CHS: Small: Collaborative Research: Community Success: Advancing Knowledge of Collaborative Organization Through Large-Scale Empirical Comparison of Peer Production Systems

Peer production – the form of community-based online collaboration used to create public information goods like Wikipedia and Linux – has transformed knowledge production, management, and innovation (Benkler, 2006; Benkler et al., 2015). Inspired by large, successful peer production projects, thousands of wikis and free/libre open source software (FLOSS) projects are started each year. The vast majority of these efforts – and even those that build sustainable and useful information goods – never become more than the work of a single individual (Healy and Schussman, 2003; Schweik and English, 2012).

Why do some peer production systems mobilize large communities of contributors and create valuable information goods while most do not? Although there has been an enormous amount of research on peer production (see reviews by Benkler et al., 2015; Crowston et al., 2010; Jullien, 2012; von Krogh et al., 2012), most has described the inner workings of projects that are among the largest and most collaborative communities and can only provide limited insight into what leads projects to thrive in the first place.

We propose a systematic comparative approach to the study of peer production communities that will test core theories of organizational behavior and advance both theoretical and practical understanding of collaboration in peer production systems. We will pursue this analysis through large-scale empirical comparisons across the wikis hosted by the Wikimedia Foundation and Wikia Inc., the largest populations of peer production wikis. In the process, we will combine insights from socio-technical systems and organization science to understand the determinants of community success.

Our empirical research will test three of the most pervasive claims about the organization of peer production: (1) that peer production is facilitated by low transaction costs; (2) that peer production is limited by competition for volunteer resources; and (3) that social interaction drives sustained participation and growth of peer production projects. Contrary to received wisdom, we anticipate that peer production organizations will be robust to small increases in transaction costs, will benefit from spillovers in volunteer effort across similar projects, and that increased social interaction will drive greater quantity and quality of participation. In testing these claims, we will design, build, and publish software to engage in organizational-level analysis of wikis and a unique research dataset of nearly 80,000 wiki projects.

The **intellectual merit** of this project is that it advances knowledge of collaborative organization through the empirical comparison of many public peer production systems using a common software platform. We contribute to knowledge at the intersection of social computing and human collaboration by using organizational theory to draw inferences about the factors that predict the growth and survival of peer production systems. We also contribute to the study of organizational behavior by capitalizing on the extraordinarily granular sources of data generated by peer production and comparing across many thousands of communities.

The **broader impact** of this project is two-fold. First, we will develop actionable insights that communities, system designers, firms, and movements engaged in online collaboration can use to achieve their collaborative goals. Additionally, we will generate a set of freely licensed and publicly available computational research systems and datasets. In both ways, we will contribute to the ability to design for better peer production projects.

1. Project Goals and Prior Work

Peer production has given rise to new types of organizations engaged in the creation of culturally and economically valuable information goods. For example, free/libre open source software (FLOSS) now accounts for some of the foundational infrastructure of the Internet and has become a mainstay of embedded systems and enterprise computing environments. Wikipedia, perceived as Utopian and unrealistic when it began in 2001, has transformed the encyclopedia industry, eliminating most for-profit encyclopedia producers and building a body of articles found to be of equal or higher quality than the leading for-profit competitor (e.g., Giles, 2005; Benkler et al., 2015). The success and impact of these projects have inspired many thousands of attempts to create peer production communities and harness the potential of distributed online collaboration (Crowston et al., 2010). However, the vast majority of these efforts fail to mobilize contributors or produce collaboration (e.g., Hill and Monroy-Hernandez, 2013b; Schweik and English, 2012). Our work tests three of the most influential explanations of why some attempts at peer production succeed while most do not.

In previous research, scholars have sought to understand the foundations of the success and failure of peer production projects through a variety of approaches and methods. Some have approached the study of peer production primarily through the lens of socio-technical systems research, while others have oriented their work toward testing and elaborating theories from organization science. At the same time, most studies have focused on analyzing a single community. In Figure 1, we provide a schematic visualization of these previous approaches along two dimensions in order to illustrate the contributions of our work.

Our work is strongly situated in the relatively underexplored left side of the graph. Like many of the most successful examples of peer production research, but unlike almost all comparative multi-project research, our proposal draws heavily both on socio-technical systems and organization science theories. To understand community success, we will compare many thousands of peer production communities and test the role of three of the most widespread explanations for the success of peer production systems drawn from both socio-technical systems research and classical organizational theory: transaction costs, resource competition, and social interaction.

1.1. *Building a Comparative Organization Science of Peer Production Systems*

In order to understand why peer production organizations succeed or fail, researchers must compare successful and failed projects. A major part of organization science in the twentieth century turned away from models of organizations as self-contained entities to models that placed organizations in their social environments (Scott and Davis, 2006). One important empirical consequence was a move from intra-organizational studies of internal processes to cross-organizational comparisons of similar organizations. This has led to a rich tradition of testing theories at the level of populations of comparable firms (Hannan and Freeman, 1977; Hannan and Carroll, 1992; Steinfield et al., 2005) and social movements (McCarthy and Zald, 1977; Soule and King, 2008). With the large-scale, public, and exhaustive data available from many peer production projects, socio-technical systems researchers can increasingly embrace similar, comparative approaches. Because peer production communities frequently use common software platforms to mediate their work and communication, they will often provide data that is deeper, more granular, and more directly comparable across projects than is possible in traditional organizational analyses.

Both socio-technical systems research and organization science have produced numerous insights into organizational aspects of peer production communities and wikis (especially Wikipedia). We have reviewed this extensive literature in prior work (Benkler et al., 2015) and elaborated how existing accounts document Wikipedia's early growth, more recent decline, editor contribution patterns

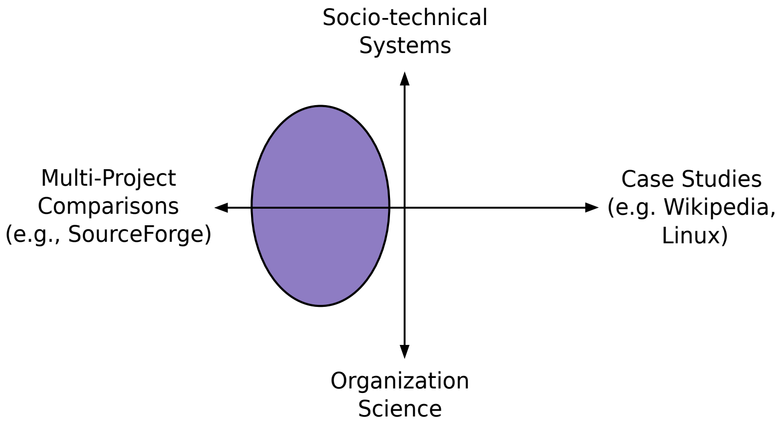


Figure 1: Schematic visualization of prior organizational research on peer production. We combine socio-technical systems approaches with organization science in large-scale comparative analysis. Most existing work has occurred in the upper and lower right quadrants. The purple oval locates our proposed research.

socialization practices, leadership dynamics, bureaucratic structure, governance procedures, community culture, gender relations, coordination dynamics, social roles, and more.

