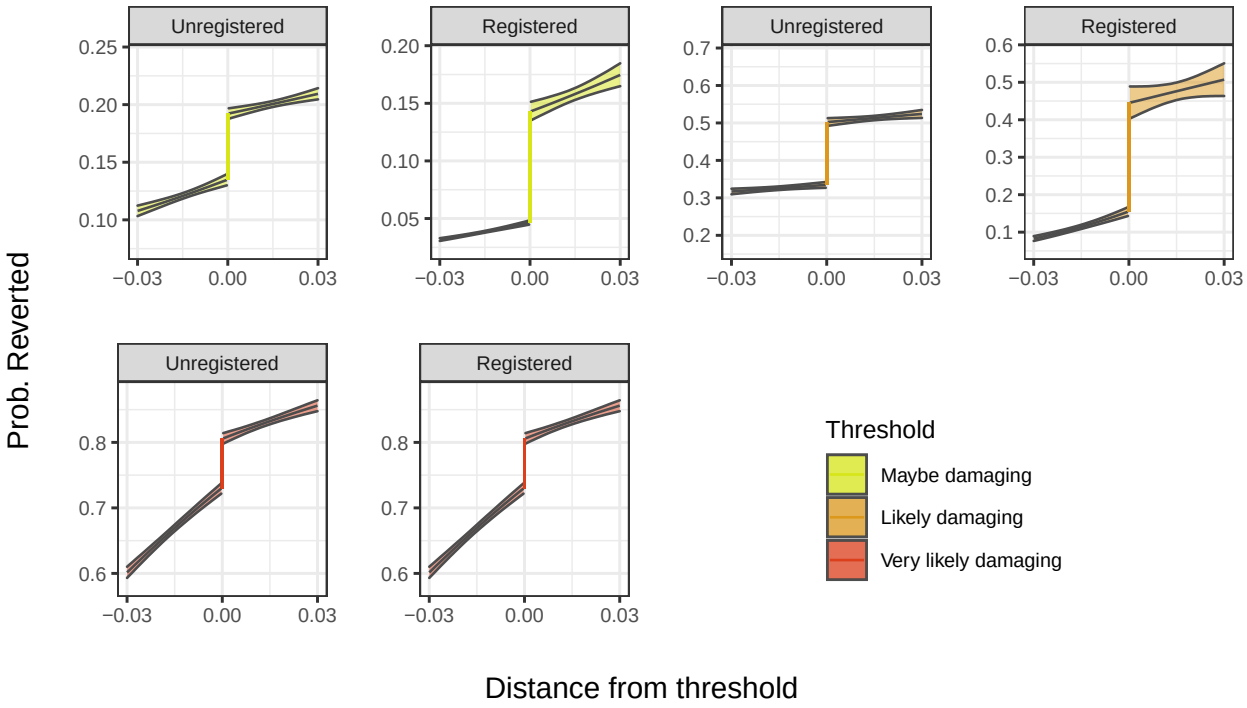


Figure C.3: Results for RQ1 comparing unregistered and registered contributors are displayed in a marginal effects plot showing the model predicted relationship with 95% credible intervals between ORES scores and reverts around the thresholds that trigger flags.

C.7. Results

RQ1: Effect of Flagging on Sanctioning

In our first research question (RQ1), we seek to understand how the increase in sanctioning caused by flagging affects discrimination against overprofiled users. If algorithmic flagging reduces overprofiling, as some computer scientists have argued (Kleinberg et al., 2018), the effect of flagging will be more scrutiny on users who are more likely to be given a pass. If algorithms simply reproduce discrimination, we will find no difference. Results for hypothesis tests answering this question are shown in Figure C.4, which visualizes the point estimates and credible intervals for differences in the causal effects of flagging on reverts between unregistered and registered contributors and between contributors with and without user pages. Values greater than 0 indicate that our estimated effect for the other users is greater than that for the overprofiled group.

In support of the idea that algorithmic flagging can reduce overprofiling bias, we find that the overall effect of flagging is to increase demographic parity between registered and unregistered editors. Aggregating our posteriors over all three thresholds shows that the average effect over the three thresholds is greater for registered editors than for unregistered editors ($\frac{1}{3}\sum_{j=1}^{3}\tau_j^{\text{Reg}} - \tau_j^{\text{Unreg}} = 0.45\ [0.16; 0.6]$). The effect of flagging on reverts of registered editors is greater than the effect for unregistered editors at both the "maybe damaging" threshold ($\tau_1^{\text{Unreg}} - \tau_1^{\text{Reg}} = 0.8\ [0.71; 0.89]$) and the "likely damaging" threshold ($\tau_2^{\text{Unreg}} - \tau_2^{\text{Reg}} = 0.78\ [0.58; 0.97]$). For an action by an unregistered contributor near to the "maybe damaging" threshold,

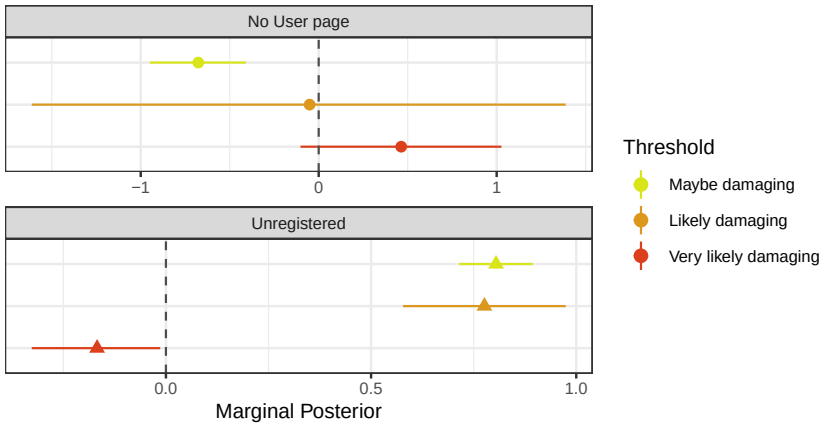


Figure C.4: Results for RQ1 showing point estimates and 95% credible intervals for differences in the causal effect of flagging on sanctioning between overprofiled contributors and others. A value greater than 0 indicates that our estimates of the effect for underprofiled contributors are greater than those for overprofiled contributors.

being flagged increases the odds of being reverted by a factor of between 1.45 and 1.6 times. This is significantly less than the increase of between 3.16 and 3.68 times for registered contributors.

However, at the "very likely damaging" threshold we find that the effects of flagging are stronger for unregistered editors than for registered editors ($\tau_2^{\text{Reg}} - \tau_2^{\text{Unreg}} = -0.17$ $[-0.33; -0.01]$). Being flagged increases the odds that an action is reverted by a factor of between 1.43 and 1.62 times for an unregistered editor and by 1.11 and 1.49 times for registered contributors. However, as Table C.1 shows, a far greater number of actions receive scores near to lower thresholds. Thus, we focus on the lower thresholds in the following discussion.

Figure C.3 lets us interpret our models by making it possible to visually compare the effects of being flagged between overprofiled and underprofiled editors at a given threshold because the y-axes in each row span an identical range. The top-left panel shows how our models' linear predictions of how the probability of sanctioning for unregistered contributors at the "maybe damaging" threshold jumps between 4.8 and 6.7 percentage points, from 13.5% to 19.2% on average. For registered editors, shown in the top-right of Figure C.3, we estimate a jump of between 9.1 and 10.3 percentage points, from 4.6% to 14.3% on average. This is between 3.3 and 4.6 percentage points greater than the jump for unregistered editors. For unflagged edits that ORES scores near the "maybe damaging" threshold, an unflagged unregistered contributor has about the same odds of being sanctioned as a flagged registered contributor.

The bottom row of Figure C.3 shows that the change in sanctioning probability at the "likely damaging" threshold is between 9.5 and 15.2 percentage points greater for registered editors than for unregistered editors. For unregistered contributors, shown in the bottom-left of Figure C.3, being flagged as "likely damaging" increases the probability of revert between 15 and 18.6 percentage points, from 33.5% to 50.2% on average. But for registered editors, shown in the bottom-right of Figure C.3, we detect an even bigger jump of between 23.7 and 34.6 percentage points, from 15.5% to 44.5% on average. For actions that ORES scores near the "likely damaging" threshold, unflagged actions by unregistered editors are far more likely to be reverted. Once flagged, actions by registered and unregistered editors are reverted at relatively similar rates.

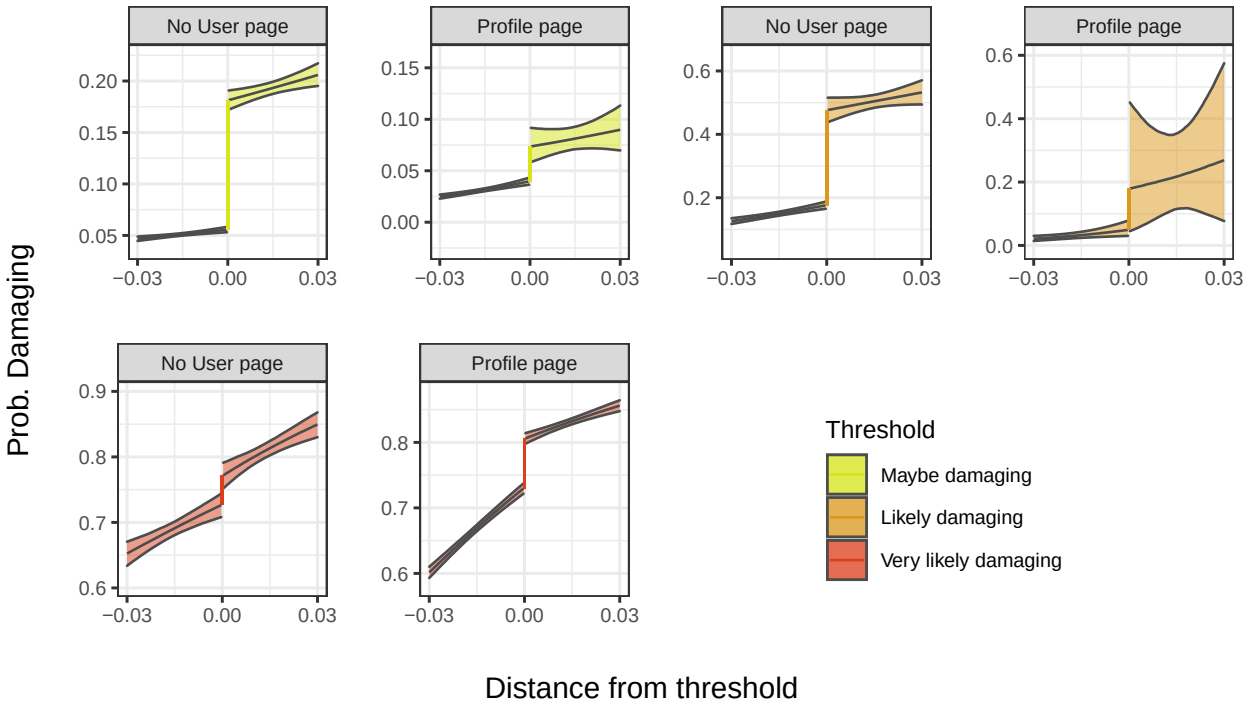
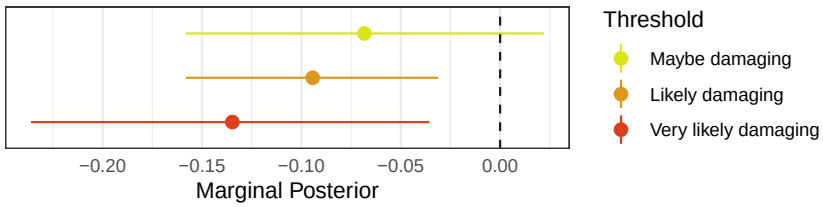


Figure C.5: Results for RQ1 comparing contributors with and without user pages. Each panel shows a marginal effects plot with 95% credible intervals of the modeled relationship between ORES scores and reverts around the thresholds that trigger flags.