Many of these studies have focused on elaborating the organizational practices and structure within a single community (the right half of Figure 1), whether from a perspective more rooted in the traditions of organization science (e.g., Ransbotham and Kane, 2011; Shaw, 2012), the approaches of socio-technical systems work (e.g., Kittur et al., 2009; Kriplean et al., 2008; Priedhorsky et al., 2007; Welser et al., 2011), or hybrid approaches (e.g., Arazy et al., 2015; Butler et al., 2008; Mugar et al., 2014; Zhu et al., 2012). The vast majority of the previous work on organizational aspects of peer production and online groups has not adopted a cross-organizational perspective.

Our project extends a comparative, inter-organizational approach to the study of peer production, social computing, and online communities. Several reviews of this literature – including ours and one by Crowston et al. (2010) – have highlighted the need for more comparative analysis that includes less successful projects. Only a handful of studies have considered wikis other than the English language version of Wikipedia. In one example, Kittur and Kraut (2010) look at a large group of wikis primarily as a way of understanding the generalizability of findings from Wikipedia. Another small group of studies by Bao et al. (2012) and colleagues has compared the content of different language versions of Wikipedia and focused on cultural and linguistic differentiation. Other work has tested theories of common pool resource management (Schweik and English, 2012) and described communication practices across a large sample of software development teams in SourceForge (Østerlund and Crowston, 2011). In prior work, we have tested a central finding from organization science across a comparative sample of over 600 wikis (Shaw and Hill, 2014).

In the project we propose here, we will design and execute a series of comparative analyses on a large population of wikis to understand how core organizational mechanisms impact peer production communities' growth and survival over time. In ways we have described, research on a large population of projects promises greater external validity than much previous work. We will also be able to test findings from previous peer production research in organization science and socio-technical systems research. This research can inform the design of future peer production systems.

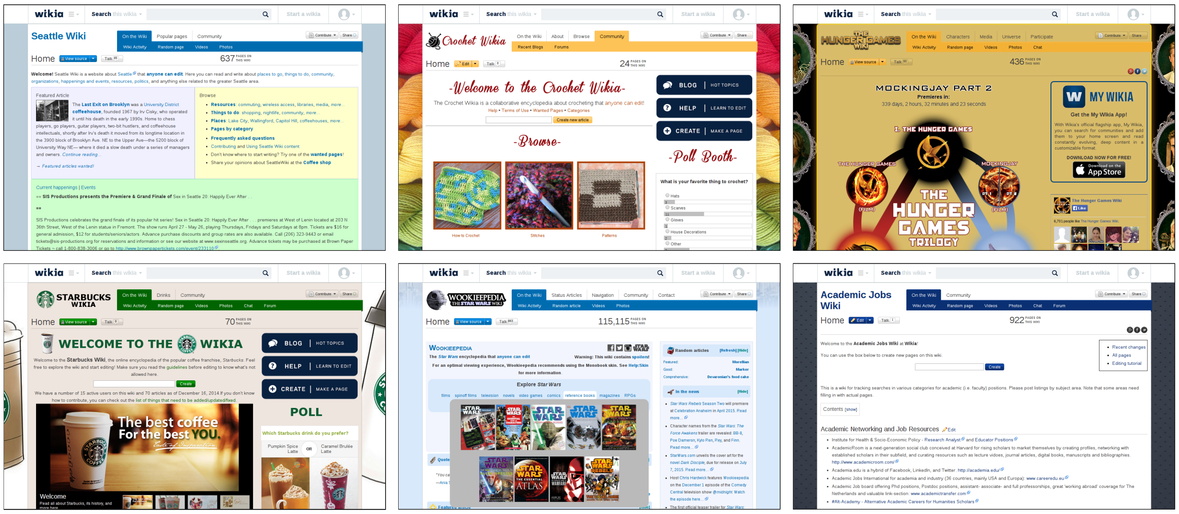


Figure 2: Examples of six wikis hosted by Wikia Inc. from our dataset including (clockwise from top left) Seattle Wiki, Crochet Wiki, The Hunger Games Wiki, the Academic Jobs Wikis, Wookiepedia (a Star Wars encyclopedia) and Starbucks wiki.

1.2. Comparing Peer Production Wikis from Wikia and the Wikimedia Foundation

We propose to build a large-scale comparative dataset of peer production wikis from the two single-largest public wiki-hosting platforms: the Wikimedia Foundation (WMF) and Wikia, Inc. WMF is the the non-profit organization that supports English Wikipedia and more than 800 different wiki communities that include efforts to build encyclopedias, dictionaries, travel guides, text books, biological taxonomic databases, and other reference works in more than 280 languages. Wikia, founded in 2004, sought to apply the Wikipedia model beyond the education-based scope of WMF. Wikia was founded by Jimmy Wales, Wikipedia's founder, and Angela Beesley, one of the most active and respected contributors to Wikipedia in its early years. Wikia's policies, structures, and technologies have been heavily influenced by Wikipedia. Although many private firms host wikis (e.g., PBWiki, WikiSpaces, and SocialText), Wikia is unique in that it only hosts publicly accessible, volunteer-produced, peer production projects and never restricts viewership or participation except to combat spam or vandalism.

Several aspects of Wikia and WMF wikis make them ideal settings to compare organizational behavior and outcomes in peer production communities. First, they are both peer production. Like Wikipedia, anybody can create an account on any WMF or Wikia wiki. The vast majority of these wikis allow contributions even without accounts. Like FLOSS, and other peer production projects, all WMF and Wikia content is distributed freely to the public.[1] More importantly, all WMF and Wikia wikis run on the same software platform – MediaWiki – making them ideally suited to a large-scale comparative analysis. As Figure 2 illustrates, Wikia wikis in particular cover diverse topics and attract widely varying degrees of participation within a common infrastructure. The majority, like Starbucks wiki, only have a handful of contributors and a few dozen pages at most. Others, such as Seattle Wiki, Crochet Wiki, or the Academic Jobs Wiki, become vibrant, medium-sized communities with hundreds or thousands of editors. A very few, like Wookiepedia (a wiki about Star Wars), attract tens of thousands of contributors and contain hundreds of thousands of pages – placing them

[1] As is common in peer production, ownership of the copyright on wiki content remains with the contributors but all material is licensed freely to the public as a condition of contribution. All WMF and Wikia content is released under the Creative Commons Attribution-ShareAlike license and is made publicly available for download.

among the largest and most active peer production communities anywhere.