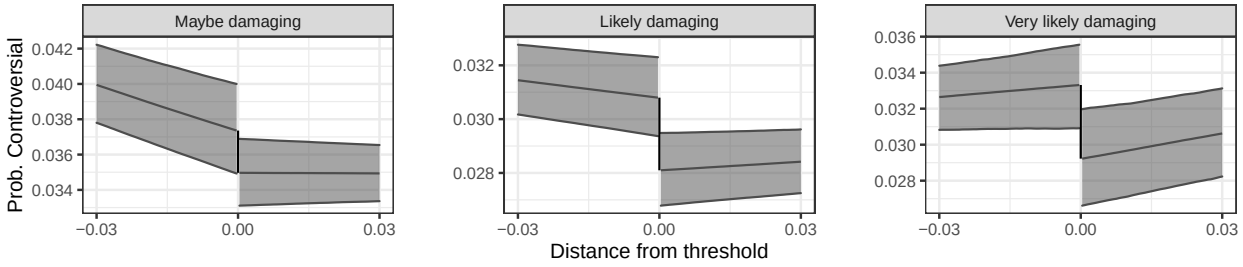
These results show that flagging causes an increase in a decision system's demographic parity concerning registration status. Actions by unregistered contributors that fall just above the cutoffs are much more likely to be reverted due to RCFilters—but the gap between actions by registered and unregistered contributors is much smaller when RCFilters has flagged an edit as "maybe damaging" or "likely damaging." In this way, our analysis suggests that algorithmic flagging can reduce overprofiling bias.

Surprisingly, our results for our second measure of over-profiling in Wikipedia suggest a dynamic that is opposite in sign to the differences we observe between registered and unregistered editors at the "maybe damaging" threshold ($\tau_1^{\text{NoUP}} - \tau_1^{\text{UP}} = -0.68\ [-0.95; -0.41]$). At the "likely damaging" ($\tau_2^{\text{NoUP}} - \tau_2^{\text{UP}} = -0.05\ [-1.61; 1.39]$) and the "very likely damaging" ($\tau_2^{\text{NoUP}} - \tau_2^{\text{UP}} = 0.46\ [-0.1; 1.03]$) thresholds, we do not detect differences in effect size between contributors with and without user pages. At the "maybe damaging" threshold, we find that flagging increases the odds that an editor without a user page is reverted between 3.47 and 4.06 times. This is significantly more than the increase of between 1.47 and 2.46 times for registered contributors.

As above, we interpret these odds ratios using marginal effects plots shown in Figure C.5. The top-left plot in the figure shows our models' linear predictions of the probability of reverting for contributors without user pages near to the "maybe damaging" threshold. For these editors, being flagged as "maybe damaging" increases the chances of sanctioning by 11.4 and 13.8 percentage points, from 5.6% to 18.1% on average. In the top-right of Figure C.5, we see a jump of between 2.2 and 4.8 percentage points, from 4% to 7.4% on average for editors that



(a) Parameter estimates and 95% credible intervals for the effects of flagging on whether reverts are controversial for unregistered editors.



(b) Marginal effects plots with 95% credible intervals for models predicting whether a revert is controversial, for unregistered editors.

Figure C.6: Results for RQ2: flagging causes a small but detectable decrease in the likelihood that an action by an unregistered contributor receives a controversial sanction.

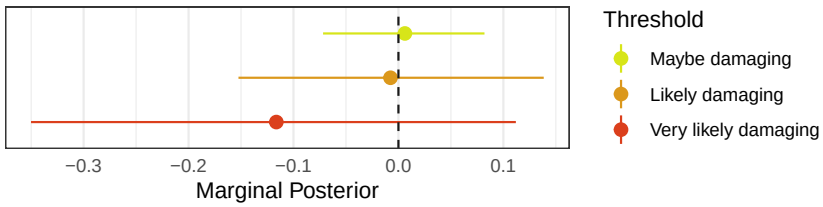
have created user pages. This is between 9.7 and 8.4 percentage points less than the jump for contributors without user pages.

RQ2: Effect of flagging on controversial sanctioning

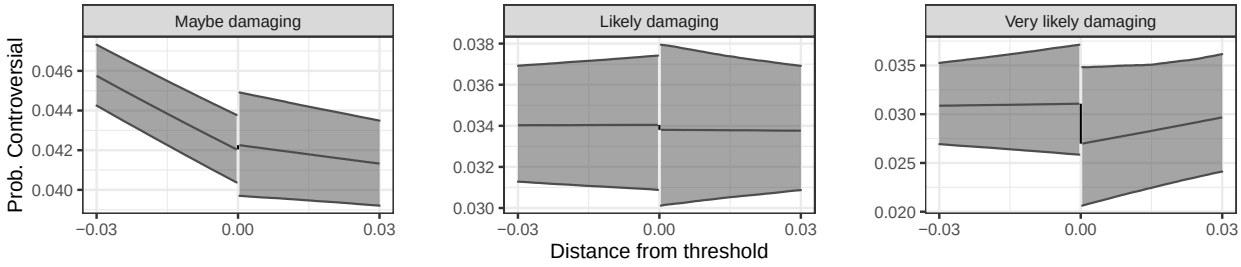
Consistent with the idea that algorithmic flagging can support fairness, we find that having an ORES score cross the "likely damaging" or "very likely damaging" thresholds decreases the chances that a revert will be controversial for unregistered editors. These results are visualized in Figure C.6a. We have less confidence in the effect at the "maybe damaging" threshold because our 95% credible interval includes 0 ($\tau_1^{\text{Unreg}} = -0.07; \text{CI} = [-0.16; 0.02]$).