1.3. Testing Key Theories of Peer Production Community Success

In order to formally test theories at a comparative, inter-organizational level, scholars of peer production need to continue to embrace the opportunity presented by the existence of many thousands of otherwise similar projects that vary in terms of the size, collaborativeness, and behavior. We propose to advance the science of organizations and socio-technical systems through just such an approach. Specifically, we will test three core theories of peer production community success:

> **Project 1 — Transaction costs:** Do small changes in the effort required for contribution to a peer production community positively affect the growth of the resource being produced? Divergent claims suggest that "transaction costs" might either facilitate peer production by lowering barriers to entry (Benkler, 2006) or open the door to enhanced abuses and undermine community growth (Resnick et al., 2000; Friedman and Resnick, 2001).
>
> **Project 2 — Resource mobilization:** Is volunteer effort in the context of peer production communities a "fixed and finite" resource for which voluntary organizations compete (McCarthy and Zald, 1977; Soule and King, 2008; Wang et al., 2013)? Or do we see patterns consistent with theories predicting a positive feedback cycle as contributors build on each others' work and interest and generate positive spillovers (Meyer and Whittier, 1994)?
>
> **Project 3 — Social interaction:** Do increases in the frequency and visibility of social interactions among contributors to peer production projects drive increases in contributions to the information resource being produced? A group of studies have proposed that such interactions will generate high quality participation in the future (Arguello et al., 2006; Butler, 2001; Cheshire and Antin, 2008; Erickson and Kellogg, 2000) and these claims mirror findings in research on the importance of interactions and social participation in collective action (Bimber et al., 2012; Bennett and Segerberg, 2012); corporate intranets (Fulk et al., 2004); and labor unions (Lipset et al., 1956).

We describe the research design and methods for these studies in detail in the corresponding subsections of the project plan in Section 2. We also elaborate how, in the process of doing this work, we will create software for parsing and extracting data from tens of thousands of wiki database files and tools to extract metadata from large numbers of public wiki application programming interfaces (APIs). We will publish these tools under FLOSS licenses and will use them to create and publish comprehensive comparative research datasets.

2. Project Plan

As we have suggested, few organization-level theories about peer production have been formalized and tested empirically across a large population of projects. Among the claims that have emerged in previous peer production and socio-technical systems research, three stand out for their clarity, tractability, and resonance with earlier perspectives in organizational analysis: Projects that (1) offer low barriers to entry, (2) successfully compete with other projects to recruit and mobilize volunteers, and (3) create an environment where social interaction is more frequent and visible, are more likely to grow and survive than projects that do not. We propose to test all three theories in comparative quantitative analyses.

Across all three of the proposed studies, we will use multilevel regression to evaluate empirical models. Where appropriate, we will combine these techniques with "quasi-experimental" methods to identify and estimate the causal impact of the different organizational mechanisms under consideration (Shadish et al., 2002). In each case, we will publish documentation and statistical code used to conduct analyses in conformance with Stodden and Miguez's (2013) best practices for ensuring reproducible computational research results.

2.1. Project 1: Testing The Effect of Transaction Costs on Participation in Peer Production

Several scholars suggest that extremely low costs of contribution are one of the central characteristics of peer production systems, and that, all else being equal, a project with lower barriers to contribution will attract a higher volume of contributions (e.g., Antin et al., 2012; Benkler, 2002; Preece and Shneiderman, 2009). Others have suggested that eliminating barriers to contribution entirely invites low quality participation and can create a negative feedback cycle driving subsequent contributors away (e.g., Halfaker et al., 2013; Hill and Monroy-Hernandez, 2013b). In support of this approach, a large body of research on reputation systems in online groups suggests that requiring users to adopt consistent identifiers can help recruit newcomers and elicit commitment (Kraut and Resnick, 2012) while also potentially deterring negative and disruptive behavior (Resnick et al., 2000). There have been almost no systematic attempts to compare the benefits and costs of either approach.

We propose to adjudicate between these perspectives on the effect of small barriers to entry in peer production communities by analyzing a series of events where 182 wikis suddenly had anonymous editing enabled or disabled over relatively short periods of time. The unannounced changes are exogenous, discontinuous shocks to the projects, and therefore natural experiments (Shadish et al., 2002). The fact that the change was implemented in software means that end-users and potential participants in the wikis had no way of avoiding exposure to the "treatment" (in this case, the absence of non-registered editing opportunities). We will estimate the effect of the discontinuous increase in transaction costs on the overall quantity and quality of incoming contributions to peer production projects by comparing edits immediately before and after the unanticipated feature change in a "regression discontinuity design" (Lee and Lemieux, 2010).

To assess contribution quality, we draw on an extensive body of research applying computational approaches to the analysis of natural language and network-based measures in the specific context of collaborative wikis (Adler and de Alfaro, 2007; Adler, 2012). Our unit of analysis is the individual wiki and we will examine the effect of the changes on the number of high and low quality contributions by comparing the five weeks immediately preceding and following each change using multilevel longitudinal models that allow for flexible specifications of time and incorporate wiki level effects (Singer and Willett, 2003).

Hypotheses While the theoretical and empirical findings described above imply divergent outcomes, we frame our hypotheses to correspond to the perspective that increased transaction costs will drive a decline in contributions across the board. Specifically, we hypothesize that:

> *$H1_A$: When wikis remove the ability to contribute without accounts, the number of low quality edits will decrease.*

> *$H2_A$: When wikis remove the ability to contribute without accounts, the number of high quality edits will decrease.*

In this sense, $H1_A$ represents the intentional effect of reducing less valuable contributions while $H2_B$ reflects the "collateral damage" of also deterring good contributions.

Measures We will classify edits by quality using Adler and de Alfaro's (2007) widely-adopted measure *WikiTrust*. In the WikiTrust model, an edit that is entirely preserved while an article is revised is considered to be of high quality while an article is that is entirely or largely removed would be classified as being of low quality. Because our hypotheses are stated in terms of the number of high quality or low quality edits, our dependent variables will be binned at the week level. To capture the number of low quality edits for every wiki in each week, we will count the number of edits made to the wiki where less than 20% of the text added in the contribution remains after six subsequent contributions.[2] Our measure of the number of high quality edits is the number of edits made to the wiki where more than 80% of the text added in the contribution is preserved over six subsequent contributions.

Because wikis might see higher or lower quality edits at different points in their lives, we will construct and include a control for each wiki's age in years at each time period. Because the calendar date can affect the way that individuals contribute to wikis (e.g., the creation of spamming robots might lead to more low quality edits), we will create a control for this as well. Finally, we will include additional controls for wiki subject area and topic that we will create by working with a team of graduate student researchers to classify the wikis in our dataset. As part of the creation and validation of our measures, we will also conduct open-ended interviews with a small, random sample of wiki administrators to evaluate the process and results through which the feature change occurred.

Anticipated Findings Based on informal qualitative observation of several of the wikis we plan to measure, we anticipate that higher transaction costs will result in less contributions from a smaller number of contributors across the wikis in our population in support of $H1_A$. However, we do not expect to find less high quality contributions and anticipate finding little or no support for $H1_B$. Regardless of the findings, we will design subsequent analyses and follow-up studies based on the initial results to identify mechanisms that support the behaviors we observe. If we find evidence of heterogeneous project behaviors (e.g., some projects with reduced barriers to participation see an increase in high-quality contributions, whereas others only encounter an influx of spam), we plan to pursue multiple case study analyses (Eisenhardt, 1989) in order to better understand the processes driving these results. Any result will provide an important first test of Benkler's (2006) basic theory of peer production.

2.2. Project 2: Evaluating Theories from Resources Mobilization in Peer Production

According to ecological models of organizational resource constraint, voluntary organizations operating in the same field seek members to engage in similar work and therefore compete for their time and effort (McCarthy and Zald, 1977, 2001; Minkoff, 1997; Soule and King, 2008). Using data from online discussion groups, scholars of online communities have shown that discussion groups covering similar topics compete for participants (Wang et al., 2013; Zhu et al., 2014b) and succeed when they have found effective niches (Zhu et al., 2014a). But volunteer labor is different from many other resources in that organizations may generate interest in volunteering (McCarthy and Zald, 1977, 2001). Indeed, volunteer mobilization efforts by one organization may even generate positive "spillovers" in other organizations (Marwell and Oliver, 1993; Meyer and Whittier, 1994). These spillovers may be even more likely in the context of peer production because potential contributors can immediately utilize and re-purpose earlier contributions due to the systematic free licensing of work products in peer production allowing content reuse.

We will compare volunteer activity across different wikis on the same topic, and articles on same topics in Wikipedia, to test whether volunteer labor behaves like the "fixed and finite" resource assumed in ecological theories. Drawing on one of the example wikis in Figure 2, we might ask whether contributions to *The Hunger Games* wiki go up, down, or remain the same when contributions to the *The Hunger Games* articles on Wikipedia pages increase.

[2] See Adler and de Alfaro (2007) and Adler et al. (2011) for details on specifics of the WikiTrust algorithm and its validation.

Hypotheses Consistent with the perspective adopted in the majority of the literature on resource mobilization, we will frame our hypotheses around the claim that volunteer effort functions like a "fixed and finite" resource:

> *$H1_B$: Contributions to active Wikia wikis on a given topic will decline as contributions to articles in the corresponding topic areas on Wikipedia go up.*

Measures To compare across active Wikia wikis and the corresponding topic areas of Wikipedia, we plan to categorize active Wikia wikis according to their subject matter (measures we describe briefly in Project A). Using this information, we will identify corresponding categories for Wikipedia articles for these topics. Our key dependent and independent variables will then consist of counts of the number of edits within a given topic on Wikipedia and Wikia for every week in our dataset.

Once again, we will include controls for the wiki's age in years at each time period. Also, because the underlying interest in a given topic might fluctuate from one week to the next, we include a control for the calendar date as well as a baseline measure of interest in a topic measured by relative search engine traffic for each topic. As part of our measure development and validation, we will conduct open-ended interviews with a small sample of contributors who work across Wikia and WMF projects in order to understand how they engage in the different communities.

Anticipated Findings In exploratory work using twenty-six wikis, we have found evidence suggesting that even during a period during which volunteer contributions to Wikipedia are shrinking overall, contributions to Wikia wikis are positively related to contributions to the overlapping topics in Wikipedia. We also find a positive relationship within the activity of editors we identify as participants in both projects, providing some evidence of spillovers.

Assuming the pattern of spillovers we have observed in our pilot analysis holds up in a larger number of wikis and over longer periods of time, we plan to investigate the causes of spillovers in greater depth. For example, spillovers may be higher among particular kinds of topics or genres of collaborative production. We will pursue follow-up analyses of wiki projects that either represent or resist the central tendencies observed across the rest of the dataset to refine and build our analyses.

2.3. Project 3: Analyzing the Impact of Increased Social Interaction on Public Goods Production

Overlapping bodies of organization science and socio-technical systems research have emphasized the role of group dynamics and social interactions in shaping organizational behavior. We will seek to understand the effect of sudden changes in the frequency, visibility, and quality of social interactions on participation patterns in peer production organizations. Existing work suggests that such sudden increases in the frequency and visibility of social interactions will drive increased participation in collective action (Bimber et al., 2012) and online communities (Arguello et al., 2006; Erickson and Kellogg, 2000), as well as enhanced performance in other kinds of organizations (Kellogg, 2009), including those engaged in the production of common pools of shared information (Butler, 2001; Cheshire and Antin, 2008; Fulk et al., 1996, 2004). To our knowledge, the causal form of this claim has not been previously tested across a large sample of organizations of any kind.

On many Wikia wikis, a threaded, forum-style discussion feature called "message walls" was suddenly switched on by project administrators. WMF is currently developing a very similar tool called "Flow"[3] which it has already rolled out on a very small number of pages and projects. In preliminary qualitative observations, we have found that the introduction of these tools facilitated more

[3] https://www.mediawiki.org/wiki/Flow

sustained conversations between participants and made these conversations more visible. We will treat the adoption of these interpersonal communication tools as another "natural experiment" and test whether the shift in social interactions driven by the introduction of these tools catalyzed different contribution patterns among new editors. Specifically, we will examine the impact of the tools on the frequency and tone of interactions between participants, and on the frequency and quality of contributions from users who join immediately before or after the introduction of the feature.

Hypotheses We plan to test four hypotheses that operationalize the findings of prior literature suggesting that enhanced ease and visibility of social interactions will result in more frequent, positive, and high quality contributions:

> *$H1_C$: The introduction of improved tools for interpersonal communication will cause the number of newcomer social interactions on wikis to increase.*
>
> *$H2_C$: The introduction of improved tools for interpersonal communication will cause the tone of newcomer social interactions on wikis to become more positive.*
>
> *$H3_C$: The introduction of improved tools for interpersonal communication will cause the number of newcomer contributions to wiki articles to increase.*
>
> *$H4_C$: The introduction of improved tools for interpersonal communication will cause the quality of newcomer contributions to wiki articles to increase.*

Measures The dependent variables in this project will all be attributes of newcomer edits to wikis that turned on the improved interpersonal communication feature. To measure the number of newcomer social interactions, we will construct a count of edits to "message wall" pages or "talk" pages (the new and the old systems for interpersonal communication) by users who created new accounts shortly before and shortly after the transition to the new system. To measure the tone of newcomer social interactions, we will apply automated sentiment analysis using the Linguistic Inquiry and Word Count (LIWC) system (Pennebaker et al., 2001) to the text of all messages sent by newcomers. To examine contributions to articles, we will construct a count of the number of edits the same newcomers have made to article pages. Finally, we will measure the quality of these article edits using the WikiTrust measure we described in Project A above. As in the first two studies described above, we also include a set of controls to account for the likelihood that wikis, as well as individual editors, may vary substantially in ways unrelated to either the introduction of message walls or our outcomes of interest, including controls for the wiki's age at each time period, calendar date, and wiki subject area. As part of our measure development and validation, we will also conduct interviews with a sample of project contributors and administrators to understand their experience of working in wikis with and without augmented interpersonal communication tools like message walls.

Anticipated Findings Based on the findings of previous research and an informal qualitative analyses of two communities' transition to message walls we have already conducted in preparation for this proposal, we anticipate that message walls likely drive a spike in the frequency of social interactions on the wikis, and result in a larger proportion of contributors engaging in on-site conversation. We also predict that the message walls increase the positive tone of these conversations. Finally, we expect that the introduction of message walls caused a positive increase in both the number of contributions by new members and the quality of those contributions. We plan to investigate whether these expectations stand up to systematic testing, and, depending on the outcomes of our initial findings, we will pursue follow-up research to uncover additional details about the mechanisms involved.