We estimate that being flagged at the "likely damaging" level results in a change in the odds that a sanction is controversial by a factor between 0.85 and 0.97. Figure C.6b shows the modeled relationship between ORES scores and the probability of a controversial sanction in the neighborhood of the thresholds for English Wikipedia. On the left plot, we see that being flagged changes unregistered contributor's likelihood of a controversial revert from a possible increase of 0.27 percentage points to a possible decrease of 0.55 percentage points, a change from 3.08% to 2.81% on average.

We observe a similar effect of flagging at the "very likely damaging" threshold ($\tau_2^{\text{Unreg}} = -0.13; \text{CI} = [-0.24; -0.04]$): the odds that a revert is controversial are between 0.79 and 0.97 times smaller. On the right side of Figure C.6b, we find that being flagged decreases the probability that a sanction to an action by an unregistered editor is controversial by between 0.11 and 0.89 percentage points, a change from from 3.33% to 2.92% on average.



(a) Parameter estimates and 95% credible intervals for effects of flagging on whether reverts are controversial for editors without user pages.



(b) Marginal effects plots with 95% credible intervals for models predicting whether a revert is controversial, for contributors without user pages.

Figure C.7: Results for RQ2 comparing contributors with user pages to those without show no detectable effect of flagging on controversial sanctioning.

However, we did not detect effects of flagging when the reverted editor lacks a user page at the "maybe damaging" ($\tau_1^{\text{NoUP}} = 0.01; \text{CI} = [-0.07; 0.08]$), "likely damaging" ($\tau_2^{\text{NoUP}} = -0.01; \text{CI} = [-0.15; 0.14]$), or "very likely damaging" ($\tau_3^{\text{NoUP}} = -0.12; \text{CI} = [-0.35; 0.11]$) thresholds. Our results for RQ2 for unregistered editors show that flagging decreases the rate of controversial sanctions. Although controversial sanctions do not precisely correspond to false-positive sanctions, we take this finding as evidence that flagging decreases the false positive rate of the decision system. We address the inconsistencies between our results for unregistered editors and editors without user pages in our discussion (§C.9).

RQ3: Social signals and effects of flagging on controversial sanctioning

To answer RQ3, we largely replicate the analysis conducted for RQ1 with the dependent variable used in RQ2. Results shown in Figure C.8 provide weak evidence that a decrease in controversial sanctioning may be greater for registered than for unregistered contributors at the "maybe damaging" ($\tau_1^{\text{Reg}} - \tau_1^{\text{Unreg}} = 0.04\ [-0.06; 0.14]$), "likely damaging" ($\tau_2^{\text{Reg}} - \tau_2^{\text{Unreg}} = 0.07\ [-0.05; 0.2]$), and "very likely damaging" ($\tau_3^{\text{Reg}} - \tau_3^{\text{Unreg}} = 0.02\ [-0.23; 0.27]$) thresholds. However, our evidence weakly suggests that the effect for contributors with user profiles is greater than those for without at the "maybe damaging" threshold ($\tau_1^{\text{UP}} - \tau_1^{\text{NoUP}} = 0.05\ [-0.08; 0.17]$) but the opposite seems true at the "likely damaging" threshold ($\tau_2^{\text{UP}} - \tau_2^{\text{NoUP}} = -0.26\ [-0.79; 0.26]$) and "very likely damaging" ($\tau_3^{\text{UP}} - \tau_3^{\text{NoUP}} = -0.16\ [-0.9; 0.56]$) thresholds. None of these estimates are statistically significant at the 95% level.

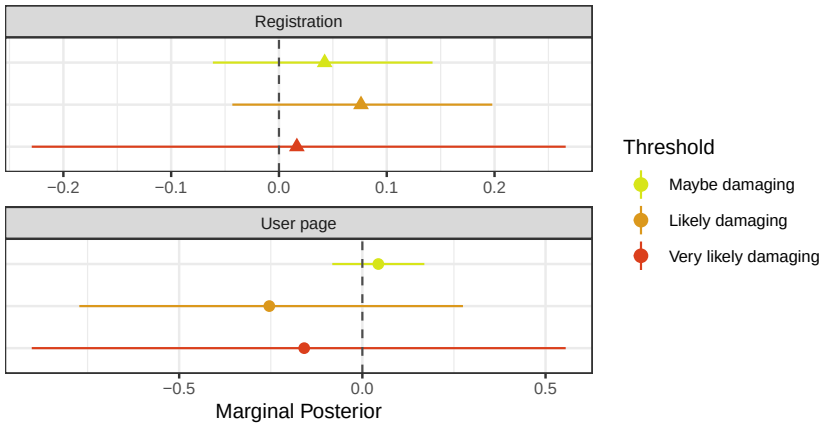


Figure C.8: Results for RQ3 showing the difference in our parameter estimates between overprofiled editors and others with 95% credible intervals. Values greater than 0 would indicate that the effect for underprofiled editors is greater than that for overprofiled editors.

C.8. Threats to Validity

Our results are subject to a range of threats to validity that pertain to our ability to make causal claims, rule out alternative explanations, and establish the generalizability of our findings. First, there are several threats to our ability to draw causal inferences that are common to RDDs. Formally, RDDs model an outcome Y as a function of a continuous "forcing variable" Z, other covariates, and a cutoff c such that $Z > c$ determines treatment assignment. In principle, treatment assignment conditional on Z is "as good as random" under two assumptions: (1) that agents have at most limited control over $Z > c$, and (2) that the relationship between Y and Z is smooth (Lee & Lemieux, 2010). Although the assumptions required for causal inference are fundamentally unverifiable, we believe that our RDD provides relatively strong evidence of causal relationships between flagging and sanctioning.