2.4. Creating Research Tools and Datasets

To complete the three studies described above, we will need to construct a set of comparative longitudinal datasets that include summaries of activity across many Wikia and WMF wikis. We will write software to build these datasets and will then document and distribute these tools for comparative wiki research under a FLOSS license in order to facilitate adoption and improvement. Currently, multiple tools exist to process Wikipedia data, but none effectively support processing multiple MediaWiki databases from multiple sources in order to create comparative, multi-community datasets. Because both WMF and Wikia use the MediaWiki software, their database dumps use an identical format and creating such datasets is realistic and achievable.

Although specific variables will differ across the three projects we have proposed, each will require a dataset that includes a variety of data and metadata that capture the rate and type of activity within wikis over periods of time (e.g., where the unit of analysis is the "wiki day" or "wiki week"). We have already described several key variables we will use and we have used other variables in our previous research. Although we will not detail every possible variable here, examples of common variables we will create include the number of contributions, the number of unique contributors, the number of contribution to administrative pages, and the number of "talk messages" on a given wiki during a given period.

In preliminary work, we have downloaded database dumps from more than 77,000 publicly available wikis hosted by Wikia Inc. The data were made available for every Wikia wiki routinely through April 2010 when we collected the raw data. We have also downloaded database dumps for WMF hosted wikis. Together, these datasets contain many terabytes of "raw" XML data including, for example, the full text of every revision of every page in every wiki and detailed metadata on contributors. Additionally, we have downloaded a complete archive of viewership logs from WMF. All of these dataset are currently stored in *Hyak*, a shared high performance computing environment at the University of Washington. In some cases, information is not available in these dumps (e.g., data on administrative rights, blocking, banning, and deleting). To collect these data, we will construct tools to query public wiki application programming interfaces (APIs) and to merge the results of these queries with the data from the database dumps.

To create comparative datasets, we will (1) parse the raw XML dataset to create tables of metadata for each individual contribution to each wiki; (2) summarize these edit histories to create datasets with summaries of each wiki; (3) bin the per-wiki edit histories by period (i.e., by days, weeks, months, and years); and (3) merge these binned and summary datasets into large comparative datasets of various size and longitudinal granularity.

As we have done in previous work, we will build software to parse and analyze these datasets in Python and R and will reuse and extend existing tools in the process. For example, we will make extensive use of the freely licensed and open source MediaWiki-Utilities library created and published by the Wikimedia Foundation.[4] Were possible, we will contribute code back to the library where we have already made contributions as part of pilot work for this proposal.

We will work together with the Wikimedia Foundation in developing this aspect of the project. First, they will host our research team for a month each year over the course of the proposed research. Second, they will provide us with access to computational resources designed for analyzing wiki data, including their Hadoop computing cluster to assist in preparation of research datasets. Third, they will work with us to arrange for access to private data stored by the Foundation but not systematically made available to the public. This access will allow us to answer substantive research questions about community success and to collaborate with them to prepare aggregate and/or anonymized

[4] https://github.com/halfak/mediawiki-utilities

versions of these data for public release. Where appropriate, WMF will also provide a venue for us to publish comparative datasets and tools to facilitate the widest dissemination and impact of our research products.

In addition to computational resources made available through WMF, we will purchase capacity in UW's Hyak high performance computing cluster to ensure that we have the processing power and memory necessary to complete this work. Finally, we will purchase a server and storage array to be hosted at Northwestern University to complete additional analysis not possible in Hyak because it is firewalled from the Internet and to host the processed public datasets for other researchers to download.

Although building software to create and analyze these data are an important challenge, we have already successfully built and used two prototype versions of the tools we describe in this proposal. An prototype used to create a dataset for an earlier paper (Shaw and Hill, 2014) was written in C++ and is published online.[5] A second prototype using the Python architecture we will use to complete the work described in this proposal was created as part of our study on redirects (Hill and Shaw, 2014) and is more flexible and adaptable. In unpublished work, we have bench-marked and tested the new system and are confident that it can adapted for building large comparative datasets with the computational resources we are requesting.

Finally, we will make our research datasets available to the public. We will release all research data under the GNU GFDL and Creative Commons Attribution Share-Alike license (the free/open source licenses used for content by Wikia and WMF). Our goal is to produce material analogous to the dataset created by the FLOSSmole project (Howison et al., 2006), which have supported a large proportion of comparative research into peer production to date.

In pilot work on redirects, we have released tools and technical documentation for processing wiki datasets.[6] Because this is a much larger project, we will issue a preliminary release of tools, documentation, and data after two years of work and will announce this release on our blogs and several popular mailing lists for Internet and peer production research (e.g., AIR-l, wiki-research-l). We will solicit feedback, suggestions, and patches to prepare an updated version of the software, documentation, and dataset for release in the final year of the project. We will also place this version of the dataset in a public archive for scientific data (e.g., the Harvard Dataverse) and will seek to publish documentation for the data in a peer reviewed venue (e.g., *Nature Scientific Data* or *PLOS ONE*).

3. Dissemination Plan, Education, and Outreach

We will disseminate the outcomes of this project through scholarly communication channels including scientific workshops, conferences, and journals. We have a history of, and a commitment to, publication in venues including Computer-Supported Cooperative Work (CSCW), Human-Computer Interaction (CHI), the International Conference on Weblogs and Social Media (ICWSM), the International Symposium on Open Collaboration (OpenSym, formerly WikiSym), as well as peer-reviewed journals on communication and sociology (e.g., the *Journal of Communication*; *Information, Communication and Society*; *Politics and Society*; and *American Behavioral Scientist)*. Whenever possible, we will also release public and freely licensed versions of our research products. In the past, we have published in open access scientific journals (*PLOS ONE*), released work under open access licenses (e.g. Hill and Shaw, 2014), and published pre-print versions on our professional websites. As described in preceding sections of this proposal, we will also disseminate the tools and datasets used in our

[5] Available at: https://github.com/makoshark/wikiq

[6] http://communitydata.cc/wiki-redirects

research under free and open licenses. These activities will both ensure the reproducibility of our findings, and facilitate the widest dissemination and scientific impact of our work.

The PIs will both include methods and results from this research directly in the courses they teach at Northwestern University and the University of Washington. PI Shaw teaches an undergraduate course on online communities and graduate seminars on peer production and research design which will both incorporate the results of the projects proposed here. PI Hill also teaches a class on online communities and graduate courses on designing Internet research and data science with online community data. All of his teaching has a strong conceptual and practical engagement with peer production and he will incorporate findings, tools, and data into all three courses.

As both PIs have done in their previous research, the research team will work closely with organizations supporting peer production in ways that go beyond simply using these communities as a source of data. Both PIs are active contributors to Wikipedia and PI Hill has been a member of the Wikimedia Foundation advisory board since 2007. In collaboration with PI Shaw, Hill also delivers an annual talk on academic research at Wikipedia's yearly International conference, Wikimania. Both PIs have regularly given talks on their research at Wikimania, the Wikimedia Foundation, Wikia Inc., and WikiHow (another commercial wiki hosting firm). As part of their support for this project, WMF has committed to hosting talks on findings of this research for the WMF staff. Hill is also a founding board member of the Wikimedia Cascadia User Group, and a regular attendee and organizer of events related to Wikipedia in the Seattle area. PI Shaw serves on the Citizen Science Advisory Board of the Adler Planetarium in Chicago, where he assists in the design and assessment of the Zooniverse, another widely successful peer production platform. We will use these relationships with the peer production contributor and business communities to disseminate our research and to ensure that it is relevant, useful, and usable.