Our treatment, being flagged in RCFilters, is an ideal candidate for an RDD from the perspective of assumption (1) because editors are unlikely to have much control over the scores that their edits receive. Although attempts to evade sanction by specially crafting edits to evade algorithmic detection are hypothetically possible, the authors of ORES and RCFilters believe they are unrealistic and very unlikely to be widespread. Assumption (2) would be violated if any unobserved treatments affect our outcomes at discrete levels of ORES scores. This is certainly possible because ORES makes scores available via a public API. Indeed, we are aware of bots that automatically revert edits triggered by the "very damaging" threshold on some of the Wikipedia language editions in our sample and therefore have more reason to doubt results at this threshold. Despite this threat, our conclusions regarding how algorithmic flagging shapes fairness are substantively similar whether we consider this threshold or not. Although we identified one anti-vandalism tool—a system called Huggle discussed in §C.9—that collects ORES damaging scores, it uses ORES scores as one feature in its own algorithmic model and, by default, presents predictions from this model to users as a list of edits sorted in order of likelihood of vandalism. Given these facts, we believe that it is unlikely that Huggle users will drive discontinuities in the relationship between ORES scores and our outcomes.

A limitation of RDD analysis is that it estimates effects for observations in the neighborhood

of the cutoff and results may not generalize far away from the cutoff. Compared with most RDD analysis, ours has the advantage of multiple different thresholds. Although our results for the "likely damaging" and "maybe damaging" thresholds are substantively similar, causal effects may diverge more at operating points we have not considered. Future work on algorithmic bias using RDD should consider that results may depend on the choice of operating points used as RDD cutoffs.

An additional threat to validity is raised by the extent to which the ORES models encode biases concerning editors who are unregistered or without profile pages. To assess this threat, we analyzed the bias of ORES models for each wiki that had deployed the system on December 19th 2020 using their human-labeled training data according to the *conditional calibration* approach to evaluating model bias (Mitchell et al., 2020).[5] In our case, this involves comparing the rate of damaging edits predicted by the model to the true rates for each type of editor. We find that ORES exhibits bias against both unregistered editors and editors without user pages but that the extent of bias against unregistered editors is much greater. These findings are opposite in sign to what we would expect if model bias were driving our results. We present detailed results from this analysis in our online supplement.

Our study design is also limited in that we cannot present causal evidence of the impact of social signals. Although RCFilters's algorithmic flags are distributed in a quasi-experimental way, overprofiled status is not. There are a range of possible systematic differences between overprofiled users and others that might be driving our results for RQ1 and RQ3. For example, if damaging edits by contributors who are unregistered or lack user pages are more difficult for ORES to detect, that might drive our findings of a decrease in overprofiling for RQ1. Although we believe that this particular threat is unlikely because it would require that overprofiled contributors be systematically more sophisticated than others—something our experience with ORES suggests is unlikely—we cannot rule out either the specific threat or a range of other possibilities. A promising direction for future work might involve experiments or quasi-experiments that can jointly vary social signals and algorithmic flagging.

Additionally, system designers will likely want to know how overall rates of sanctioning and controversial sanctions change before and after a system such as RCFilters is launched. Unfortunately, our analysis cannot answer this question directly. In preliminary work, we attempted to draw a statistical comparison between Wikipedia governance before and after the introduction of ORES but high temporal variation in sanctioning behaviors made this type of aggregate change difficult to measure. Future studies should organize with communities to conduct planned and principled field experiments to study the causal effects of introducing such systems in online communities using the model being pioneered by Matias and Mou (2018).

Finally, a set of largely unanswerable threats involves questions of generalizability across our measures and empirical contexts. Although our theory of interactions between algorithmic flags and social signals is general, and although we study RCFilters across 23 distinct communities, languages, and cultures, we study a single moderator tool on one platform. We cannot claim that our findings generalize beyond the specific pool of communities that we study. Additionally, we have considered only a small subset of possible social signals that may be used in online community moderation. Clearly, we also cannot claim that our settings are representa-

[5]We chose conditional calibration as our fairness metric because it does not depend on the choice of threshold. This simplifies the analysis of a decision system with multiple thresholds.

tive of moderation in online communities in general. Like most other empirical studies in social computing, we must sadly leave these questions for further research.

C.9. Discussion

In the broadest strokes, our work is potentially good news for advocates of algorithmic flagging in social computing systems. It provides some evidence supporting the idea that algorithmic flagging can reduce discrimination in the form of overprofiling bias and that it can increase fairness. Our adoption check (§C.6) provides strong evidence that RCFilters drives behavior and our answers to RQ1 (§C.7) suggests that flagging can level the playing field by increasing decision system demographic parity between unregistered and registered Wikipedia editors. Flagged edits by these contributors are reverted at similar rates, but unflagged edits of comparable quality by registered editors are reverted relatively infrequently. More good news comes in the form of our answer to RQ2 (§C.7) that suggests that flagging is associated with a decrease in controversial sanctions among some overprofiled users and provides evidence that algorithmic flagging systems can help moderators more accurately issue sanctions.

When it comes to the details, however, the picture that emerges from our results is much more contingent and mixed. Our analysis used two different measures of overprofiling in Wikipedia but the pattern of our results diverged substantially between the two. The optimistic story about the effects of algorithmic flagging on overprofiled users only describes our results for unregistered Wikipedia users. Our evidence on overprofiled users without user pages is much weaker and points, in part, in the direction of algorithmic flagging increasing discrimination. Why do these results diverge? What do these divergent results mean for theory?

One possible explanation is that editors without user pages are, quite simply, not particularly overprofiled. Of the two social signals we consider, registration status attracts far more attention from academics and community members in discussions of Wikipedia vandalism (e.g., Hill & Shaw, 2020). Our analysis for RQ2, where we did not detect changes in controversial sanctions for editors without user pages, is also consistent with the notion that contributors without user pages may not be overprofiled. If algorithmic flagging systems help moderators more accurately issue sanctions by reducing overprofiling, then flagging would not decrease controversial sanctioning for editors that are not overprofiled. However, this alone does not explain why the effect for editors without profile pages was larger than for editors with them.

Our results might be explained if model bias against contributors without user pages means that the set of flagged edits from these users are less damaging than flagged edits by contributors who have profile pages. As discussed in §C.8 and documented in our online supplement, ORES models are sometimes biased against contributors without user pages, but they are even more biased against anonymous contributors. Our results make sense if the overprofiling of anonymous editors outweighs model bias against them, but the reverse is true for editors without user pages.