In this project, we will work particularly closely with the research, analytics, and development teams at the WMF. In support of this proposal, the WMF has offered to host both PIs as well as graduate student research assistants for multiple weeks in their San Francisco offices. Our relationship with WMF will ensure that we access to knowledge of the MediaWiki software (of which WMF is the primary author and will ensure that our work is integrated into the technical analytics and research infrastructure used by peer production and wiki researchers worldwide.

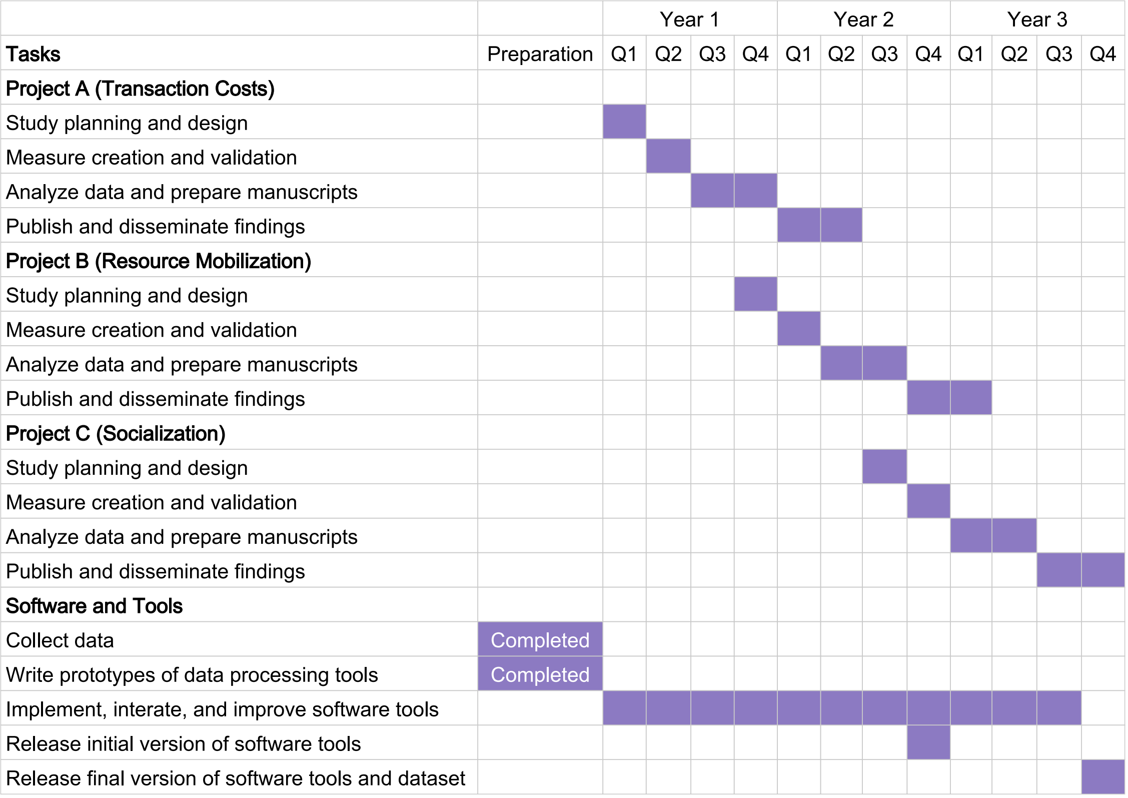
At both institutions, graduate students will be involved in the project as research assistants, collaborators, and advisees, and will interact with both PIs through research activities and monthly team meetings over video conference. These experiences will provide them with a supportive learning environment incorporating personnel from both Northwestern and the University of Washington.

4. Project Timeline

The timeline in Figure 4 describes our empirical research in terms of the three projects proposed in Section 2 as well as a supporting work in terms of software systems and dataset preparation.

5. Intellectual Merit

The **intellectual merit** of the proposed work stems from the way our research will will bridge the rich traditions of computing research and organization science and use comparative data from populations of peer production communities to test three of the most important theories explaining community success. In particular, we will test competing theories about the effect of small transaction costs in the form of required accounts, evaluate resource mobilization theory and ecological models of competition in the context of online communities, and consider the effect of socializing on



		Year 1				Year 2				Year 3			
Tasks	Preparation	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Project A (Transaction Costs)													
Study planning and design		■											
Measure creation and validation			■										
Analyze data and prepare manuscripts				■	■								
Publish and disseminate findings						■	■						
Project B (Resource Mobilization)													
Study planning and design					■								
Measure creation and validation						■							
Analyze data and prepare manuscripts							■	■					
Publish and disseminate findings									■	■			
Project C (Socialization)													
Study planning and design								■					
Measure creation and validation									■				
Analyze data and prepare manuscripts										■	■		
Publish and disseminate findings												■	■
Software and Tools													
Collect data	Completed												
Write prototypes of data processing tools	Completed												
Implement, interate, and improve software tools		■	■	■	■	■	■	■	■	■	■	■	
Release initial version of software tools									■				
Release final version of software tools and dataset													■

Figure 3: Project timeline broken up in the quarter system used by both Northwestern and UW. Q1 is the start of the academic year on the quarter system (i.e., Oct-Dec).

communal public good production. In doing so, we will offer tests of three of the most important theories from social computing, organization science, social movement research, and collective behavior in the context of a large population of online communities. Contrary to prior expectations, we anticipate finding that peer production will be robust to small increases in transaction costs and will overcome resource constraints experienced by traditional volunteer organizations. We also anticipate that social interaction and communication will play a key role in support the quality and quantity of contributions to peer production projects. Testing these theories will build toward a broader theory of community success. This work will facilitate subsequent comparisons with other kinds of organizations, including offline organizations that pre-date the rise of peer production.

6. Broader Impacts of the Proposed Work

The **broader impacts** of this work focus on the initiators and managers of virtual communities and the designers of community software and systems. A growing number of organizations rely on distributed and often ad-hoc teams whose works is organized into virtual communities. Moreover, a growing number of organizations rely on new online communities for content, quality assurance, support, and innovation. Testing three of the key theories of community success will have direct impact on firms building these communities. In particular, it will provide prescriptions on whether and when to require accounts, how to approach new communities on popular topics, and how to invest in technological systems for improving social communication around work products. Our goal is for these results to shape the decisions of firms, non-profits, and governments working with social media and user generated content and to improve the experience of every day users of these systems. In this process, our project may help make virtual work more effective resulting in general economic benefits and improved satisfaction among knowledge workers.