It is also plausible that our mixed results are evidence that algorithmic flags will substitute for some social signals used in overprofiling while reinforcing others. Our study analyzes only two of many possible social signals that online community moderators might use. A better understanding of which signals drive sanctioning misbehavior can help explain if and when algorithmic triage systems can increase fairness. Our results suggest that algorithmic flags can substitute for some social signals and reduce overprofiling in online community moderation.

Our results also suggest that they might reinforce social signals, make overprofiling worse, or introduce new forms of unfairness through encoded bias. Unfortunately, outcomes resulting from myriad factors acting at once are likely contingent on details of sociotechnical arrangements and difficult to know *ex ante*.

Although RQ2 suggests that algorithmic flagging can increase fairness for overprofiled contributors, our null results for RQ3 mean that we could not detect a difference in this effect between overprofiled editors and others. Uncertainty in our models for RQ3 is high enough that parameter values consistent with a substantive average effect that is either positive or negative are plausible. A null effect for RQ3 might also be explained if meta-norms and improved information are more important to controversial sanctioning than bias introduced by algorithmic flags or social signals acting as cues.

Our work has several important implications for designers of algorithmic flagging systems and sociotechnical systems. Scholars of human computer interaction, science and technology studies, and the law have all called for analyses of algorithmic fairness to move beyond biases inherent in algorithms to consider the systemic and downstream effects of algorithms in use (Selbst et al., 2019; Stevenson, 2017; Zhu et al., 2018). Ultimately, we recommend that operators of algorithmic flagging systems should continuously evaluate decision system fairness metrics and seek to improve them according to their values. In that the ORES model is, itself, biased against overprofiled users, our results suggest that evaluating the fairness of model predictions is only one piece of understanding how an algorithmic system shapes fairness in contexts such as online community moderation.

Future work should rigorously construct and critique decision system fairness criteria in terms of their consequences. The algorithmic fairness literature often treats algorithmic predictions as equivalent to final decisions. Our work shows that sociotechnical decision systems with humans in the loop face distinctive and contextually sensitive epistemic, ontological, and ethical questions about how decision system fairness should be defined or measured (Kleinberg et al., 2018; Selbst et al., 2019).

Decision system fairness is particularly important in open production communities such as Wikipedia because of the trade-offs between quality control and the essential tasks of supporting newcomers and encouraging contribution (Halfaker et al., 2013; Morgan et al., 2013). Past work has shown that increased quality control efforts correspond to a decrease in newcomer engagement and have hypothesized that one mechanism is increased scrutiny of newcomers (Halfaker et al., 2013; TeBlunthuis et al., 2018). Similarly, although blocking anonymous edits to wikis has shown been shown to cause a decrease in reverted edits, it also leads to a decrease in positive contributions (Hill & Shaw, 2020). While it may be intuitive to think about edits that get sanctioned as obvious vandalism, many of the edits flagged at the "maybe damaging" threshold are authored by well-meaning newcomers (Halfaker et al., 2013). There's a potentially high cost to sanctioning these low quality but well-intentioned contributions. We believe that our results point to the benefit of tracking changes in the rate of sanctions to sensitive groups of community members in order to assure that such well-meaning contributors are not being driven away.

There are also lessons to learned from the impressive degree with which RCFilters shapes behavior. Although the choice of operating points in algorithmic systems is often framed as purely about trading off precision and recall, our work demonstrates that these choices can have a range of other important consequences. Our disparate findings at the "very likely damaging"

threshold for overprofiling based on registration status reveal that an algorithmic tool might improve fairness at a given operating point but decrease it at another. Although thresholds allowed us to explore the effects of flagging on sanctioning behavior, this arbitrary flagging of actions applied by RCFilters brought disproportionate attention to contributions just above the thresholds compared to contributions just below. Designers should think about whether using thresholds to trigger flagging in moderation interfaces is a fair practice at all. Our results show that this leads to sanctioning behavior that is, like the thresholds, arbitrary.

What types of designs might support quality control support models that scrutinize contributions in proportion to the likelihood that the contributions deserve to be sanctioned? We see some inspiration in Huggle, a counter-vandalism tool for Wikipedia which sorts actions by the likelihood that they are damaging.[6] Huggle users are encouraged to review the highest likelihood edits first and only move onto lower likelihood edits once those reviews are complete. Such a user experience might increase efficiency and fairness by better concentrating moderator attention wherever it can have the greatest benefits.

C.10. Conclusion

As algorithmic flagging becomes more integrated into online community moderation, it is important to understand its effects and consequences on overprofiling and fairness. We use a regression discontinuity analysis of the RCFilters to find and sanction misbehavior by volunteers on Wikipedia to consider how the use of algorithmic flagging and social signals interact. We find that by drawing moderator attention to misbehavior by registered participants, algorithmic flagging can reduce overprofiling in certain contexts. We also find that algorithmic flagging can support fairness by decreasing controversial sanctions of unregistered contributors. Our results also suggest that the same system may have much less effect, and might even increase discrimination, for other types of overprofiled users.

Studies of machine learning in high-stakes settings like employment, education, and criminal justice trace how algorithms can encode discriminatory patterns in human behavior but might also improve fairness compared with human biases. Although the stakes are much lower, such questions are also pertinent to the use of machine predictions for online community moderation. We find that tools for predictive governance in a sociotechnical system can reduce overprofiling but their effects are also difficult to anticipate.

Although our analysis of overprofiling based on registration status supports a rosy account of algorithmic flagging, our analysis of overprofiling based on user pages does not. While contributors without user pages may be less overprofiled compared to unregistered contributors, our results also suggest that the interaction between algorithmic flagging and social signals is complex and contingent. We suggest a need for future work that describes the kinds of social signals that are used in practice and explains how different types of information may be used alongside algorithmic flags. Finally, we present a methodological approach that we hope future studies of algorithmic tools in real-world sociotechnical systems might build upon to establish the causal effects of algorithmic systems without experimental intervention.