7. PI Preparation and Preliminary Results

7.1. Personnel

Aaron Shaw, Ph.D. (PI), is an Assistant Professor in the Department of Communication Studies at Northwestern University and a Faculty Associate of the Berkman Center for Internet and Society at Harvard University. At Northwestern, he is also affiliated with the Sociology Department, the Institute for Policy Research, and the Buffett Center. His research has been supported by FUSE Labs at Microsoft Research, the Ford Foundation, the Ewing and Marion Kaufman Foundation, the Berkman Center for Internet and Society, the University of California Office of the President, and the United States Department of Education. Shaw holds graduate degrees in Humanities and Sociology from both Stanford University and UC Berkeley.

Shaw studies collective action, collaboration, and mobilization in peer production and crowds, and has published widely on these topics (e.g., Antin and Shaw, 2012; Hargittai and Shaw, 2013; Kittur et al., 2013; Shaw et al., 2011; Shaw and Benkler, 2012; Shaw, 2012). Shaw has received awards from the American Political Science Association, American Sociological Association, the International Communication Association, and the Association for Computer Machinery Conference on Computer Supported Cooperative Work (CSCW) for several papers and his dissertation research. He brings expertise in organizational research, peer production, quantitative and quasi-experimental methods, empirical analysis of online communities, and research design.

Benjamin Mako Hill, Ph.D. (PI), is an Assistant Professor of Communication at the University of Washington. He is also a faculty affiliate at the Berkman Center for Internet and Society and an affiliate at the Institute for Quantitative Social Science – both at Harvard University. Hill holds a Masters degree from the MIT Media Lab and a Ph.D. from MIT in Management and Media Arts and Science from an interdepartmental program overseen by HCI faculty at the MIT Media Lab and social science faculty at the MIT Sloan School of Management. He has published numerous articles in peer reviewed journals and conference proceedings (e.g., Buechley and Hill, 2010; Hill and Monroy-Hernandez, 2013a,b; Monroy-Hernández et al., 2011), and has received awards from the International Communication Association, the Association for Computing Machinery Conferences on Computer Supported Cooperative Work (CSCW) and Human Factors in Computing Systems (CHI), MTV, and Cisco.

Prior to his graduate studies, PI Hill worked full time as a software engineer and has received his masters degree from MIT for software development and HCI related research. Hill has a strong background in technology management, data-driven statistical analyses of online communities, computational research, management science, and peer production. In particular, Hill has been a leader, developer, and contributor to the free and open source software community for more than a decade as part of the Debian and Ubuntu projects, two of the most popular Linux distributions with millions of users worldwide and is the author of several best-selling technical books (e.g., Hill et al., 2008). He is a member of the Free Software Foundation board of directors, and is has been on the Advisory Board of the Wikimedia Foundation since 2007.

If funded, PI Hill will be assisted by a graduate research assistant at the University of Washington. PI Shaw is already assisted by graduate student (funded by Northwestern) who holds expertise in peer production, quantitative research methods, and computational analysis.

7.2. Related Experience and Prior Work

Both PIs possess advanced training in organizational research, statistical analysis, and quantitative methodology and have published peer reviewed research employing large-scale, empirical data analysis techniques. Over the past four years, we have been working collaboratively on research projects and pilot studies to build the experience and skills to successfully complete these projects. As collaborators, we have undertaken multiple research projects on peer production, resulting in three co-authored peer reviewed publications (Hill and Shaw, 2013; Shaw and Hill, 2014; Hill and Shaw, 2014) and multiple co-authored working papers and conference presentations (e.g., Hill et al., 2012; Shaw et al., 2014). In collaboration with Yochai Benkler at Harvard Law School, we have co-authored a book chapter for a forthcoming edited volume from MIT Press that lays out out the need for the kind of comparative organizational research described in this proposal (Benkler et al., 2015).

As discussed in Section 2.4, we have already collected the raw data necessary to complete the studies described in this proposal. Although analyzing these data in the way have described is challenging, we have the preparation to complete this research with the resources requested. Using the prototype tools we describe in Section 2.4, we have already created a dataset of the largest 1% of Wikia wikis which we have used to test several core theories of organizational democracy. The results of this research have been published in a 2014 issue of the *Journal of Communication* on "Big Data" research (Shaw and Hill, 2014). In another pilot project, we wrote software to analyze the full text history of English Wikipedia (the largest and most complicated project we will analyze) to create a comphensive longitudinal dataset of "redirects" and to use the dataset to show that conclusions from several of the most highly cited papers on peer production are subtantively changed when our dataset is taken into account (Hill and Shaw, 2014). We have published both our software and dataset alongside the substantive findings in the published manuscript.[7]

8. Results from Prior NSF Support

PI Hill has one active NSF award (Collaborative Research: New Pathways into Data Science: Extending the Scratch Programming Language to Enable Youth to Analyze and Visualize Their Own Learning; DRL-1417663; $124,374.00; September 1, 2014 – August 31, 2016) studying the impact of data and analytic tools on learning in the Scratch online community. The award is too recent to report results. PI Hill has no other prior NSF awards. PI Shaw has received no previous support from NSF.

[7] http://communitydata.cc/wiki-redirects/

References

Adler, B. T. (2012). *Wikitrust: Content-driven Reputation for the Wikipedia*. Ph.d. dissertation, University of California at Santa Cruz, Santa Cruz, CA, USA. AAI3521730.

Adler, B. T., Alfaro, L. d., Mola-Velasco, S. M., Rosso, P., and West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, number 6609 in Lecture Notes in Computer Science, pages 277–288. Springer Berlin Heidelberg.

Adler, B. T. and de Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 261–270, New York, NY, USA. ACM.

Antin, J., Cheshire, C., and Nov, O. (2012). Technology-mediated contributions: editing behaviors among new wikipedians. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 373–382, New York, NY, USA. ACM.

Antin, J. and Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of amazon mechanical turk in the US and india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2925–2934, New York, NY, USA. ACM.

Arazy, O., Ortega, F., Nov, O., Yeo, L., and Balila, A. (2015). Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of the 2015 ACM conference on Computer supported cooperative work*, CSCW '15, pages 233–236, New York, NY, USA. ACM.

Arguello, J., Butler, B. S., Joyce, E., Kraut, R., Ling, K. S., Rosé, C., and Wang, X. (2006). Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 959–968, New York, NY, USA. ACM.

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. (2012). Omnipedia: bridging the wikipedia language gap. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, pages 1075–1084, New York, NY, USA. ACM.

Benkler, Y. (2002). Coase's penguin, or, linux and the nature of the firm. *Yale Law Journal*, 112(3):369–446.

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

Benkler, Y., Shaw, A., and Hill, B. M. (2015). Peer production: A form of collective intelligence. In Malone, T. and Bernstein, M., editors, *Handbook of Collective Intelligence*. MIT Press, Cambridge, MA.

Bennett, W. L. and Segerberg, A. (2012). The logic of connective action. *Information, Communication & Society*, 15(5):739–768.

Bimber, B. A., Flanagin, A. J., and Stohl, C. (2012). *Collective Action in Organizations: Interaction and engagement in an Era of Technological Change*. Communication, society and politics. Cambridge University Press, New York.

Buechley, L. and Hill, B. M. (2010). LilyPad in the wild: how hardware's long tail is supporting new engineering and design communities. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, pages 199–207, New York, NY, USA. ACM.

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., Joyce, E., and Pike, J. (2008). Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1101–1110, New York, NY, USA. ACM.

Cheshire, C. and Antin, J. (2008). The social psychological effects of feedback on the production of internet information pools. *Journal of Computer–Mediated Communication*, 13(3):705–727.

Crowston, K., Wei, K., Howison, J., and Wiggins, A. (2010). Free/libre open source software: What we know and what we do not know. *ACM Computing Surveys*, 44(2):2012.

Eisenhardt, K. M. (1989). Building theories from case study research. *The Academy of Management Review*, 14(4):532–550.

Erickson, T. and Kellogg, W. A. (2000). Social translucence: An approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(1):59–83.

Friedman, E. J. and Resnick, P. (2001). The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, 10(2):173–199.

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

Fulk, J., Heino, R., Flanagin, A. J., Monge, P. R., and Bar, F. (2004). A test of the individual action model for organizational information commons. *Organization Science*, 15(5):569–585.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.

Halfaker, A., Geiger, R. S., Morgan, J. T., and 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.

Hannan, M. T. and Carroll, G. (1992). *Dynamics of organizational populations density, legitimation, and competition*. Oxford University Press, New York.

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

Hargittai, E. and Shaw, A. (2013). Digitally savvy citizenship: The role of internet skills and engagement in young adults' political participation around the 2008 presidential election. *Journal of Broadcasting & Electronic Media*, 57(2):115–134.

Healy, K. and Schussman, A. (2003). The ecology of open-source software development.

Hill, B. M., Burger, C., Jesse, J., and Bacon, J. (2008). *Official Ubuntu Book*. Prentice Hall, 3 edition.

Hill, B. M. and Monroy-Hernandez, A. (2013a). The cost of collaboration for code and art: evidence from a remixing community. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, pages 1035–1046, New York, NY, USA. ACM.

Hill, B. M. and Monroy-Hernandez, A. (2013b). The remixing dilemma the trade-off between generativity and originality. *American Behavioral Scientist*, 57(5):643–663.

Hill, B. M. and Shaw, A. (2013). The wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS ONE*, 8(6):e65782.

Hill, B. M. and Shaw, A. (2014). Consider the redirect: A missing dimension of wikipedia research. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, pages 28:1–28:4, New York, NY, USA. ACM.

Hill, B. M., Shaw, A., and Benkler, Y. (2012). Status, social signalling and collective action in a peer production community. Working Paper.

Howison, J., Conklin, M., and Crowston, K. (2006). FLOSSmole. *International Journal of Information Technology and Web Engineering*, 1(3):17–26.

Jullien, N. (2012). What we know about wikipedia: A review of the literature analyzing the project(s). SSRN Scholarly Paper ID 2053597, Social Science Research Network, Rochester, NY.

Kellogg, K. (2009). Operating room: Relational spaces and micro-institutional change inside two surgical teaching hospitals. *American Journal of Sociology*, (Forthcoming).

Kittur, A. and Kraut, R. E. (2010). Beyond wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW)*, pages 215–224, Savannah, Georgia, USA. ACM.

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

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, pages 1301–1318, New York, NY, USA. ACM.

Kraut, R. E. and Resnick, P. (2012). *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.

Kriplean, T., Beschastnikh, I., and McDonald, D. W. (2008). Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW2008)*, pages 47–56, San Diego, CA, USA. ACM.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

Lipset, S. M., Trow, M. A., and Coleman, J. S. (1956). *Union democracy: The Internal Politics of the International Typographical Union*. Free Press, Glencoe, Ill.

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

McCarthy, J. D. and Zald, M. N. (1977). Resource mobilization and social movements: A partial theory. *The American Journal of Sociology*, 82(6):1212–1241.

McCarthy, J. D. and Zald, M. N. (2001). The enduring vitality of the resource mobilization theory of social movements. In Turner, J. H., editor, *Handbook of Sociological Theory*, Handbooks of Sociology and Social Research, pages 533–565. Springer US. 10.1007/0-387-36274-6_25.

Meyer, D. S. and Whittier, N. (1994). Social movement spillover. *Social Problems*, 41(2):277–298.

Minkoff, D. C. (1997). The sequencing of social movements. *American Sociological Review*, 62(5):779–799.

Monroy-Hernández, A., Hill, B. M., Gonzalez-Rivero, J., and boyd, d. (2011). Computers can't give credit: How automatic attribution falls short in an online remixing community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3421–3430, New York, NY, USA. ACM.

Mugar, G., Østerlund, C., Hassman, K. D., Crowston, K., and Jackson, C. B. (2014). Planet hunters and seafloor explorers: Legitimate peripheral participation through practice proxies in online citizen science. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 109–119, New York, NY, USA. ACM.

Østerlund, C. and Crowston, K. (2011). What characterize documents that bridge boundaries compared to documents that do not? an exploratory study of documentation in FLOSS teams. In *2011 44th Hawaii International Conference on System Sciences (HICSS)*, pages 1–10.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Preece, J. and 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., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 259–268, New York, NY, USA. ACM.

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

Resnick, P., Kuwabara, K., Zeckhauser, R., and Friedman, E. (2000). Reputation systems. *Communications of the ACM*, 43(12):45–48.

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

Scott, W. R. and Davis, G. F. (2006). *Organizations and organizing: Rational, natural and open systems perspectives*. Prentice Hall.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, New York, NY.

Shaw, A. (2012). Centralized and decentralized gatekeeping in an open online collective. *Politics & Society*, 40(3):349–388.

Shaw, A. and Benkler, Y. (2012). A tale of two blogospheres: Discursive practices on the left and right. *American Behavioral Scientist*, 56(4):459–487.

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

Shaw, A., Zhang, H., Monroy-Hernández, A., Munson, S., Hill, B. M., Gerber, E., Kinnaird, P., and Minder, P. (2014). Computer supported collective action. *interactions*, 21(2):74–77.

Shaw, A. D., Horton, J. J., and Chen, D. L. (2011). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 275–284, New York, NY, USA. ACM.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, USA, 1 edition.

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

Steinfield, C. W., Markus, M. L., and Wigand, R. T. (2005). Exploring interorganizational systems at the industry level of analysis: evidence from the US home mortgage industry. *Journal of Information Technology*, 20(4):224–233.

Stodden, V. and Miguez, S. (2013). Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. SSRN Scholarly Paper ID 2322276, Social Science Research Network, Rochester, NY.

von Krogh, G., Haefliger, S., Spaeth, S., and Wallin, M. W. (2012). Carrots and rainbows: Motivation and social practice in open source software development. *MIS Quarterly*, 36(2):649–676.

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

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M. (2011). Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA. ACM.

Zhu, H., Chen, J., Matthews, T., Pal, A., Badenes, H., and Kraut, R. E. (2014a). Selecting an effective niche: An ecological view of the success of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 301–310, New York, NY, USA. ACM.

Zhu, H., Kraut, R., and Kittur, A. (2012). Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 935–944, New York, NY, USA. ACM.

Zhu, H., Kraut, R. E., and Kittur, A. (2014b). The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 281–290, New York, NY, USA. ACM.