[6] See discussion in (Halfaker et al., 2014)

acknowldgements

We are grateful to the anonymous CSCW reviewers and associate chairs for their keen insights and feedback. We would also like to thank the Wikimedia Foundation for its support, members of the WMF analytics team including Andrew Otto, Luca Toscano, and Joal Allemandou for help with data access and computing infrastructure and members of the WMF research team including Jonathan Morgan for feedback early in project development. Thanks also go to members of the Community Data Science Collective who provided multiple rounds of feedback and contributed to copyediting including Kaylea Champion, Charles Kiene, Stefania Druga, Sohyeon Hwang, Jeremy Foote, and Aaron Shaw. We also thank the WMF staff and volunteers who developed the systems we analyze including Roan Kattouw, the main developer of RCFilters, and the developers of ORES including Amir Sarabadani and Andy Craze. Special thanks to Amanda TeBlunthuis. Finally we owe an extra special thanks to the Wikipedia contributors whose digital traces we analyze. Portions of this work were facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington. Financial support for this work came from the Wikimedia Foundation, from the National Science Foundation graduate research fellowship program #2016220885, and from the University of Washington.

Data Access

A replication dataset including ORES scores, thresholds, and our sample of Wikipedia revisions, along with all of our code has been placed in the Harvard Dataverse archive and is available at the following URL: https://doi.org/10.7910/DVN/E0RYJ4

References

Adler, B. T., & de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. *Proceedings of the 16th International Conference on World Wide Web*, 261–270.

Allison, P. (2004). Convergence Problems in Logistic Regression. *Numerical Issues in Statistical Computing for the Social Scientist* (pp. 238–252). John Wiley & Sons, Ltd.

Ayers, P., Matthews, C., & Yates, B. (2008). *How Wikipedia works and how you can be a part of it*. No Starch Press.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness in Machine Learning*. fairmlbook.org.

Barocas, S., & Nissenbaum, H. (2015). Big Data's End Run around Anonymity and Consent. In J. I. Lane (Ed.), *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press
OCLC: 882939943.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago Press
OCLC: 859759499.

Bernstein, M. S., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., & Vargas, G. (2011). 4chan and /b/: An analysis of anonymity and ephemerality in a large online community. *Fifth International AAAI Conference on Weblogs and Social Media*.

Bertrand, M., & Duflo, E. (2016). *Field Experiments on Discrimination* (tech. rep. w22014). National Bureau of Economic Research. Cambridge, MA.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Salience Theory of Choice Under Risk. *The Quarterly Journal of Economics*, *127*(3), 1243–1285.

Broughton, J. (2008). *Wikipedia the missing manual*. Pogue Press/O'Reilly
OCLC: 708321411.

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

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI now 2017 report. *AI Now Institute at New York University*.

Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019). 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.

Chandrasekharan, E., Gandhi, C., Mustelier, M. W., & Gilbert, E. (2019). Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–30.

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, *94*, S95–S120.

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428.

de Laat, P. B. (2015). The use of software tools and autonomous bots against vandalism: Eroding Wikipedia's moral order? *Ethics and Information Technology*, *17*(3), 175–188.

de Laat, P. B. (2016). Profiling vandalism in Wikipedia: A Schauerian approach to justification. *Ethics and Information Technology*, *18*(2), 131–148.

Donath, J. (2007). Signals in Social Supernets. *Journal of Computer-Mediated Communication*, *13*(1), 231–251.

Donath, J. (2014). *The social machine: Designs for living online*
OCLC: 1139880278.

Dubrovsky, V. J., Kiesler, S., & Sethna, B. N. (1991). The Equalization Phenomenon: Status Effects in Computer-Mediated and Face-to-Face Decision-Making Groups. *Human–Computer Interaction*, *6*(2), 119–146
_eprint: https://www.tandfonline.com/doi/pdf/10.1207/s15327051hci0602_2.

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication*, *11*(2), 415–441.

Ellison, N. B., Steinfield, C., & Lampe, C. (2011). Connection strategies: Social capital implications of Facebook-enabled communication practices. *New Media & Society*, *13*(6), 873–892.

Frey, S., Krafft, P. M., & Keegan, B. C. (2019). "This place does what it was built for": Designing digital institutions for participatory change. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 32:1–32:31.

Friedman, E. J., & Resnick, P. (2001). The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, *10*(2), 173–199.

Gan, E. F., Hill, B. M., & Dasgupta, S. (2018). Gender, feedback, and learners' decisions to share their creative computing projects. *Proceedings of the ACM on Human-Computer Interaction*, *2*, 54:1–54:23.

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

Geiger, R. S., & Ribes, D. (2010). The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 117–126.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press
OCLC: on1005113962.

Grabner-Kräuter, S., & Bitter, S. (2015). Trust in online social networks: A multifaceted perspective. *Forum for Social Economics*, *44*(1), 48–68
_eprint: https://doi.org/10.1080/07360932.2013.781517.

Grimmelmann, J. (2015). The Virtues of Moderation. *Yale Journal of Law and Technology*, *17*, 42–109.

Halfaker, A., & Geiger, R. S. (2020). ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *4*(148), 37.

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

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

Halfaker, A., Kittur, A., & Riedl, J. (2011). Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*, 163–172.

Hancock, J. T., & Dunham, P. J. (2001). Impression Formation in Computer-Mediated Communication Revisited: An Analysis of the Breadth and Intensity of Impressions. *Communication Research*, *28*(3), 325–347.

Hara, N., Shachaf, P., & Hew, K. F. (2010). Cross-cultural analysis of the Wikipedia community. *Journal of the American Society for Information Science and Technology*, *61*(10), 2097–2108.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]*.

Hecht, B., & Gergle, D. (2010). The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 291–300.

Herring, S. C. (2000). Gender differences in CMC: Findings and implications. *Computer Professionals for Social Responsibility Journal*, *18*(1), 0.

Hill, B. M., & Shaw, A. (2020). The Hidden Costs of Requiring Accounts: Quasi-Experimental Evidence from Peer Production. *Communication Research*, 30.

Horne, C. (2001). The Enforcement of Norms: Group Cohesion and Meta-Norms. *Social Psychology Quarterly*, *64*(3), 253–266.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Jacob, R. T., Zhu, P., Somers, M.-A., & Bloom, H. (2012). A practical guide to regression discontinuity. *MDRC Working Papers on Research Methodology*.

Jacobson, D. (1999). Impression Formation in Cyberspace: Online Expectations and Offline Experiences in Text-based Virtual Communities. *Journal of Computer-Mediated Communication*, *5*(1).

Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019). "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 192:1–192:33.

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Trans. Comput.-Hum. Interact.*, *26*(5), 31:1–31:35.

Jiang, J. ", Kiene, C., Middler, S., Brubaker, J. R., & Fiesler, C. (2019). Moderation challenges in voice-based online communities on Discord. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 23.

Jouhki, J., Lauk, E., Penttinen, M., Sormanen, N., & Uskali, T. (2016). Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics. *Media and Communication*, *4*(4), 75–85.

Kiene, C., & Hill, B. M. (2020). Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Kiene, C., Jiang, J. ", & Hill, B. M. (2019). Technological frames and user innovation: Exploring technological change in community moderation teams. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 44:1–44:23.

Kiene, C., Monroy-Hernández, A., & Hill, B. M. (2016). Surviving an "Eternal September": How an online community managed a surge of newcomers. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1152–1156.

Kiesler, S. E., Kraut, R. E., Resnick, P., & Kittur, A. (2012). Regulating behavior in online communities. In R. E. Kraut & P. Resnick (Eds.), *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*.

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

Kreiss, D., Finn, M., & Turner, F. (2011). The limits of peer production: Some reminders from Max Weber for the network society. *New Media & Society*, *13*(2), 243–259.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4066–4076). Curran Associates, Inc.

Lam, S. (K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., & Riedl, J. (2011). WP:clubhouse?: An Exploration of Wikipedia's Gender Imbalance. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 1–10.

Lampe, C. (2012). The Role of Reputation Systems in Managing Online Communities. In H. Masum & M. Tovey (Eds.), *The Reputation Society*. The MIT Press.

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 543–550.

Lampe, C. A., Ellison, N., & Steinfield, C. (2007). A familiar face(book): Profile elements as signals in an online social network. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 435–444.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Lindebaum, D., Vesa, M., & den Hond, F. (2019). Insights From "The Machine Stops" to Better Understand Rational Assumptions in Algorithmic Decision Making and Its Implications for Organizations. *Academy of Management Review*, *45*(1), 247–263.

Litschig, S., & Morrison, K. M. (2013). The Impact of Intergovernmental Transfers on Education Outcomes and Poverty Reduction. *American Economic Journal: Applied Economics*, *5*(4), 206–240.

Ma, X., Hancock, J. T., Lim Mingjie, K., & Naaman, M. (2017). Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2397–2409.

Marlow, J., Dabbish, L., & Herbsleb, J. (2013). Impression formation in online peer production: Activity traces and personal profiles in github. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 117–128.

Matias, J. N. (2016). Going dark: Social factors in collective action against platform operators in the Reddit blackout. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 1138–1151.

Matias, J. N., & Mou, M. (2018). Civilservant: Community-led experiments in platform governance. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 9:1–9:13.

McDonald, N., Hill, B. M., Greenstadt, R., & Forte, A. (2019). Privacy, Anonymity, and Perceived Risk in Open Collaboration: A Study of Service Providers. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12.

Merchant, A., Shah, D., Bhatia, G. S., Ghosh, A., & Kumaraguru, P. (2019). Signals Matter: Understanding Popularity and Impact of Users on Stack Overflow. *The World Wide Web Conference*, 3086–3092.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2020). Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv:1811.07867 [stat]*.

Morgan, J. T., Bouterse, S., Walls, H., & Stierch, S. (2013). Tea and sympathy: Crafting positive new user experiences on wikipedia. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 839–848.

Narayan, S., TeBlunthuis, N., Hale, W. S., Hill, B. M., & Shaw, A. (2019). All Talk: How Increasing Interpersonal Communication on Wikis May Not Enhance Productivity. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 101:1–101:19.

O'Neil, C. (2018). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Books
OCLC: 1039545320.

Pentland, A. (2008). *Honest Signals How They Shape Our World*. The MIT Press
OCLC: 8162307241.

Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review*, *62*(4), 659–661.

Piskorski, M. J., & Gorbatâi, A. D. (2017). Testing Coleman's social-norm enforcement mechanism: Evidence from Wikipedia. *American Journal of Sociology*, *122*(4), 1183–1222.

Potthast, M., Stein, B., & Gerling, R. (2008). Automatic Vandalism Detection in Wikipedia. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. W. White (Eds.), *Advances in Information Retrieval* (pp. 663–668). Springer.

Reagle, J. M. (2010). "Be Nice": Wikipedia norms for supportive communication. *New Review of Hypermedia and Multimedia*, *16*(1-2), 161–180
_eprint: https://doi.org/10.1080/13614568.2010.498528.

Ridgeway, C. L. (2019). *Status: Why is it everywhere? why does it matter?*
OCLC: 1104214327.

Roberts, S. (2016). Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Studies Publications*.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678.

Seering, J., Kraut, R., & Dabbish, L. (2017). Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 111–125.

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, 1461444818821316.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

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

Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L., & Tan, C. (2019). Content Removal As a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 163:1–163:21.

Stevenson, M. T. (2017). Assessing Risk Assessment in Action. *SSRN Electronic Journal*.

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

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Wallach, H. (2019). Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency. *Medium*.

Weber, M. (1978). *Economy and society*. University of California Press.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, 1391–1399.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12.

Zhu, H., Yu, B., Halfaker, A., & Terveen, L. (2018). Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), 194:1–194:23.